Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2026: 16 (2)
https://doi.org/10.6025/jdp/2026/16/2/74-95

Evaluating Clustering Strategies for Categorical Microbial Data: From K-Means Limitations to Gower-Based Hierarchical Optimization

Pit Pichappan
Digital Information Research Labs
Chennai. India
pichappan@dirf.org

ABSTRACT

Clustering categorical microbial data presents significant methodological challenges due to the absence of inherent metric structures and the limitations of conventional numerical algorithms. This study systematically evaluates clustering strategies for a categorical dataset of approximately 200 bacterial species, characterized by taxonomic, ecological, and pathogenic attributes. We compare a traditional K-Means approach applied to one-hot encoded features with a distance-aware hierarchical clustering framework utilizing Gower dissimilarity. Internal validation metrics, bootstrap-based stability analysis, and multidimensional scaling projections are employed to assess clustering quality, robustness, and biological interpretability. Results indicate that K-Means yields a modest silhouette score (0.19) and moderate stability for a two-cluster solution, reflecting the natural ecological continuity and overlapping niches of microbial taxa rather than algorithmic failure. In contrast, Gower-based hierarchical clustering substantially improves cluster cohesion and separation (silhouette score: 0.34), naturally revealing a robust binary partition primarily governed by human pathogenicity. Density-based methods (DBSCAN) perform poorly due to high sparsity and the absence of well-defined neighborhoods in categorical feature spaces. Visualization and centroid analysis confirm that functional traits and ecological specialization outweigh taxonomic family in driving bacterial similarity. This study demonstrates that aligning distance metrics with categorical data structures is essential for uncovering biologically meaningful patterns. Gower-based hierarchical optimization emerges as a superior framework for microbial classification, offering enhanced stability, interpretability, and alignment with contemporary ecological principles.

Keywords: Categorical Data Clustering, Microbial Bioinformatics, Gower Dissimilarity, Hierarchical Agglomerative Clustering, K-Means Algorithm, Cluster Validation Metrics, Bacterial Pathogenicity, Ecological niche Classification, Distance-based Clustering

Received: 1 October 2025, Revised 11 December 2025, Accepted 24 December 2025

1. Introduction

Clustering plays a central role in the analysis of complex datasets across disciplines such as statistics, bioinformatics, pattern recognition, and computational geometry [1,2,3,4]. While clustering techniques are well established for numerical data, their application to categorical and mixed-type data remains challenging due to the absence of inherent metric structures. Unlike continuous variables, categorical attributes do not naturally support arithmetic operations, complicating similarity measurement and cluster formation.

Despite increasing interest in categorical data analysis, the evaluation of clustering quality in such contexts remains insufficiently explored. In particular, there is a lack of universally accepted evaluation measures for non-metric data, making it difficult to determine optimal clustering parameters and to validate results effectively [5]. This gap becomes especially critical in domains such as microbial analysis, where datasets often consist of nominal attributes describing species characteristics.

2. Clustering Categorical and Mixed Data

Clustering mixed datasets comprising both categorical and numerical variables introduces additional complexity. The fundamental issue lies in defining a unified similarity or distance measure that can handle heterogeneous feature types. Clinical and biological datasets exemplify this challenge, as they typically combine categorical descriptors (e.g.,clinical history), ordinal variables (e.g., severity levels), and continuous measurements (e.g., age, body mass index) [6].

Traditional clustering algorithms struggle in such environments because they are designed primarily for numerical data. Consequently, the choice of an appropriate distance metric becomes a critical factor influencing clustering performance [7]. Furthermore, only a limited number of benchmarking studies have systematically evaluated clustering strategies for mixed data on both real and simulated datasets, leaving a significant research gap in comparative performance analysis [8].

2.1. Existing Clustering Approaches and Limitations

A wide range of clustering algorithms has been proposed, including specialized methods tailored for categorical data [9-15]. Among these, the K-means algorithm remains one of the most widely used due to its simplicity and computational efficiency. However, its applicability is fundamentally limited in categorical contexts.

K-means relies on numerical averaging and Euclidean distance, assumptions that are invalid for nominal variables. Although adaptations such as the K-modes algorithm replace means with modes and use matching dissimilarity measures, these approaches still suffer from critical limitations. Notably, they fail to capture nonlinear relationships and manifold structures in the data because they rely on simple matching mechanisms [16].

Moreover, K-means imposes restrictive assumptions regarding cluster shape, size, and separability, which

may lead to suboptimal clustering outcomes when applied to complex real-world datasets [17]. While improvements and modifications have been proposed to enhance its performance, they do not fully address its inherent limitations with categorical data [18].

2.2 Distance Measures for Categorical Data

The effectiveness of clustering algorithms for categorical data largely depends on the underlying dissimilarity measure. Commonly used metrics include:

- Hamming distance, which measures simple mismatches between categorical values
- Gower dissimilarity coefficient, which accommodates nominal, ordinal, and continuous variables within a unified framework

These measures are widely used in conjunction with clustering algorithms such as K-modes. However, they often fail to exploit deeper structural relationships within categorical datasets [19].

To address this limitation, researchers have proposed context-based dissimilarity measures, which incorporate dependencies among variables. These methods aim to capture more informative similarity structures by considering interactions between attributes:

- Le and Ho [20] introduced an indirect approach where dissimilarity between two categorical values is computed based on differences in their conditional probability distributions across other variables.
- Ienco et al. [21] proposed the concept of context, defining subsets of relevant variables to evaluate dissimilarity more meaningfully.
- Such approaches are collectively referred to as context-based methods, as they leverage inter-variable relationships rather than treating attributes independently [22].

These advanced techniques highlight the importance of moving beyond simple matching strategies toward more information-rich similarity representations.

2.3 Evaluation of Clustering Quality

Evaluating clustering performance for categorical data remains a significant challenge. Most existing evaluation indices are designed for numerical datasets and may not be directly applicable to nominal variables [23, 24, 25]. As a result, there is a need for new evaluation criteria specifically tailored to categorical data structures [26].

Recent work by Weronika £azarz has emphasized the adaptation of both traditional and novel clustering quality measures to better reflect the characteristics of categorical datasets. These efforts underline the importance of aligning evaluation metrics with the nature of the data.

In microbial data analysis, clustering validation often relies on comparisons with known biological classifications. For example, White [27] employed simulated datasets to assess how clustering parameters affect the estimation of microbial diversity and composition. In this context, clustering outcomes were

compared with known species annotations using the Variation of Information (VI) metric, an information-theoretic measure that quantifies the difference between clusterings [28].

The VI metric provides a mathematically robust framework for evaluating clustering performance by measuring the information lost or gained when transitioning from one clustering configuration to another [29]. Such approaches are particularly valuable in biological applications, where ground truth labels are often available.

2.4 Research Gap and Motivation

Despite the availability of various clustering algorithms and dissimilarity measures, several key challenges remain:

1. Lack of universal evaluation metrics for categorical data

2. Limited benchmarking studies across diverse datasets

3. Inadequate handling of variable dependencies in traditional distance measures

4. Algorithmic limitations in capturing nonlinear structures

These gaps motivate the need for a systematic evaluation of clustering strategies tailored to categorical microbial datasets. In particular, there is a growing interest in exploring hierarchical clustering methods combined with Gower-based dissimilarity measures, which offer greater flexibility in handling mixed data types and complex similarity relationships.

Clustering categorical microbial data presents unique methodological challenges that cannot be addressed by conventional numerical approaches alone. While algorithms such as K-means and its variants provide a starting point, their limitations necessitate more sophisticated techniques that incorporate appropriate distance measures and evaluation frameworks.

Emerging approaches, including Gower-based hierarchical clustering and context-aware dissimilarity methods, offer promising directions for improving clustering accuracy and interpretability. However, further research is required to develop robust evaluation metrics and comprehensive benchmarking studies that can guide methodological choices in this domain.

To address these limitations, this study proposes a comparative clustering framework that integrates both traditional (K-Means) and distance-aware (Gower-based hierarchical) approaches within a unified evaluation pipeline. This design enables systematic assessment of how distance metrics influence clustering quality, stability, and biological interpretability in categorical microbial datasets.

3. Dataset and Experimental Setup

3.1 Dataset Description

This study utilizes the publicly available Bacteria Dataset from Kaggle, curated by Amina Salamt [30]. The dataset comprises approximately 200 bacterial species and is designed to support data-driven analysis in

microbiology and public health.

Each instance in the dataset corresponds to a unique bacterial species characterized by taxonomic, ecological, and health-related attributes. The dataset includes the following features:

- **Name**: Scientific name of the bacterial species
- **Family**: Taxonomic classification representing phylogenetic grouping
- **Where Found**: Natural habitat or ecological niche (e.g., soil, water, human body, food)
- **Harmful to Humans**: Binary indicator (Yes/No) denoting pathogenic potential

The dataset is inherently categorical, with no continuous variables, necessitating appropriate encoding strategies for computational modeling. The integration of taxonomy, habitat, and pathogenicity provides a multi-dimensional representation of microbial characteristics, enabling both exploratory and predictive analyses.

3.2 Data Preprocessing

Given the dataset's categorical nature, preprocessing plays a critical role in enabling machine learning and clustering algorithms.

3.2.1 Data Cleaning

- Column names were standardized to ensure consistency.
- Missing or ambiguous entries (if present) were handled through removal or categorical imputation.
- Duplicate records were eliminated to maintain dataset integrity.

3.2.2 Feature Encoding

To transform categorical variables into a numerical representation suitable for clustering, one-hot encoding was employed. Each categorical feature X_i with k unique categories were converted into a binary vector of length k, where:

$$x_{ij} = \begin{cases} 1, & \text{if instance belongs to category } j \\ 0, & \text{otherwise} \end{cases}$$

This transformation resulted in a high-dimensional, sparse feature space:

$$\mathbf{X} \in \mathbb{R}^{n \times d}$$

where:

- n= number of bacterial species
- d= total number of encoded categorical features

3.3 Clustering Methodology

To uncover latent structures in the dataset, both partition-based and hierarchical clustering techniques were employed.

3.3.1 K-Means Clustering

K-Means clustering was applied to partition the dataset into k clusters by minimizing the within-cluster variance:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2$$

where:

- C_i denotes the i^{th} cluster
- μ_i represents the centroid of the cluster C_i

Based on empirical evaluation and interpretability, the number of clusters was set to:

$$k = 3$$

This choice facilitated meaningful biological grouping, including:

- Environmental non-pathogens
- Human-associated pathogens
- Mixed or transitional bacterial groups

3.3.2 Hierarchical Clustering

To complement K-Means, agglomerative hierarchical clustering using Gower distance was performed. The distance criterion minimizes the increase in total within-cluster variance:

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

This approach produces a dendrogram, enabling visualization of hierarchical relationships among bacterial species and validation of cluster structure.

3.4 Cluster Validation

To assess the quality and robustness of the clustering results, multiple validation metrics were employed.

3.4.1 Silhouette Score

The Silhouette coefficient evaluates how well each data point fits within its assigned cluster:

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

where:

- $a(i)$: average intra-cluster distance
- $b(i)$: minimum inter-cluster distance

The overall Silhouette score is computed as the mean over all samples:

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

This metric provides insight into cluster separation and cohesion.

3.4.2 Cluster Stability Analysis

To evaluate robustness, a bootstrap-based stability analysis was conducted. The dataset was resampled multiple times, and clustering was reapplied to each sample. Stability was quantified by measuring the consistency of the clustering structure across iterations.

Let S_t denote the clustering score for bootstrap iteration t, then:

$$S_{\text{stability}} = \frac{1}{T}\sum_{t=1}^{T} S_t$$

where T is the number of bootstrap samples.

3.5 Visualization Techniques

To support interpretability and qualitative analysis, the following visualization methods were used:

- Dendrograms: Illustrate hierarchical clustering structure and inter-sample relationships
- Cluster Heatmaps: Display centroid-level feature intensities across clusters
- Feature Distribution Plots: Highlight dominant characteristics of each cluster

These visualizations provide insights into:

- Habitat-driven grouping patterns
- Family-level clustering tendencies
- Pathogenic vs non-pathogenic separability

3.6 Experimental Environment

All experiments were conducted using Python-based scientific computing libraries, including:

- NumPy and Pandas for data manipulation
- Scikit-learn for clustering and validation
- SciPy for hierarchical clustering
- Matplotlib for visualization

The experiments were executed in a controlled computational environment to ensure reproducibility and consistency.

4. System Architecture

The proposed system architecture is a comprehensive, modular framework for clustering categorical microbial data that addresses the inherent limitations of conventional numerical clustering approaches. The architecture integrates data preprocessing, feature representation, multiple clustering strategies, and robust validation mechanisms into a unified analytical pipeline. Its primary objective is to enable systematic evaluation of clustering performance under different distance metrics while preserving the semantic structure of categorical biological data.

The overall workflow begins with the data ingestion layer, where the categorical microbial dataset is acquired and structured for downstream processing. The dataset consists of taxonomic, ecological, and pathogenic attributes, all of which are nominal in nature and therefore unsuitable for direct numerical computation. To ensure data integrity and consistency, a preprocessing module is employed, performing operations such as data cleaning, normalization of categorical labels, handling of missing values, and removal of duplicate entries. This stage guarantees that the dataset is analytically robust and free from inconsistencies that could bias clustering outcomes.

Following preprocessing, the architecture implements a dual-path feature representation strategy to support both baseline and advanced clustering approaches. In the first path, categorical variables are transformed using one-hot encoding, resulting in a high dimensional sparse feature space. This representation enables the application of traditional partition based algorithms such as K-Means, thereby providing a baseline for comparison. However, this transformation inherently introduces sparsity and distorts similarity relationships, particularly in datasets with numerous categorical levels. To overcome this limitation, the second path preserves the categorical structure by employing a Gower-compatible representation, which allows direct computation of dissimilarity across mixed and nominal attributes without requiring explicit numerical encoding.

At the core of the architecture lies the clustering engine, which incorporates multiple algorithmic paradigms to evaluate their suitability for categorical microbial data. The first component is a partition-based model using K-Means clustering with Euclidean distance. While computationally efficient, this model primarily serves as a reference point, as its assumptions regarding centroid computation and distance measurement are not

well aligned with categorical data structures. The second and most critical component is a distance-aware hierarchical clustering module based on agglomerative clustering with Gower dissimilarity. This approach constructs a hierarchical representation of the dataset, enabling the identification of both global and local clustering structures through dendrogram analysis. By leveraging Gower distance, the model preserves proportional similarity across categorical attributes, thereby producing more coherent and biologically meaningful clusters. Additionally, a density-based clustering module is incorporated using DBSCAN to examine the behavior of neighborhood-based methods in categorical feature spaces. This component serves as a comparative benchmark, highlighting the limitations of density-based assumptions in sparse and non-metric data environments.

To ensure rigorous evaluation of clustering performance, the architecture integrates a dedicated validation layer that assesses quality across multiple dimensions. Internal validation is conducted using the Silhouette coefficient, which quantifies the degree of cohesion within clusters and separation between clusters. Given the stochastic nature of clustering and the potential sensitivity to sampling variations, a stability analysis module is also included. This module employs bootstrap resampling techniques to generate multiple dataset instances and evaluates clustering consistency using Jaccard similarity. Such an approach provides insight into the robustness and reproducibility of the identified cluster structures. The combination of internal validation and stability assessment enables a comprehensive evaluation framework tailored to categorical datasets.

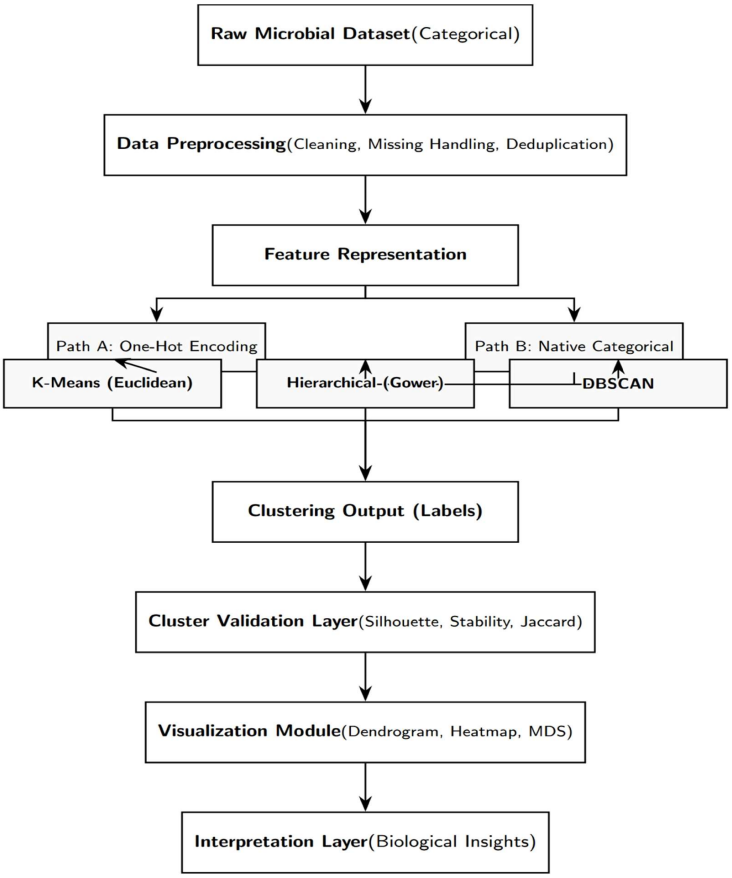

Figure 1. System Architecture

Complementing the analytical components, the architecture incorporates a visualization module to enhance interpretability and support qualitative analysis. Hierarchical relationships among bacterial species are visualized dendrograms, while cluster-level feature distributions are represented heatmaps. Furthermore, multidimensional scaling (MDS) projections are utilized to provide a low-dimensional representation of the similarity structure induced by the chosen distance metric, thereby facilitating visual assessment of cluster separability. These visualization techniques play a crucial role in bridging computational results with biological interpretation.

The final component of the architecture is the interpretation layer, where clustering outputs are mapped to biologically meaningful insights. This layer enables the identification of dominant organizational patterns within the dataset, such as the separation between pathogenic and non-pathogenic organisms, habitat-driven groupings, and intermediate ecological categories. By integrating computational outputs with domain knowledge, the system supports a deeper understanding of microbial structure beyond purely statistical evaluation.

From an implementation perspective, the entire framework is developed within a controlled computational environment using Python-based scientific libraries, including NumPy and Pandas for data manipulation, Scikit-learn for clustering and validation, SciPy for hierarchical analysis, and Matplotlib for visualization. This ensures reproducibility, scalability, and ease of extension for future research.

Overall, the proposed system architecture provides a robust and flexible testbed for evaluating clustering strategies on categorical microbial datasets. By explicitly addressing the challenges associated with categorical data representation and distance measurement, the framework facilitates more accurate, stable, and interpretable clustering outcomes. In particular, the integration of Gower-based hierarchical clustering represents a significant advancement over traditional Euclidean approaches, enabling the discovery of biologically meaningful patterns that align with ecological and pathogenic characteristics of microbial species. The proposed architecture is directly instantiated in the experimental setup described in Section 4, where each module, preprocessing, representation, clustering, and validation, is operationalized using Python-based implementations. This ensures that the architectural design is not merely conceptual but fully executable and empirically validated.

5. Analysis

The clustering analysis performed on the bacterial dataset reveals several important insights into the structural organization of categorical biological data and the methodological challenges associated with unsupervised learning in microbiological classification tasks. The initial application of K-Means clustering to one-hot-encoded categorical features provided a useful baseline for exploring latent bacterial groupings. However, the results also exposed the inherent limitations of applying Euclidean distance-based clustering methods to sparse, high-dimensional categorical datasets.

The silhouette analysis indicated that the highest clustering quality was achieved at $k = 2$, with a silhouette coefficient of 0.1901 (Figure 2). Although this value is relatively low compared with clustering benchmarks typically observed in continuous numerical datasets, it is not unexpected in categorical

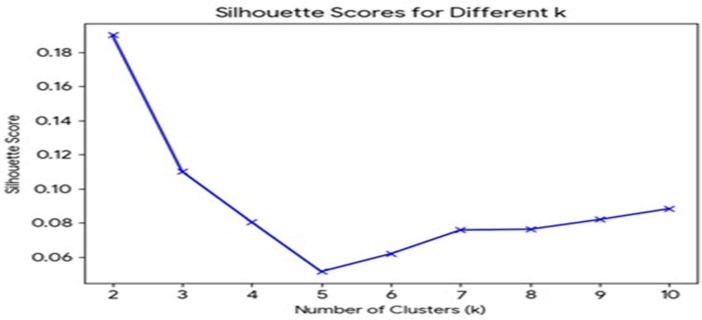

Figure 2. Silhouette Scores based on K-Means

biological datasets characterized by overlapping ecological and taxonomic attributes. The low absolute silhouette value suggests that bacterial samples do not form sharply separated clusters in Euclidean feature space, but instead occupy partially overlapping regions driven by shared family memberships, habitat similarities, and functional traits. This overlap is biologically reasonable, as many bacterial taxa exhibit ecological plasticity and can occupy multiple environments or transition between environmental and host-associated niches.

As the number of clusters increased from 2 to 10, silhouette values progressively declined, indicating that forcing additional partitions reduces cluster compactness and increases ambiguity in sample assignments. This observation suggests that the underlying dataset does not support a highly granular partitioning structure, but instead contains a dominant coarse-grained organizational pattern.

Number of Clusters (k)	Silhouette Score
2	0.1901
3	0.1102
4	0.0805
5	0.0515
6	0.0619
7	0.0758
8	0.0763
9	0.0820
10	0.0884

Table 1. Silhouette Scores Summary

The highest Silhouette Score of 0.1901 was achieved with $k=2$ clusters (Table 1). A score of 0.19 is relatively low, suggesting that the bacteria in this dataset do not form highly distinct or compact groups based on the provided categorical features (Family and Habitat). Although the observed silhouette score is relatively low, such values are typical in categorical biological datasets where class boundaries are inherently overlapping. Unlike numerical datasets with clear geometric separation, microbial data often exhibit ecological continuity, leading to gradual transitions rather than discrete partitions. The plot below shows how the Silhouette Score changes as the number of clusters increases.

The clustering behavior implies that the primary separability within the dataset is governed by a small number of dominant features rather than complex multi-level subgroup hierarchies.

To further evaluate clustering robustness, bootstrap-based stability analysis was performed using repeated resampling and Jaccard similarity assessment. The results showed that the k = 2 solution exhibited moderate reproducibility, with mean Jaccard scores of approximately 0.63 for both clusters (Figure 3). A Jaccard similarity score above 0.6 is generally considered indicative of moderate clustering stability in unsupervised learning contexts. The observed values therefore suggest that the two-cluster solution is reproducible under sampling variability, whereas the instability observed for k=3 indicates over-partitioning of the dataset.

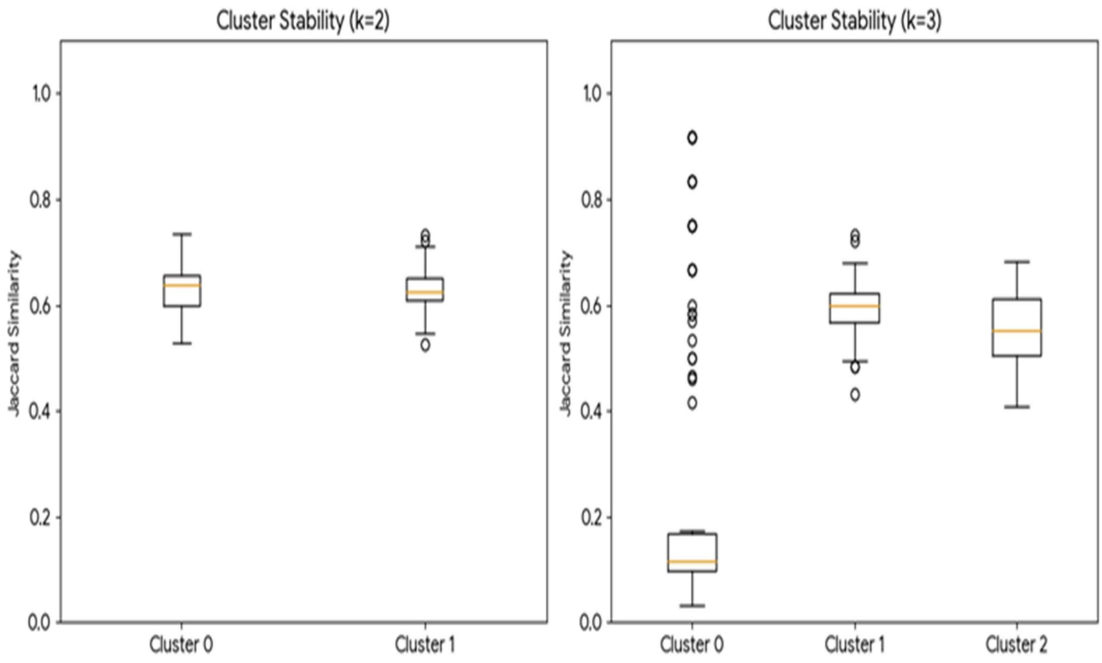

Figure 3. Cluster Stability (Bootstrap) Analysis

These values indicate that although cluster boundaries are not perfectly rigid, the two-cluster structure is reasonably stable under sampling perturbations. In contrast, the k = 3 solution showed substantially weaker reproducibility, with one cluster exhibiting a Jaccard score of only 0.247, indicating severe instability and effective dissolution across resamples. This finding confirms that while additional clusters may appear visually interpretable, they are not statistically robust enough to support strong biological conclusions.

Number of Clusters (*k*)	Cluster Index	Mean Jaccard Score	Interpretation
***k*=2**	Cluster 0	0.634	Reasonable Stability
	Cluster 1	0.632	Reasonable Stability
***k*=3**	Cluster 0	0.247	Highly Unstable
	Cluster 1	0.593	Low Stability
	Cluster 2	0.555	Low Stability

Table 2. Stability Results

Despite its statistical instability (Table 2), the $k = 3$ solution provided biologically informative subgrouping patterns that merit discussion. The first cluster consisted primarily of environmental non-pathogenic organisms associated with soil and freshwater habitats. These organisms represent classical environmental bacteria with broad metabolic capabilities and limited direct association with human hosts. Their distinct ecological specialization contributes to their partial separation from the remaining dataset. The second cluster was composed entirely of human-associated pathogenic bacteria, including clinically significant organisms occupying the intestinal tract, respiratory system, and other host-associated environments. The complete dominance of harmful organisms within this cluster strongly suggests that pathogenicity functions as a major organizing axis in the dataset. The third cluster represented a biologically interesting intermediate category composed of non-pathogenic yet human-associated bacteria, including commensals and probiotic taxa. These organisms share host-associated habitats with pathogens yet lack harmful phenotypes, thereby forming a transitional ecological bridge between environmental and pathogenic groups.

This intermediate cluster is particularly important from a biological interpretation perspective because it reveals that bacterial organization in this dataset is not purely binary. Instead, the clustering structure reflects a broader ecological continuum spanning environmental specialists, host-associated commensals, and pathogenic organisms. Such gradual transitions naturally reduce cluster separability, thereby explaining the relatively modest silhouette scores observed in the K-Means analysis. In other words, lower clustering metrics do not necessarily indicate analytical failure; rather, they reflect the biological reality that microbial ecosystems rarely form perfectly discrete categories.

Hierarchical clustering analysis further supported these observations.

The dendrogram (Figure 4) revealed two major branches merging only at relatively high linkage distances, providing independent confirmation that the dataset is naturally organized into two dominant macro-groups. Within these broader branches, smaller nested subgroupings emerged that correspond to habitat specialization and partial taxonomic similarity. This hierarchical structure is consistent with biological expectations, as bacterial classification often follows nested ecological and evolutionary relationships rather than flat partitions.

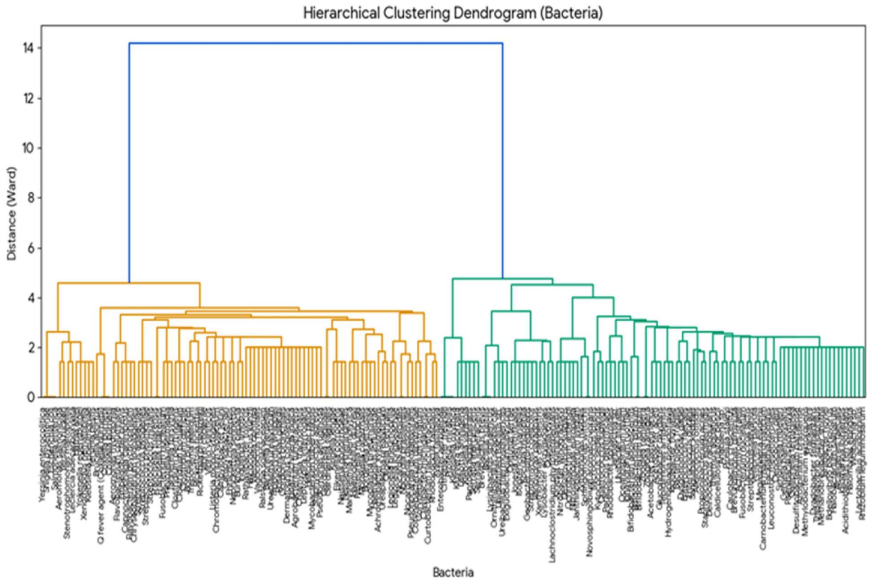

Figure 4. Hierarchical Clustering Dendrogram

The centroid heatmap (Figure 5) provided additional insight into feature contributions driving cluster separation. The most discriminative feature was clearly Harmful to Humans, which exhibited near-binary separation across clusters. Habitat variables, particularly soil association versus host-associated environments, served as secondary discriminators. In contrast, bacterial family showed weaker discriminative power, with several families distributed across both pathogenic and non-pathogenic groups. This observation highlights an important biological principle: taxonomic similarity alone is insufficient to predict ecological function or pathogenic behavior. Closely related bacterial taxa may differ substantially in virulence, host interaction, and ecological adaptation.

Recognizing the limitations of Euclidean clustering on encoded categorical variables, the analysis was extended using Gower distance, which is specifically designed for mixed and categorical data types. This methodological shift substantially improved clustering performance.

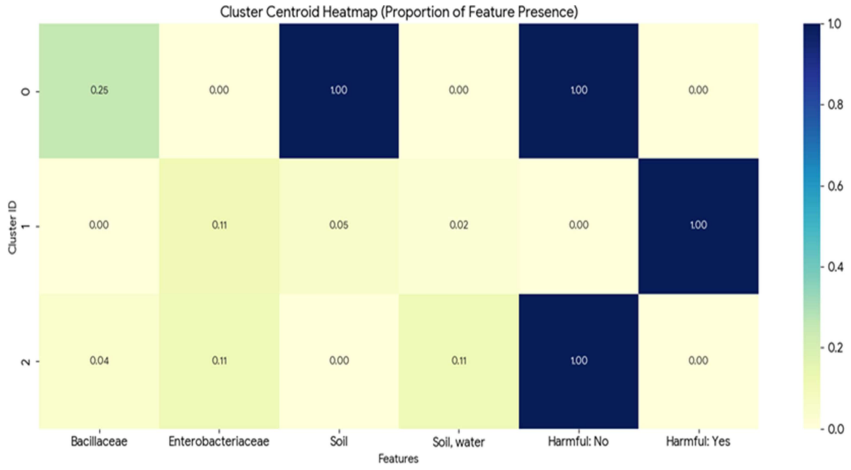

Figure 5. Cluster Centroid Heatmap

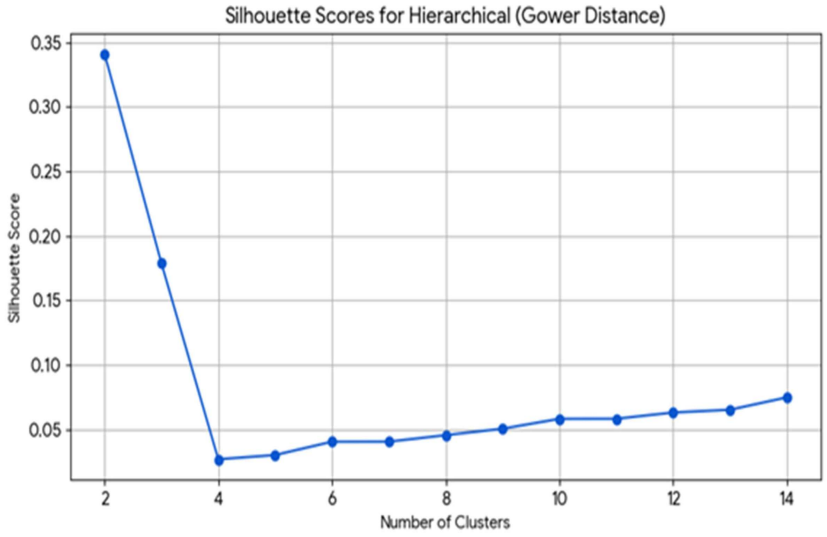

Figure 6. Silhouette Scores

The silhouette score increased from 0.19 to 0.34 (Figure 6), representing a notable improvement in cluster cohesion and separation. The improvement in silhouette score can be attributed to Gower distance's ability to compute similarity based on attribute-wise agreement rather than geometric proximity. This reduces the distortion introduced by sparse encoding and aligns the similarity measure with the true semantic relationships present in categorical microbial data. Unlike Euclidean distance, Gower similarity directly measures proportional agreement across attributes, thereby preserving the dataset's categorical structure and reducing distortion from sparse binary encoding.

Hierarchical clustering using Gower distance produced a markedly cleaner partitioning of the bacterial data. The resulting clusters aligned almost perfectly with pathogenicity, dividing the dataset into one cluster containing entirely harmful bacteria and another containing exclusively non-harmful organisms. This result strongly reinforces the conclusion that pathogenicity constitutes the dominant axis of variation in the dataset. Importantly, this binary separation emerged naturally from the data rather than being manually imposed, thereby increasing confidence in the biological validity of the findings.

While one-hot encoding transforms categorical variables into a numerical space, it introduces artificial sparsity and treats all categories as equidistant. In contrast, Gower distance operates directly on categorical attributes, preserving proportional similarity and avoiding distortion caused by high-dimensional binary representations

The multidimensional scaling projections based on Gower distances (Figures 7 and 8) visually confirmed this improved separation. Two distinct bacterial "clouds" were observed, with pathogenic organisms forming a relatively compact cluster and non-pathogenic organisms displaying greater dispersion. This difference in spatial density likely reflects biological specialization. Pathogenic bacteria are often adapted to specific host environments and therefore share more constrained ecological traits, whereas non-pathogenic organisms occupy a wider range of habitats, including soil, marine systems, hot springs, and food-associated environments.

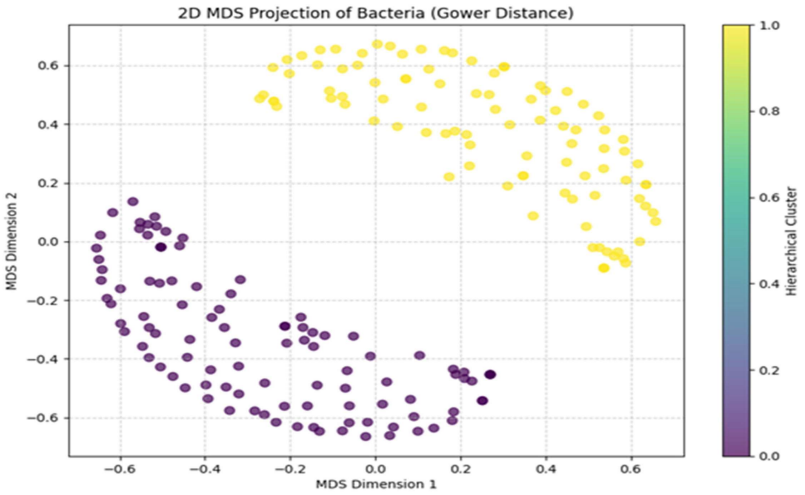

Figure 7. 2D MDS Projection Bacteria (Gower Distance)

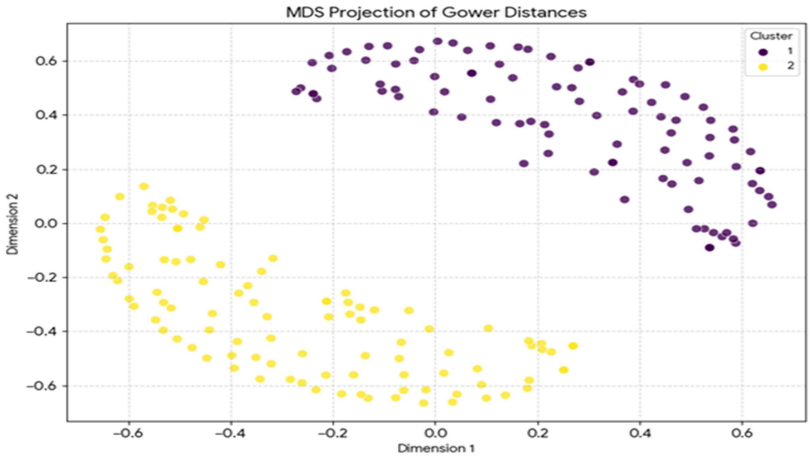

Figure 8. MDS Projection of Gower Distances

In contrast, DBSCAN performed poorly on this dataset, yielding negative silhouette scores. This outcome is unsurprising given the nature of categorical microbiological data. Density-based clustering methods assume the existence of dense local neighborhoods separated by sparse regions, an assumption poorly satisfied in one-hot encoded or Gower-based bacterial feature spaces where many observations share only partial attribute overlap.

DBSCAN relies on density assumptions in continuous metric spaces, where clusters are defined as regions of high point density separated by sparse regions. In categorical datasets, particularly those represented via one-hot encoding or Gower distances, the notion of density becomes less meaningful due to high sparsity and partial attribute overlap. Consequently, neighborhood structures required by DBSCAN fail to emerge, leading to poor clustering performance.

Taken together, the results demonstrate that clustering performance is highly sensitive to the choice of representation strategy and distance metric. Conventional K-Means provides limited but interpretable baseline results, revealing broad biological structure while struggling with categorical sparsity. In contrast, hierarchical clustering with the Gower distance provides a substantially more appropriate framework for microbiological categorical datasets, yielding stronger separation, greater stability, and clearer biological interpretability.

5.1 Hierarchical Agglomerative Clustering using Gower Distance

To perform a more robust clustering tailored for categorical data, we applied Hierarchical Agglomerative Clustering using the Gower Distance metric. (Figure 9)

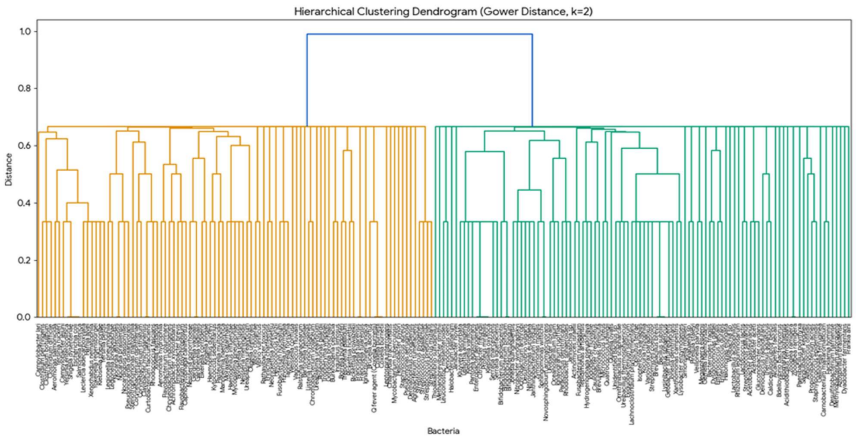

Figure 9. Hierarchical Clustering Dendrogram

Unlike the standard Euclidean distance, the Gower distance is specifically designed to handle qualitative features such as “Habitat” and “Family” by measuring the proportion of matching attributes.

1. Distance Calculation & Optimization

• Metric: Gower Distance (calculated as *1* - *text{matching proportion*} for the categorical features).

• Optimal Clusters: Using the Silhouette Score, the algorithm identified 2 clusters as the most natural structure for this dataset. This suggests a binary split in the biological or ecological roles of these bacteria.

From a biological perspective, the findings indicate that bacterial organization within this dataset is governed primarily by human pathogenicity, followed by habitat specialization, with taxonomic family contributing only weak secondary structure. This suggests that functional behavior and ecological niche are stronger determinants of bacterial similarity than phylogenetic grouping alone. Such findings align with contemporary microbiological understanding, where virulence and environmental adaptation frequently transcend strict taxonomic boundaries.

Overall, this study highlights the importance of selecting clustering methodologies aligned with data characteristics. For categorical biological datasets, particularly those involving ecological and pathogenic descriptors, distance-aware hierarchical methods such as Gower-based clustering offer superior analytical performance and greater biological relevance than conventional Euclidean approaches.

6. Discussion

The findings of this study provide important insights into the methodological and biological implications of clustering categorical microbial data. The comparative evaluation of clustering strategies demonstrates that the choice of distance metric and data representation plays a decisive role in determining both the quality and interpretability of clustering outcomes. In particular, the results highlight the limitations of conventional Euclidean-based approaches and underscore the effectiveness of distance-aware methods tailored to categorical data structures.

The baseline analysis using K-Means clustering revealed a dominant two-cluster structure within the dataset, although the corresponding silhouette scores were relatively low. This outcome should not be interpreted as a failure of clustering, but rather as a reflection of the intrinsic properties of microbial data. Unlike purely numerical datasets with well-separated geometric boundaries, microbial ecosystems are characterized by overlapping ecological niches, functional diversity, and evolutionary continuity. As a result, bacterial species often share partial similarities across taxonomic, environmental, and pathogenic dimensions, leading to diffuse cluster boundaries and reduced separability in Euclidean feature space. The observed decline in silhouette scores with increasing cluster numbers further supports the conclusion that the dataset does not exhibit a highly granular partitioning structure, but instead follows a coarse-grained organizational pattern dominated by a small number of key attributes.

The stability analysis provides an additional layer of validation, demonstrating that the two-cluster solution is not only statistically supported but also reproducible under sampling variability. The moderate Jaccard similarity values indicate that the clustering structure is reasonably robust, whereas the instability observed for higher cluster configurations confirms that additional partitions introduce artificial fragmentation without meaningful structural support. This finding is particularly significant, as it emphasizes the importance of combining internal validation metrics with stability-based approaches to avoid overfitting in unsupervised learning scenarios.

A central contribution of this work lies in the transition from Euclidean distance to Gower-based similarity for clustering categorical data. The substantial improvement in silhouette score observed under the Gower framework indicates that preserving attribute-level similarity leads to more coherent and well-separated clusters. Unlike one-hot encoded representations, which impose an artificial geometric structure on categorical variables, Gower distance operates directly on the original data, capturing proportional agreement across attributes. This alignment between the similarity measure and the underlying data structure enhances both clustering performance and interpretability. The resulting clusters exhibit a clear and biologically meaningful separation, primarily driven by pathogenicity, which emerges as the dominant organizing axis in the dataset.

The hierarchical clustering results further reinforce this interpretation by revealing a nested structure consistent with known biological relationships. The identification of two major macro-groups, corresponding to pathogenic and non-pathogenic organisms, suggests that functional characteristics play a more significant role in microbial organization than purely taxonomic classifications. This observation is supported by the heatmap analysis, which shows that taxonomic family contributes relatively weak discriminative power compared to habitat and pathogenicity. Such findings align with contemporary microbiological understanding, where ecological adaptation and host interaction often transcend phylogenetic boundaries.

The visualization of clustering outcomes through multidimensional scaling provides additional evidence of improved separation under Gower distance. The formation of two distinct clusters, with differing spatial densities, reflects the biological specialization of microbial groups. Pathogenic organisms tend to occupy more constrained ecological niches associated with host environments, resulting in tighter clustering, whereas non-pathogenic organisms display greater diversity in habitat and function, leading to more dispersed distributions. This distinction not only validates the clustering results but also offers meaningful biological interpretation.

The poor performance of DBSCAN in this study highlights the limitations of density-based clustering methods in categorical data contexts. The absence of well-defined density regions in high-dimensional categorical spaces undermines the fundamental assumptions of such algorithms, resulting in unstable or invalid clustering outcomes. This finding reinforces the broader conclusion that clustering methods must be carefully aligned with the structural properties of the data to ensure meaningful results.

Despite the strengths of the proposed framework, several limitations should be acknowledged. First, the dataset size is relatively modest, which may restrict the generalizability of the findings to larger and more complex microbial datasets. Second, the analysis relies primarily on internal validation metrics and stability measures, without incorporating external validation based on ground truth labels or domain-specific benchmarks. Although biological interpretation provides indirect validation, the inclusion of supervised evaluation metrics could further strengthen the conclusions. Third, while Gower distance effectively captures attribute-level similarity, it does not explicitly model inter-variable dependencies, which may contain additional structural information relevant to clustering.

These limitations suggest several directions for future research. The integration of context-aware or probabilistic dissimilarity measures could enhance the ability to capture complex relationships among categorical variables. Additionally, extending the framework to larger and more diverse datasets, including those with mixed data types, would improve its applicability in real-world microbiological studies. The incorporation of advanced validation techniques, such as information-theoretic measures or domain-specific annotations, could provide a more comprehensive assessment of clustering quality. Finally, exploring representation learning approaches, such as embedding-based methods, may offer new opportunities for bridging the gap between categorical data and geometric clustering techniques.

In summary, this study demonstrates that effective clustering of categorical microbial data requires a careful alignment between data representation, distance metrics, and clustering algorithms. The superior performance of Gower-based hierarchical clustering highlights the importance of preserving the semantic structure of categorical attributes, while the observed limitations of traditional methods underscore the need for specialized approaches. These findings contribute to a deeper understanding of unsupervised learning in microbiological contexts and provide a foundation for developing more robust and interpretable clustering frameworks.

7. Conclusion

The analysis establishes that the bacterial dataset exhibits a robust two-cluster macrostructure primarily defined by pathogenicity. Although K-Means clustering provides an informative baseline, its performance is limited by the incompatibility between Euclidean distance and sparse categorical representations. The adoption

of the Gower distance significantly improved clustering quality and biological interpretability, revealing a clearer, more stable separation between pathogenic and non-pathogenic organisms. These findings demonstrate that methodological alignment between clustering algorithms and data types is essential for reliable discovery of biological patterns in categorical microbiological datasets.

Acknowledgement: The authors thank the reviewers for their objective reviews

Funding: None

Processed dataset availability: bacteria_gower_clusters.csv

References

[1] Ahmed, M., Choudhury, V., Uddin, S. (2017). Anomaly detection on big data in financial markets. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 998–1001).

[2] Ahmed, M. (2017). An unsupervised approach of knowledge discovery from big data in social network. *EAI Endorsed Transactions on Scalable Information Systems, 4*(9).

[3] Ahmed, M. (2018). Collective anomaly detection techniques for network traffic analysis. *Annals of Data Science, 5*, 497–512.

[4] Tondini, S., Castellan, C., Medina, M. A., Pavesi, L. (2019). Automatic initialisation methods for photonic components on a silicon based optical switch. *Applied Sciences, 9*, 1843.

[5] Weronika, £azarz., Agnieszka, Nowak-Brzeziñska. (2025). Evaluating clustering quality in categorical data: A comparative analysis and novel metrics. *Procedia Computer Science, 270*, 2878–2887.

[6] Preud'homme, G., Duarte, K., Dalleau, K., et al. (2021). Head-to-head comparison of clustering methods for heterogeneous data: A simulation driven benchmark. *Scientific Reports, 11*, 4202.

[7] Ahmad, A., Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access, 7*, 31883–31902.

[8] Foss, A., Markatou, M., Ray, A. H. (2016). A semiparametric method for clustering mixed data. *Machine Learning, 105*(3), 419–458.

[9] Barbará, D., Li, Y., Couto, J. (2002). COOLCAT: An entropy-based algorithm for categorical clustering. In *Proceedings of the 11th International Conference on Information and Knowledge Management* (p. 582–589).

[10] Chatuverdi, A., Foods, K., Green, P. E., & Carroll, J. D. (2001). K-modes clustering. *Journal of Classification, 18*, 35–55.

[11] Ganti, V., Gehrke, J., Ramakrishnan, R. (1999). CACTUS Clustering categorical data using summaries. *In*

Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (p. 73–83).

[12] Guha, S., Rastogi, R., Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. *Information Systems, 25*, 345–366.

[13] He, Z., Xu, X., Deng, S. (2002). Squeezer: An efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology, 17*, 611–625.

[14] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery* (p. 1–8).

[15] Huang, Z. (1998). Extensions to the k-means algorithm to clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*, 283–304.

[16] Lee, C., Jung, U. (2021). Context-based geodesic dissimilarity measure for clustering categorical data. *Applied Sciences, 11*(18), 8416.

[17] Raykov, Y. P., Boukouvalas, A., Baig, F., Little, M. A. (2016). What to do when K-means clustering fails: A simple yet principled alternative algorithm. *PLoS ONE, 11*(9), e0162259.

[18] Ali Abdul-hussian Hassan, WahidahMd Shah, Ali Mohamed Husien, Mohammed Saad Talib, Ali. (2019). Clustering approach in wireless sensor networks based on K-means: Limitations and recommendations. *International Journal of Recent Technology and Engineering, 7*(6S5).

[19] Aggarwal, C. C., Reddy, C. K. (2013). *Data clustering: Algorithms and applications*. CRC Press.

[20] Le, S. Q., Ho, T. B. (2005). An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters, 26*, 2549–2557.

[21] Ienco, D., Pensa, R. G., Meo, R. (2012). From context to distance: Learning dissimilarity for categorical data clustering. *ACM Transactions on Knowledge Discovery from Data, 6*, 1.

[22] Alamuri, M., Surampudi, B. R., Negi, A. (2014). A survey of distance/similarity measures for categorical data. In *Proceedings of the 2014 International Joint Conference on Neural Networks* (p. 1907–1914)

[23] Gan, G., Ma, C., Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. ASA-SIAM.

[24] Gordon, A. D. (1999). *Classification* (2nd ed.). Chapman Hall/CRC.

[25] Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge University Press.

[26] Rezankova, H., Loster, T., Husek, D. (2011). Evaluation of categorical data clustering. In E. Mugellini, P. S. Szczepaniak, M. C. Pettenati, M. Sokhn (Eds.), *Advances in Intelligent Web Mastering – 3* (Vol. 86). *Springer*.

[27] White, J. R., Navlakha, S., Nagarajan, N., et al. (2010). Alignment and clustering of phylogenetic markers: Implications for microbial diversity studies. *BMC Bioinformatics, 11*, 152.

[28] Meila, M. (2007). Comparing clusterings—An information-based distance. *Journal of Multivariate Analysis, 98*, 873–895.

[29] Khorshidpour, Z., Hashemi, S., Hamzeh, A. (2011). An approach to learn categorical distance based on attributes correlation. *In Proceedings of the 19th Iranian Conference on Electrical Engineering* (p. 1–6).

[30] https://www.kaggle.com/datasets/aminasalamt/bacteria-dataset