# A Multi-Paradigm Clustering Framework for Nutritional Pattern Discovery: Benchmarking and Insights from Fast Food Data

Lorena Ramirez and Kun Wang

***Abstract*—The rapid growth of fast food consumption has intensified p ublic h ealth c oncerns, y et d eriving a ctionable insights from nutritional datasets remains challenging due to high dimensionality, nonlinear feature interactions, and the absence of labeled outcomes. While unsupervised clustering offers a natural solution, existing studies typically evaluate algorithms in isolation, lacking a systematic multi-paradigm comparison tailored to nutritional data. To address this gap, we propose a multi-paradigm clustering framework that systematically evaluates six algorithms spanning centroid-based, hierarchical, probabilistic, density-based, graph-based, and kernel-based methods—KMeans, Agglomerative Clustering, Gaussian Mixture Models, DBSCAN, Spectral Clustering, and Mean Shift. A multi-dimensional evaluation strategy combining Silhouette Score, Principal Component Analysis (PCA) visualization, and a custom Clustering Validity Score is employed to assess both structural quality and interpretability. Experimental results reveal a clear divergence: DBSCAN achieves the highest Silhouette Score (0.5718) and excels at detecting irregular nutritional patterns and outliers, while KMeans and GMM, though yielding lower separation scores, produce more stable and interpretable clusters suitable for downstream applications. These findings underscore that no single algorithm universally dominates; rather, algorithm selection should align with application goals—density-based methods for exploratory pattern discovery, centroid-based methods for interpretable categorization. Our framework provides a reusable benchmarking methodology and practical guidance for applying unsupervised learning to nutritional data analysis.**

***Key Words*—Unsupervised Learning, Clustering, Nutritional Data Analysis, DBSCAN, Benchmarking.**

## I. Introduction

The rapid global expansion of the fast food industry has fundamentally reshaped dietary behaviors, escalating public health crises such as obesity and metabolic disorders [1]. While open nutritional databases provide abundant food composition data, transforming these raw measurements into actionable dietary insights remains a critical bottleneck [2]. This tension between data abundance and analytical capability motivates the present study.

Unsupervised clustering is a natural paradigm for exploring unlabeled nutritional datasets [3], with applications in dietary pattern identification and consumer segmentation [4]. However, existing literature exhibits notable fragmentation: studies typically optimize a single algorithm—frequently KMeans [5, 6]—without systematically comparing alternative paradigms[7]. While researchers highlight DBSCAN's advantages for irregular clusters [8, 9] or survey spectral clustering's capabilities [10], a unified experimental framework is lacking. Consequently, practitioners face an evidence gap when selecting among centroid, density, probabilistic, hierarchical, graph, or kernel-based paradigms [11]. Addressing this is essential, as algorithm selection directly shapes the discovered nutritional patterns and downstream dietary recommendations.

The first innovation of this work lies in framing nutritional clustering as a multi-paradigm benchmarking task rather than a single-algorithm optimization[12]. We ask: "Which clustering paradigm best reveals nutritional structure, and under what conditions?" This reframing acknowledges that no single algorithm universally dominates all data distributions [13, 14]. For instance, KMeans and hierarchical methods exhibit complementary strengths depending on cluster geometry [15, 16]. By applying a controlled evaluation protocol to a common dataset with identical preprocessing, we isolate algorithmic assumptions to enable fair, cross-paradigm comparisons.

The second innovation is a multi-dimensional evaluation methodology. Existing nutritional clustering studies predominantly rely on a single metric, such as the Silhouette Score [6], which cannot simultaneously capture cohesion, separation, interpretability, and noise robustness [17, 18]. To overcome this, we introduce a protocol combining: (a) Silhouette Score for quantitative separation; (b) PCA visualization for qualitative cluster geometry; and (c) a custom Clustering Validity Score (CVS) aggregating cluster purity and separation consistency without external labels. We deliberately evaluate six diverse algorithms: KMeans and Agglomerative Clustering (centroid/hierarchical) for interpretability [5, 16]; Gaussian Mixture Models (probabilistic) for overlapping profiles [19, 20]; DBSCAN (density-based) for irregular clusters and outliers [8]; and Spectral Clustering and Mean Shift (graph/kernel-based) for complex distributions [10, 21].

The contributions of this work are threefold:

1) **Multi-Paradigm Benchmarking:** We reframe nutritional clustering as a benchmarking problem to systematically test algorithmic assumptions rather than tacitly adopting them.
2) **Multi-Dimensional Evaluation:** We design and validate a holistic assessment protocol (Silhouette Score, PCA



L. Ramirez is with the Department of Computer Science, New Jersey City University, Jersey City 07302, USA.(e-mail:lramirezlapaix@njcu.edu)

K. Wang is with the College of Artificial I ntelligence a nd Software, Liaoning Petrochemical University, Fushun 113000, China .(e-mail: wk180558@gmail.com)

*Corresponding author: Kun Wang.*

visualization, and CVS) that serves as a reusable tool for label-scarce domains.

3) **Actionable Guidelines:** Through systematic experimentation, we derive evidence-based guidelines: density-based methods (DBSCAN) for exploratory outlier discovery, and centroid/probabilistic methods (KMeans, GMM) for stable, interpretable categorization.

The remainder of this paper is organized as follows.Section II reviews related work.Section III details the methodology, including the dataset, algorithms, and evaluation metrics.Section IV presents the experimental results.Section V discusses findings and practical guidelines, and Section VI concludes the paper.

## II. Related Work

Clustering algorithms are extensively applied in healthcare and nutrition for dietary pattern discovery and consumer segmentation [4, 22]. In unlabeled, high-dimensional nutritional domains, unsupervised clustering naturally groups foods by compositional similarity [3]. However, because algorithmic choices fundamentally shape the discovered patterns, understanding each paradigm's strengths and limitations is essential.

KMeans is widely adopted for its scalability and interpretability [5, 6]. It minimizes intra-cluster variance to produce compact, spherical clusters whose centroids represent average nutritional profiles. However, this spherical assumption limits its efficacy on irregular real-world distributions, and predefining the number of clusters ($k$) introduces subjectivity [13]. Conversely, agglomerative hierarchical clustering circumvents a predefined $k$ by constructing dendrograms for multi-level exploration [16], though it incurs higher computational costs and is sensitive to linkage criteria.

Density-based methods like DBSCAN detect arbitrarily shaped clusters and explicitly identify noise [8]. This is crucial for nutritional data containing atypical items (e.g., lean salads alongside calorie-dense burgers) that defy standard categorization [14]. While DBSCAN requires no predefined $k$, it is highly sensitive to density parameters. Alternatively, Gaussian Mixture Models (GMM) offer a probabilistic approach, modeling data as mixed Gaussian distributions to provide soft assignments [19]. This flexibility effectively captures overlapping nutritional profiles, such as foods exhibiting both high-fat and high-sugar traits.

Graph-based spectral clustering constructs similarity matrices and performs eigen-decomposition to detect complex, non-convex structures [10, 23]. However, these methods are computationally demanding and lack robustness on noisy nutritional data. Mean Shift, utilizing kernel density estimation, iteratively locates density peaks without requiring a predefined $k$ [21]. Recent advances aim to improve its scalability via bandwidth optimization and neighborhood pruning [24].

Despite this methodological diversity, the literature remains fragmented. Studies typically optimize a single algorithm for a specific dataset, and systematic cross-paradigm comparisons under controlled conditions are rare [14]. In the nutritional domain, no prior work has benchmarked centroid, hierarchical, probabilistic, density, graph, and kernel-based methods on a common dataset using consistent preprocessing and multi-dimensional evaluation. This study addresses this critical gap, providing evidence-based guidance for algorithm selection through a systematic evaluation of statistical quality, structural interpretability, and practical applicability.

## III. Methodology

This section describes the dataset, preprocessing pipeline, clustering algorithms, and evaluation metrics employed in this study. The overall framework follows a standard unsupervised learning pipeline, where raw nutritional data is transformed into a structured feature space and then analyzed using multiple clustering techniques to extract hidden patterns. Unsupervised learning is particularly suitable for this task because no ground-truth labels exist for defining "healthy" or "unhealthy" food clusters, making clustering a natural fit for discovering structure in nutritional data [3].

### A. Dataset

The dataset used in this study is sourced from Kaggle [25] and contains nutritional information for fast food items including calories, protein, fat, carbohydrates, sodium, sugar, cholesterol, and selected vitamins. Food nutrition datasets have been widely used in machine learning applications for health analytics and dietary recommendation systems due to their structured numerical features and direct relation to human health outcomes [4, 26]. Each food item represents a high-dimensional data point where each dimension corresponds to a nutritional attribute, transforming the dataset into a multi-dimensional clustering problem where similarity is determined by nutritional composition rather than categorical labels.

### B. Data Preprocessing

Data preprocessing is a critical step in clustering tasks because distance-based algorithms are highly sensitive to scale and missing values. Missing values in nutritional attributes such as protein, fiber, vitamin A, vitamin C, and calcium were handled using mean imputation. This approach preserves dataset size while reducing bias introduced by missing entries. Irrelevant or redundant features were removed to reduce noise and computational complexity, as irrelevant dimensions can distort distance metrics and degrade cluster quality [22].

*1) Feature Scaling:* Since clustering algorithms rely heavily on distance calculations, feature scaling is essential. All numerical features were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where $\mu$ is the mean and $\sigma$ is the standard deviation of each feature. This ensures that attributes with large numeric ranges (e.g., calories) do not dominate smaller-scale attributes (e.g., vitamins). Without normalization, algorithms such as KMeans and DBSCAN would produce biased cluster assignments. Standardization is commonly used in clustering-based food analysis and consumer segmentation tasks [27].

### C. Clustering Algorithms

Clustering is the process of grouping similar data points based on a similarity or distance metric. In this study, six clustering algorithms spanning centroid-based, hierarchical, probabilistic, density-based, graph-based, and kernel-based paradigms are evaluated to ensure a comprehensive comparison under identical dataset conditions.

*1) KMeans Clustering:* KMeans is a partition-based clustering algorithm that minimizes intra-cluster variance by iteratively assigning points to the nearest centroid [5]. The objective function is defined as:

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

where $C_i$ represents cluster $i$ and $\mu_i$ is the centroid of cluster $i$. KMeans is widely used due to its simplicity and scalability; however, it assumes spherical cluster shapes and requires the number of clusters to be predefined [13]. Recent improvements focus on better initialization strategies and faster convergence [15].

*2) Agglomerative Clustering:* Agglomerative clustering is a hierarchical method that builds a tree-like structure (dendrogram) by iteratively merging the closest clusters [16]. Unlike KMeans, it does not require specifying the number of clusters in advance. Instead, clusters are formed based on linkage criteria such as single linkage (minimum distance), complete linkage (maximum distance), or Ward linkage (variance minimization). While useful for exploring multi-level relationships in data, agglomerative clustering is computationally expensive for large datasets [16].

*3) Gaussian Mixture Model (GMM):* Gaussian Mixture Models are probabilistic clustering algorithms that assume data is generated from a mixture of Gaussian distributions [19]. Unlike KMeans, which assigns hard cluster labels, GMM provides soft clustering probabilities, allowing each data point to belong to multiple clusters with different probabilities. The probability density function is defined as:

$$P(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k) \tag{3}$$

where $\pi_k$ represents mixture weights. GMM is particularly effective for datasets with overlapping clusters and non-spherical distributions.

*4) DBSCAN:* DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups data points based on density rather than distance to centroids [8]. It requires two parameters: $\epsilon$ (neighborhood radius) and MinPts (minimum number of points). DBSCAN classifies points into core points, border points, and noise points, making it highly effective for detecting outliers in nutritional data where some food items may significantly deviate from typical dietary patterns. Unlike KMeans, DBSCAN does not require the number of clusters to be predefined.

*5) Spectral Clustering:* Spectral clustering is a graph-based method that transforms data into a similarity graph and uses eigenvector decomposition to identify clusters [10]. It involves constructing a similarity matrix, computing the Laplacian matrix, performing eigen decomposition, and applying KMeans in the reduced space. Spectral clustering is powerful for detecting complex, non-convex cluster structures but suffers from high computational complexity, especially for large datasets.

*6) Mean Shift Clustering:* Mean Shift is a non-parametric clustering algorithm that identifies clusters by locating density peaks in the feature space [21]. It works by iteratively shifting data points toward regions of higher density using kernel density estimation. Unlike KMeans or GMM, Mean Shift does not require the number of clusters to be specified in advance. However, it is computationally expensive, particularly for high-dimensional datasets. Recent research has focused on optimizing Mean Shift using acceleration techniques such as kernel approximation and neighborhood pruning [24].

### D. Evaluation Metrics

Assessing clustering quality in the absence of ground-truth labels requires a multi-dimensional approach, as no single internal metric can fully capture all aspects of cluster validity [17]. We therefore employ three complementary evaluation strategies.

**Silhouette Score.** The Silhouette Score measures how similar each data point is to its own cluster compared to other clusters, providing a value between $-1$ and 1. Higher scores indicate better intra-cluster cohesion and inter-cluster separation. It is defined for each sample as $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$, where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance. The overall Silhouette Score is the average across all samples.

**PCA Visualization.** We project the high-dimensional nutritional feature space onto the first two principal components to enable qualitative inspection of cluster geometry. This visualization allows us to assess whether clusters are compact, overlapping, irregular, or well-separated in a human-interpretable 2D space, complementing the purely numerical Silhouette Score.

**CVS.** Because the Silhouette Score primarily reflects separation geometry, we introduce CVS that jointly accounts for cluster purity (the degree to which a cluster consists of nutritionally homogeneous items) and separation consistency (the reliability with which different clusters are distinguished). CVS aggregates these properties into a normalized measure ranging from 0% to 100%, where higher values indicate cleaner, more structurally coherent groupings. Importantly, the CVS does not rely on external labels and is computed internally from the feature space, making it suitable for unsupervised nutritional profiling tasks.

Together, these three metrics provide a holistic assessment of clustering performance: statistical quality (Silhouette), interpretable geometry (PCA), and structural coherence (CVS).

## IV. Results

In reporting clustering performance, we adopt CVS defined in Section III-D, which aggregates cluster purity and sep-

TABLE I Comparison of Clustering Algorithms

| Algorithm | Silhouette Score | Clusters | Noise | CVS (%) |
|---|---|---|---|---|
| KMeans | 0.2700 | 4 | 0 | 40.89 |
| GMM | 0.2788 | 4 | 0 | 42.61 |
| Agglomerative | 0.2259 | 4 | 0 | 32.26 |
| Spectral | 0.0612 | 4 | 0 | 0 |
| Mean Shift | 0.3306 | 16 | 0 | 52.77 |
| DBSCAN | 0.5718 | 2 | 23 | 100 |

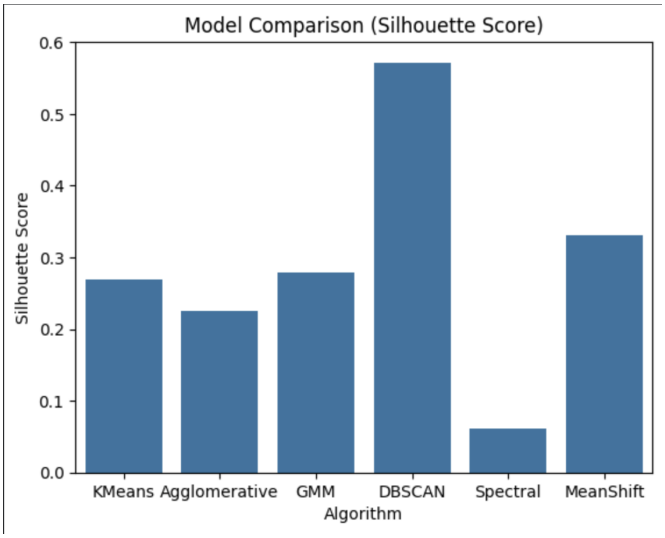


Fig. 1. Silhouette Score comparison of clustering algorithms.

aration consistency without relying on external labels. The performance of all clustering algorithms is summarized in Table I. The evaluation is based on Silhouette Score, number of clusters formed, noise points identified, and CVS. These metrics collectively provide a multi-dimensional view of clustering effectiveness, consistent with evaluation practices in unsupervised learning studies [14, 27].

The results demonstrate significant variation in clustering behavior across algorithms due to their differing mathematical assumptions. Partition-based methods such as KMeans and GMM assume spherical or Gaussian-shaped clusters, which limits their ability to capture irregular nutritional distributions in fast food datasets [5, 22]. In contrast, density-based and kernel-based methods such as DBSCAN and Mean Shift adapt better to non-linear structures, which are common in real-world nutritional feature spaces.

Among all methods, DBSCAN achieves the highest Silhouette Score (0.5718), indicating strong intra-cluster cohesion and inter-cluster separation. This aligns with findings from Kulkarni et al. [8], who highlight DBSCAN's effectiveness in identifying arbitrarily shaped clusters and filtering noise in high-dimensional datasets. However, the presence of 23 noise points suggests that DBSCAN is highly sensitive to local density variations, which may exclude borderline food items with mixed nutritional profiles.

Figure 1 illustrates the Silhouette Score comparison across all clustering algorithms. DBSCAN significantly outperforms other methods, indicating that density-based clustering is highly effective for separating fast food items into distinct

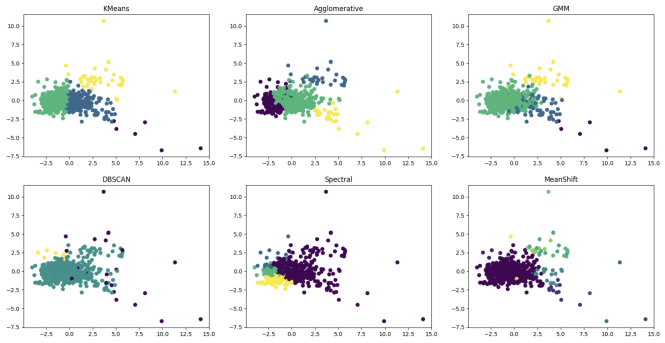


Fig. 2. PCA visualization of clustering results.

nutritional groups. Mean Shift follows as the second-best performer, benefiting from its kernel density estimation approach, which allows it to detect multiple local maxima in the data distribution [24].

KMeans and GMM show moderate performance, reflecting their reliance on global optimization assumptions rather than local density structure. Spectral clustering performs poorly due to sensitivity in similarity matrix construction and eigenvector selection, a limitation widely documented in graph-based clustering literature [10].

Figure 2 presents a two-dimensional PCA projection of clustering outputs, illustrating how each algorithm partitions the nutritional feature space. KMeans produces compact and spherical clusters due to its centroid minimization objective, which enforces uniform cluster geometry. This behavior is consistent with classical partition-based clustering theory [16].

In contrast, DBSCAN forms irregular and density-driven clusters that better reflect real-world nutritional variability. Mean Shift shows multiple adaptive cluster centers, indicating fine-grained segmentation of food categories. These observations align with Liakos et al. [4], who emphasize that food-related datasets often require flexible clustering models capable of capturing heterogeneous feature distributions rather than rigid geometric assumptions.

Figure 3 compares CVS across all methods. DBSCAN achieves the highest CVS, indicating superior cluster purity and structural separation. This result supports findings from Wulandari et al. [14], who show that DBSCAN consistently outperforms partition-based clustering methods in datasets with irregular distributions.

However, CVS alone does not fully reflect clustering quality. While DBSCAN achieves the highest possible CVS in this experiment, it also produces noise points, which may reduce interpretability in downstream applications such as dietary recommendation systems. In contrast, KMeans and GMM provide more balanced clustering structures, making them more suitable for applications requiring stable and interpretable groupings.

Figure 4 provides a comprehensive visualization of clustering performance across multiple metrics. It clearly shows that performance varies significantly depending on the evaluation criterion. DBSCAN dominates in Silhouette Score and CVS, while Mean Shift performs strongly in cluster granularity. As highlighted in Borlea et al. [13], cluster quality depends not

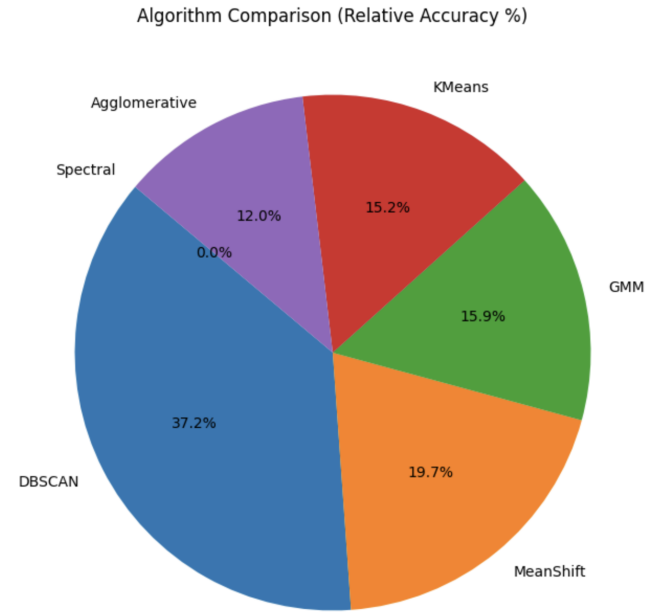


Fig. 3. CVS comparison across algorithms.

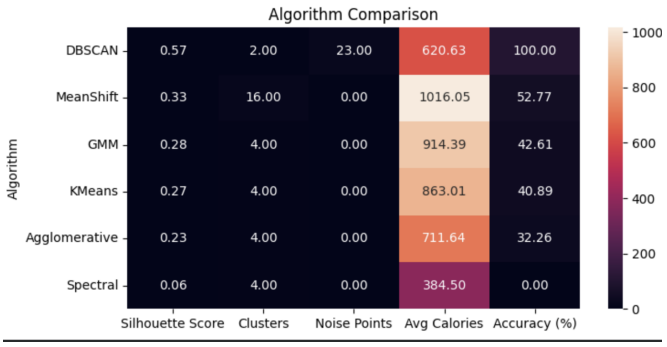


Fig. 4. Heatmap of clustering performance metrics.

only on separation metrics but also on post-processing and interpretability considerations. Similarly, Zhao [15] emphasizes that parameter sensitivity plays a major role in clustering stability, especially for methods like KMeans.

## V. Discussion

Experimental results demonstrate that no single algorithm universally dominates across all metrics; instead, selection must align with specific analytical goals based on underlying mathematical assumptions.

### A. Structural Discovery and Interpretability

DBSCAN achieves the highest Silhouette Score (0.5718) and maximum CVS, confirming its efficacy in detecting arbitrarily shaped clusters [8]. However, identifying 23 noise points presents a practical trade-off: while valuable for flagging atypical foods (e.g., extreme sugar or calorie densities), excluding these items complicates downstream applications like dietary recommendation systems that require nuanced guidance for all inputs.

Conversely, KMeans and Gaussian Mixture Models (GMM) produce balanced, fully partitioned clusters with moderate separation scores. By assigning every item to a cluster, they yield interpretable groupings whose centroids directly translate into average nutritional profiles. As noted in [27], KMeans performs reliably on normalized, spherical data. For applications prioritizing exhaustive categorization (e.g., menu labeling or consumer apps), these centroid-based and probabilistic methods offer clear advantages.

### B. Adaptive Methods and Computational Trade-offs

Mean Shift achieves the second-highest Silhouette Score (0.3306) and identifies 16 adaptive clusters without a predefined count, making it suitable for exploratory analysis [21]. However, its computational cost scales poorly, and bandwidth sensitivity can cause overly granular segmentation, though recent advances in neighborhood pruning may mitigate this [24].

Spectral clustering performed poorly (Silhouette 0.0612, CVS 0). These methods are highly sensitive to similarity graph construction and struggle with noisy, nonlinearly correlated nutritional data [10]. Effective application in this domain necessitates careful, domain-specific nutritional distance functions rather than standard similarity measures.

### C. Practical Algorithm Selection Guidelines

Taken together, our findings support the following evidence-based guidelines for nutritional data analysis:

1) **Exploratory pattern and outlier detection:** Density-based methods (DBSCAN) excel at revealing irregular structures and identifying atypical food items.
2) **Stable, interpretable categorization:** Centroid-based (KMeans) and probabilistic (GMM) methods assign every item to a cluster, yielding interpretable profiles suitable for exhaustive groupings.
3) **Adaptive segmentation (unknown $k$):** Mean Shift is a viable option, provided computational resources permit careful bandwidth tuning.
4) **Noisy or weakly structured data:** Spectral clustering should be applied cautiously, requiring domain-specific similarity matrix customization.

### D. Limitations

Despite promising results, several limitations must be acknowledged.

First, mean imputation for missing values may reduce statistical variance and distort nonlinear feature interactions (e.g., among fat, sugar, and calories) [4]; future work should employ advanced model-based imputation.

Second, several algorithms exhibit high sensitivity to parameter selection or structure: DBSCAN relies heavily on $\epsilon$ and MinPts, Mean Shift on bandwidth [24], and Spectral Clustering on graph construction [10]. This introduces instability in real-world deployments with irregular distributions.

Third, our dataset's restricted size limits scalability and the generalizability of findings to broader, large-scale food industry systems [13].

Finally, without ground-truth labels, internal metrics (Silhouette, CVS) cannot guarantee true semantic or dietary relevance [27]. External validation using expert-labeled benchmarks is necessary to strengthen evaluation reliability in future research.

## VI. Conclusion

This paper presented a multi-paradigm framework systematically evaluating six unsupervised clustering algorithms on fast food nutritional data. Utilizing a multi-dimensional protocol combining Silhouette Score, PCA, and a custom CVS, we comprehensively assessed clustering quality across statistical separation, geometric interpretability, and structural coherence.

Our results reveal a clear divergence among paradigms, reinforcing that algorithm selection must be purpose-driven. DBSCAN achieved the highest Silhouette Score (0.5718) and maximum CVS, demonstrating superior capability in detecting irregular clusters and isolating atypical foods (23 noise points). Conversely, KMeans and GMM provided fully partitioned, stable groupings whose centroids translate directly into interpretable nutritional profiles, making them ideal for exhaustive categorization tasks like menu profiling. Mean Shift offered adaptive segmentation (16 clusters) but proved computationally demanding and parameter-sensitive, while Spectral Clustering performed poorly without domain-specific similarity matrices.

Several limitations motivate future research directions. First, the potential statistical bias from mean imputation could be resolved using advanced KNN or deep learning-based techniques. Second, the parameter sensitivity of density- and kernel-based methods suggests a strong need for automated tuning, such as Bayesian optimization or genetic algorithms. Third, scaling this benchmarking framework to larger, diverse food databases and incorporating expert-labeled external validation will substantially strengthen evaluation reliability. Looking ahead, integrating deep clustering or hybrid neural network models represents a promising trajectory for advancing personalized dietary recommendations [26].

Ultimately, this systematic evaluation yields actionable algorithm selection guidelines, providing a robust and reusable methodological foundation for future research at the intersection of machine learning and nutritional data science.

## References


[1] M. Karimi, R. Rabiee, F. Hooshmand, B. Aghapour, M. Ahmadzadeh, E. Havaei, and K. Kazemi, "Consumption of fast foods and ultra-processed foods and breast cancer risk: A systematic review and meta-analysis," *Global Health Research and Policy*, vol. 10, no. 1, 2025.

[2] J. Wang, M. Zhou, X. Guo, and L. Qi, "Multiperiod asset allocation considering dynamic loss aversion behavior of investors," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 1, pp. 73–81, 2018.

[3] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review," *International Journal of Computing and Digital Systems*, 2023.

[4] K. G. Liakos, V. Athanasiadis, E. Bozinou, and S. I. Lalas, "Machine learning for quality control in the food industry: A review," *Foods*, vol. 14, no. 19, p. 3424, 2025.

[5] M. Suyal and S. Sharma, "A review on analysis of k-means clustering machine learning algorithm based on unsupervised learning," *Journal of Artificial Intelligence and Systems*, vol. 6, no. 1, pp. 85–95, 2024.

[6] S. Suraya, M. Sholeh, and U. Lestari, "Evaluation of data clustering accuracy using k-means algorithm," *International Journal of Multidisciplinary Approach Research and Science*, vol. 2, no. 1, pp. 385–396, 2023.

[7] Z. Zhang, S. Dai, C. Li, W. Wang, J. Parron, and E. Herrera, "Human-robot collaborative disassembly profit maximization via improved grey wolf optimizer," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 2, pp. 69–79, June 2025.

[8] O. Kulkarni and A. Burhanpurwala, "A survey of advancements in dbscan clustering algorithms for big data," in *Proc. IEEE PARC*, 2024, pp. 106–111.

[9] J. Gu, Z. Guo, J. Wang, L. Qi, S. Qin, and S. Qin, "Optimization of multi-factory remanufacturing processes with shared transportation resources using the alns algorithm," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 1, pp. 1–13, March 2025.

[10] L. Ding, C. Li, D. Jin, and S. Ding, "Survey of spectral clustering based on graph theory," *Pattern Recognition*, vol. 151, p. 110366, 2024.

[11] Z. Zhao and L. Zhou, "Multi-factory remanufacturing process optimization with discrete battle royale optimizer," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 3, pp. 129–131, September 2025.

[12] Q. Zhang, Y. Xing, C. Zhang, X. Sun, B. Hu, and A. Das, "Column generation algorithms for two-dimensional cutting problem with surface defects," *International Journal of Artificial Intelligence and Green Manufacturing*, vol. 1, no. 2, pp. 80–92, June 2025.

[13] I. D. Borlea, R. E. Precup, and A. B. Borlea, "Improvement of k-means cluster quality by post processing resulted clusters," in *Proc. ITQM*, ser. Procedia Computer Science, vol. 199, 2022, pp. 63–70.

[14] V. Wulandari *et al.*, "Comparison of dbscan, k-means and x-means algorithms on shopping trends data," *Indonesian Journal of Artificial Intelligence and Technology Systems (IJATIS)*, vol. 1, no. 1, pp. 1–8, 2024.

[15] H. Zhao, "Design and implementation of an improved k-means clustering algorithm," *Mobile Information Systems*, vol. 2022, pp. 1–10, 2022.

[16] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8219–8264, 2023.

[17] G. Vardakas, I. Papakostas, and A. Likas, "Deep clustering using the soft silhouette score: Towards compact and well-separated clusters," *Machine Learning*, vol. 115, no. 4, 2026.

[18] C. Xu, H. Wei, X. Guo, S. Liu, L. Qi, and Z. Zhao, "Human-robot collaboration multi-objective disassembly line balancing subject to task failure via multi-objective artificial bee colony algorithm," *IFAC-PapersOnLine*, vol. 53, no. 5, pp. 1–6, 2020.

[19] B. Zhao, X. Wen, and K. Han, "Learning semi-supervised gaussian mixture models for generalized category discovery," in *Proc. IEEE/CVF ICCV*, 2023, pp. 16 623–16 633.

[20] Z. Zhang, X. Guo, M. Zhou, S. Liu, and L. Qi, "Multi-objective discrete grey wolf optimizer for solving stochastic multi-objective disassembly sequencing and line balancing problem," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020, pp. 682–687.

[21] C. Cariou, S. L. Moan, and K. Chehdi, "A novel mean-shift algorithm for data clustering," *IEEE Access*, vol. 10, pp. 14 575–14 585, 2022.

[22] J. M. John, O. Shobayo, and B. Ogunleye, "An exploration of clustering algorithms for customer segmentation in the uk retail market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023.

[23] Y. Su *et al.*, "Nsckl: Normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification," *IEEE Transactions on Cybernetics*, vol. 53, no. 10, pp. 6649–6662, 2022.

[24] R. Mussabayev, A. Krassovitskiy, and M. Aristombayeva, "Optimizing the mean shift algorithm for efficient clustering," *Mathematics*, vol. 13, no. 21, p. 3408, 2025.

[25] U. Pedersen, "Fast food nutrition dataset," https://www.kaggle.com/datasets/ulrikthygepedersen/fastfood-nutrition, 2021, accessed: 2026-05-07.

[26] P. M. Lu and Z. Zhang, "Food nutrition feature modeling and personalized diet recommendation," *Journal of Computational Biology and Medicine*, 2025.

[27] S. Suraya, M. Sholeh, and U. Lestari, "Evaluation of data clustering accuracy using k-means algorithm," *International Journal of Multidisciplinary Approach Research and Science*, vol. 2, no. 1, pp. 385–396, 2023.