(Article)

A Computational Framework for Nutrient Density Assessment and Food Categorization

Rafael Julio Suseno[1], Kevin Matthew Siregar[2] and Devi Dwi Purwanto[3,*]

Received: 23-02-2026
Revised: 08-03-2026
Accepted: 27-03-2026
Published: 27-03-2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Correspondence: devi.dp@ukwms.ac.id

[1] Widya Mandala Surabaya Catholic University, Surabaya, Indonesia; rafael-j.inf24@ukwms.ac.id
[2] Widya Mandala Surabaya Catholic University, Surabaya, Indonesia; kevin-m.inf24@ukwms.ac.id
[3] Widya Mandala Surabaya Catholic University, Surabaya, Indonesia; devi.dp@ukwms.ac.id

Highlights

PCA-enhanced Agglomerative Clustering significantly outperforms DBSCAN in food nutrient density assessment, providing a superior computational framework for accurate nutritional categorization and precise data-driven dietary evaluation.

What are the main findings?

- This clustering process reveals that the combination of using the Agglomerative Clustering method and the PCA dimension reduction technique was more effective compared to the DBSCAN method.
- The Silhouette Scores and Calinski-Harabasz Index values obtained from the clustering process using the proposed method (0.41 and 85.97, respectively) are higher compared to the values obtained from the DBSCAN method (0.32 and 8.39, respectively), suggesting that the clusters are more distinct and separate.

What are the implications of the main findings?

- The research indicates that by using Agglomerative Clustering in conjunction with PCA, a more reliable way of clustering foods based on their nutritional content is achieved.
- This stronger clustering could be useful in personalized nutrition, where foods can be more easily classified for more precise dietary advice, and could even be used to refine food labelling with a better understanding of nutritional content.

Abstract

In this paper, we aim to explore the potential of clustering in creating a nutritional map of different foods based on the nutritional elements present in the food. We evaluated two clustering algorithms: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and Agglomerative Clustering. We used these algorithms to cluster the different foods in the Kaggle Food-101 dataset, which contains nutritional features such as proteins, carbohydrates, fats, and energy density. In order to enhance the efficiency of the clustering process and reduce the complexity of the data, we used the PCA (Principal Component Analysis) technique for data reduction. The Agglomerative Clustering technique with PCA demonstrated superior clustering quality compared to DBSCAN. This was based on the fact that the Agglomerative Clustering technique with PCA produced a higher Silhouette Score (0.41) and Calinski-Harabasz Index (85.97) than the DBSCAN technique. In our research, it was also found that the clusters produced by the Agglomerative Clustering technique with PCA could separate the different foods based on nutritional elements, which included high-protein foods, high-carb foods, and balanced diet foods.

Keywords: Agglomerative clustering, clustering algorithms; density-based spatial clustering of applications with noise (DBSCAN), nutrient density; principal component analysis (PCA).

1. Introduction

Computational nutrition is an increasingly important aspect of modern nutrition science, driven by the fact that the data on the composition of foods is very high-dimensional and difficult to understand. The traditional methods of dietary assessment, which are those methods that do not require complex categorizations, may not capture the broader nutritional aspects that more fully inform outcomes. Data-driven methods are recommended as a means of enhancing the ordering of the nutrient list and our understanding of the food profiles contained within the food composition databases, which are the foundation of nutritional assessment and public health policy formation [1], [2]. However, studies using modern computational methods, particularly for assessing nutrient density and clustering foods, are scarce. Machine learning, and particularly unsupervised clustering, is a means of identifying structural patterns within rich nutrition data sets, which may reveal non-linear and intricate relationships between nutritional variables, providing insights beyond those available through traditional statistical approaches [3].

In nutritional science, clustering can be used for grouping the patterns of diet, for studying the patterns of human nutrition, and for discovering the foods that have similar nutritional properties [4], [5]. In nutritional science, some of the clustering algorithms used are DBSCAN and Agglomerative Clustering. DBSCAN can be used for nutritional science because it does not require the number of clusters to be specified in advance, and it can also be used for the detection of noise points, which may not belong to any cluster due to different nutritional properties. In Agglomerative Clustering, the most similar items are joined based on a given distance, and a dendrogram is used to show the relationship of the food groups with the nutritional properties of the foods. This

type of clustering has been used for grouping foods based on nutritional properties in food composition data [6].

While several studies have already explored clustering, often combining statistics and ML to cluster foods or identify dietary patterns, head-to-head comparisons of the two algorithms, DBSCAN and Agglomerative Clustering, on the specific task of nutrient density evaluation are still scarce [7]. Emerging data-driven research on clustering promises to simplify the understanding of food groups based on their nutritional profiles, potentially filling in missing nutrient values based on cluster imputations. In the realm of precision nutrition, ML algorithms are cited as having the potential to improve nutrient predictions and tailor dietary recommendations, though the gap remains in the comparison of density-based clustering versus hierarchical clustering algorithms on the specific task of nutrient density evaluation [8].

The main innovation here is that it presents a comparison of DBSCAN and agglomerative clustering algorithms, but in the specific context of food nutrient data, with the intention of determining nutrient density. In addition, it incorporates dimension reduction techniques like Principal Component Analysis to make the clustering more effective. In essence, the research evaluates how the algorithms can identify clusters and outliers, linking them to the development of a Nutrient Density Index for precision nutrition, based on a data set of 101 foods that are commonly consumed. In other words, the research is advancing nutritional data analysis techniques and grounding the science of nutritional computing even further.

2. Materials and Methods

2.1. Data Collection

A public repository accessible through Kaggle, under the title "Computational Nutritional Science" by Alexander Clarke (2024), served as a source of nutritional data for this research. This data set was selected for its ability to provide a comprehensive overview of food composition, making it appropriate for computational analysis of nutritional content. It contains 101 unique food items, with each item describing nutritional content per 100 grams of food's edible portion. The number of retained PCA components (n=2) was determined by analyzing the variance ratio and scree plot to ensure maximum information retention while reducing noise. It covers a variety of foods, from baked goods to sea foods, meats, vegetables, desserts, and mixed dishes, ensuring a wide variety of foods for clustering beyond conventional food groups.

The data set is sufficiently broad in scope while still being deep in content, offering a wide variety of foods while maintaining a significant number of nutritional content variables to allow for clustering and evaluation of nutrient density. Table I lists the eight major nutritional content variables utilized in this research.

TABLE I

Dataset Nutritional Variables.

Variable	Unit	Description
Weight	grams (g)	Serving size
Calories	kcal	Total energy content
Protein	grams (g)	Total protein
Carbohydrates	grams (g)	Total carbohydrates

Variable	Unit	Description
Fats	grams (g)	Total fat
Fiber	grams (g)	Dietary fiber
Sugars	grams (g)	Total sugars
Sodium	milligrams (mg)	Sodium content

These variables were chosen based on their relevance to the evaluation of the nutritional quality of food products. In the cluster analysis, the positive variables, including protein, fiber, and fats, play a significant role in the evaluation of the quality of food products, while the negative variables, including sugar and salt, provide insight into the possible health hazards of consuming food products. The data was obtained from the Kaggle platform in CSV format, bearing the file name "nutrition.csv".

2.2. Data Preprocessing

The preprocessing of the data plays a significant role in ensuring that the data obtained from the cluster analysis is accurate, reliable, and understandable. In the study, the preprocessing of the data involved normalization, dimensionality reduction, and removal of outliers from the data set. In the normalization of the data, the features are normalized to ensure that all features are equally significant in the cluster analysis, especially when the features are measured in different units, e.g., protein measured in grams and energy measured in units of energy. In the study, z-score normalization was used, as shown in Equation 1.

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Through z-score normalization, all the variables such as protein, fat, carbohydrates, and sodium are equally weighted during the clustering. This avoids any dominance by variables with large scales. Principal Component Analysis (PCA) minimizes dimensionality by transforming correlated variables into uncorrelated principal components that hold maximum variance. PCA works by computing the covariance matrix and then determining its eigenvalues and eigenvectors. The eigenvectors hold the new axes, while the eigenvalues hold the variance. This transformation changes the original dataset X with a dimension of n x p into a new dataset Y, as shown in Equation 2.

$$Y = X.W \quad (2)$$

Principal Component Analysis (PCA) uses the eigenvectors and eigenvalues of the covariance matrix, which is obtained from the data set. The eigenvectors indicate the direction of the principal components, while the eigenvalues show the variance associated with each principal component. The number of principal components can also be determined using the variance ratio and the scree plot, where the optimal number of principal components can be determined. The use of PCA is critical in the reduction of data complexity, retaining the most significant features, and thus enhancing the analysis and visualization of the data, especially in clustering situations [9]. In the study, PCA was used to simplify the dimensionality of the nutritional data, retaining the most significant features of the nutritional profiles. The data was analyzed and clustered based on the food items, ensuring that the clustering process focuses on the most significant features and not the redundant data [10].

2.3. Data Clustering

In this research, two clustering algorithms, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and Agglomerative Clustering, were utilized to analyze the nutritional dataset. DBSCAN is a density-based clustering algorithm that clusters data points with high density, identifying data points with low density as noise. DBSCAN is a clustering algorithm that does not require any prior knowledge of cluster numbers. This clustering algorithm is appropriate for use with datasets that do not have a regular cluster structure. DBSCAN is a clustering algorithm that can be used to identify data points with unique features, which is useful in clustering nutritional data with varying features. DBSCAN can be used to identify data points with unique features, and this is useful in clustering nutritional data with varying features [11]. The parameters of DBSCAN, epsilon (ε) and min_samples, were optimized to ensure optimal clustering of data points.

Agglomerative clustering is a hierarchical algorithm that starts with every data point being its own cluster and combines them in pairs of the closest ones based on their Euclidean distance. This method is beneficial in explaining hierarchical relationships and analyzing data in different levels of granularity. This algorithm results in a dendrogram that explains data structures and makes it easier to identify patterns in nutritional data [12], [13].

2.4. Clustering Evaluation

In order to evaluate the quality of the clusters formed by DBSCAN and Agglomerative Clustering, two of the commonly used cluster evaluation criteria, the Silhouette Score and the Calinski-Harabasz Index, have been used. The Silhouette Score calculates the similarity of an object to its own cluster in comparison to other clusters. It is computed based on the following formula, as shown in Equation 3.

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (3)$$

A high value of the silhouette score indicates that the clusters are well separated and the data points are suitably assigned to the clusters. If the score tends towards -1, it may indicate that the points have been assigned to the wrong clusters. In addition, the Calinski-Harabasz Index, which is also known as the Variance Ratio Criterion, has been used for the evaluation of the quality of the clusters. This index can be computed based on the formula shown in Equation 4.

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \cdot \frac{n-k}{k-1} \quad (4)$$

2.5. Nutrient Density Index (NDI)

The Nutrient Density Index (NDI) defines a food item's nutritional value by considering both the positive and negative impacts of the food item's nutrients on health. The positive nutrients, including protein and dietary fibers, are associated with positive health impacts, while the negative nutrients, including sugars and sodium, are associated with health risks at higher concentrations. The Nutrient Density Index can be calculated using the formula provided by equation 5.

$$NDI = \frac{\sum Positive\ Nutrients - \sum Negative\ Nutrients}{Energy} \quad (5)$$

This approach follows the nutrient profiling approach developed by Drewnowski and Fulgoni, where foods are ranked based on the nutrient content in relation to the energy content in the food items [14]. The nutrient profiling concepts, including the Nutrient-Rich Food (NRF) Index, are

designed to score food items by rewarding nutrient-dense foods and penalizing food items that are energy-dense and nutrient-poor, including foods rich in sugars and sodium [15]. The approach of using the positive and negative impacts of the food items on health creates a balanced perspective on the Nutrient Density Index, where the benefits and risks are considered in a quantitative approach.

2.6. Desain Architecture

The methodology includes data preprocessing, normalization, dimensionality reduction using Principal Component Analysis, clustering using DBSCAN and Agglomerative Clustering, cluster quality assessment using the Silhouette Score and Calinski-Harabasz Index, and lastly, the Nutrient Density Index calculation to assess the nutritional value of the food item, considering positive and negative nutrient factors. This methodology helps to achieve a more detailed insight into the relationship between the clusters of food items and the nutritional value of the items.

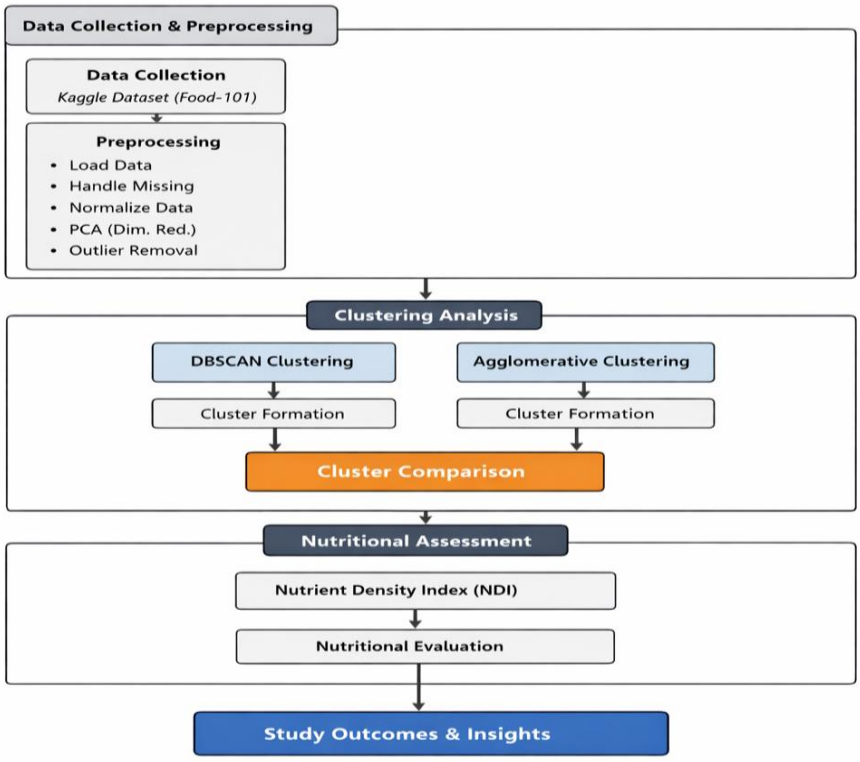

Fig. 1. Pipeline Nutritional Clustering and Assessment.

The study is commenced by data collection, where the Kaggle Food-101 dataset is used, which contains nutritional data for 101 different food items. The data is collected, and the data is preprocessed to ensure that the data is complete. Z-score normalization is used to standardize the values of each nutritional data item, reducing the impact of scale differences on the analysis. Next, the Principal Component Analysis is used to reduce the dimensionality of the data, making the data simpler while retaining the essential characteristics. Lastly, the data is processed to remove the data points that are identified as outliers, especially for the clustering algorithms, as the algorithms are sensitive to the data distribution.

In the Clustering Analysis, two different algorithms, namely DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and Agglomerative Clustering, were used for the clustering

of the data points. DBSCAN uses the density of the data points to form clusters and identify outliers. This technique does not require the number of clusters to be specified in advance, which suits the complex shapes of the data. Agglomerative Clustering uses a hierarchical technique where the clusters are iteratively combined based on the similarity of the data points. This technique can be used to generate a dendrogram for the data points. After the completion of the clustering analysis, the Silhouette Score and the Calinski-Harabasz Index were used for the evaluation of the clusters. This step provided a deeper understanding of the effectiveness of the clustering algorithms in grouping the foods based on the nutritional profiles.

In the last step of the framework, namely Nutritional Assessment, the results of the clustering analysis are combined with the Nutrient Density Index (NDI) for the assessment of the nutritional value of the clusters. In the Nutrient Density Index, the positive and negative nutrients of the foods are used to generate a score that reflects the nutritional value of the food based on the nutrient density in relation to the total calories in the food product. This step of the framework combines the nutritional value of the foods with the clustering results to provide a deeper understanding of the relationship between the clusters of foods and the nutritional value of the foods.

3. Results

3.1. Data Quality and Descriptive Statistics

The data set had no missing values, indicating that the data set is complete, as there are no missing values across the 505 data points and the eight nutritional variables. The data set is therefore ideal for analysis. The data set had some extreme values, as indicated by the outlier analysis, where the variable 'sugars' had the highest number of extreme values, at 54 (10.7%), while 'fats' had the second-highest, at 18 (3.6%), and 'protein' had the least, at 16 (3.2%).

TABLE II

Descriptive Statistics of Nutritional Variables (N=505).

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Weight (g)	227.2	102.6	50.0	150.0	200.0	300.0	700.0
Calories (kcal)	481.4	230.1	50.0	300.0	450.0	600.0	1260.0
Protein (g)	21.2	16.3	2.0	9.0	16.0	30.0	88.0
Carbohydrates (g)	41.9	30.6	0.0	18.0	38.0	60.0	150.0
Fats (g)	23.7	14.8	1.0	12.0	20.0	30.0	90.0
Fiber (g)	3.3	2.6	0.0	1.0	3.0	5.0	15.0
Sugars (g)	12.7	16.7	0.0	3.0	6.0	15.0	105.0
Sodium (mg)	595.8	373.1	10.0	300.0	600.0	875.0	1867.0

3.2. Correlation Structure

The correlation matrix showed that the nutritional variables are highly related to each other. The correlation between 'calories' and 'fats' was the highest, at 0.82, indicating that the variation in calories is primarily attributable to fat content. The correlation between 'weight' and 'sodium' is moderate, at 0.74, indicating that heavier food items tend to contain more sodium. The correlation between 'carbohydrates' and 'sugars' is moderate, at 0.53, indicating that higher carbohydrate content is associated with higher sugar content.

TABLE III

Pearson Correlation Matrix of Nutritional Variables. Lower triangle displayed.

Variable	Calories	Protein	Carbs	Fats	Fiber	Sugars	Sodium
Calories	1.00						
Protein	0.43	1.00					
Carbohydrates	0.41	-0.32	1.00				
Fats	0.82	0.48	-0.15	1.00			
Fiber	0.23	0.12	0.32	0.07	1.00		
Sugars	0.33	-0.21	0.53	0.06	-0.01	1.00	
Sodium	0.52	0.43	0.22	0.42	0.22	-0.04	1.00

3.3. Principal Component Analysis

PCA of seven nutritional variables yielded components with the following explained variance:

TABLE IV

Eigenvalues and Variance Explained by Principal Components for Nutritional Variables

Component	Eigenvalue	Variance (%)	Cumulative (%)
PC1	2.84	40.6	40.6
PC2	1.48	21.1	61.7
PC3	1.02	14.6	76.3
PC4	0.72	10.3	86.6
PC5	0.48	6.9	93.5
PC6	0.31	4.4	97.9
PC7	0.15	2.1	100.0

TABLE V

Principal Component Loadings (First Two Components).

Variable	PC1 Loading	PC2 Loading	PC1 Squared	PC2 Squared
Calories	0.85	-0.12	0.72	0.01
Protein	0.61	-0.51	0.37	0.26
Carbohydrates	0.10	0.84	0.01	0.71
Fats	0.79	-0.28	0.62	0.08
Fiber	0.28	0.34	0.08	0.12
Sugars	0.19	0.64	0.04	0.41
Sodium	0.62	0.15	0.38	0.02

Component interpretation:

- PC1 (Energy-Mass Dimension): This component has strong positive loadings for calories, fats, protein, and sodium. Food items with high PC1 scores (like desserts and processed meats) are relatively heavy in terms of energy content.
- PC2 (Carbohydrate-Sugar Dimension): This component also has strong positive loadings for carbohydrates and sugars, and a strong negative loading for protein. It distinguishes between food items that are heavy in carbohydrates and food items that are heavy in protein. The vectors for carbohydrates and sugars are almost parallel, suggesting a strong positive correlation between the two, and the protein vector points in the opposite direction.

3.4. Clustering Algorithm Comparison

Two algorithms were evaluated using the Silhouette Score and Calinski-Harabasz Index on the same standardized data (features: protein_pct, carb_pct, fat_pct, energy_density):

TABLE VI

Clustering Algorithm Performance Comparison.

Algorithm	Optimal Parameters	Silhouette Score	Calinski-Harabasz
DBSCAN	ε = 0.4, min_samples = 5	0.32	8.39
Agglomerative	n = 5	0.32	68.21
Agglomerative	N=5, PCA (n=2)	0.41	85.97

As shown in Table VI Agglomerative Clustering with PCA achieved the highest performance, where n = 2, as it has the highest Silhouette Score, 0.41, and the highest Calinski-Harabasz Index, 85.97. This means that the clusters obtained by applying the clustering operation on the transformed data, using PCA, are the most well-defined and meaningful, when compared to the other methods. Hence, it is deduced that the optimal clustering method is the Agglomerative Clustering with PCA, where n = 2. Following this, Fig. 2 shows the visualization of the clustering obtained, along with the dendrogram, providing more insights into the data.

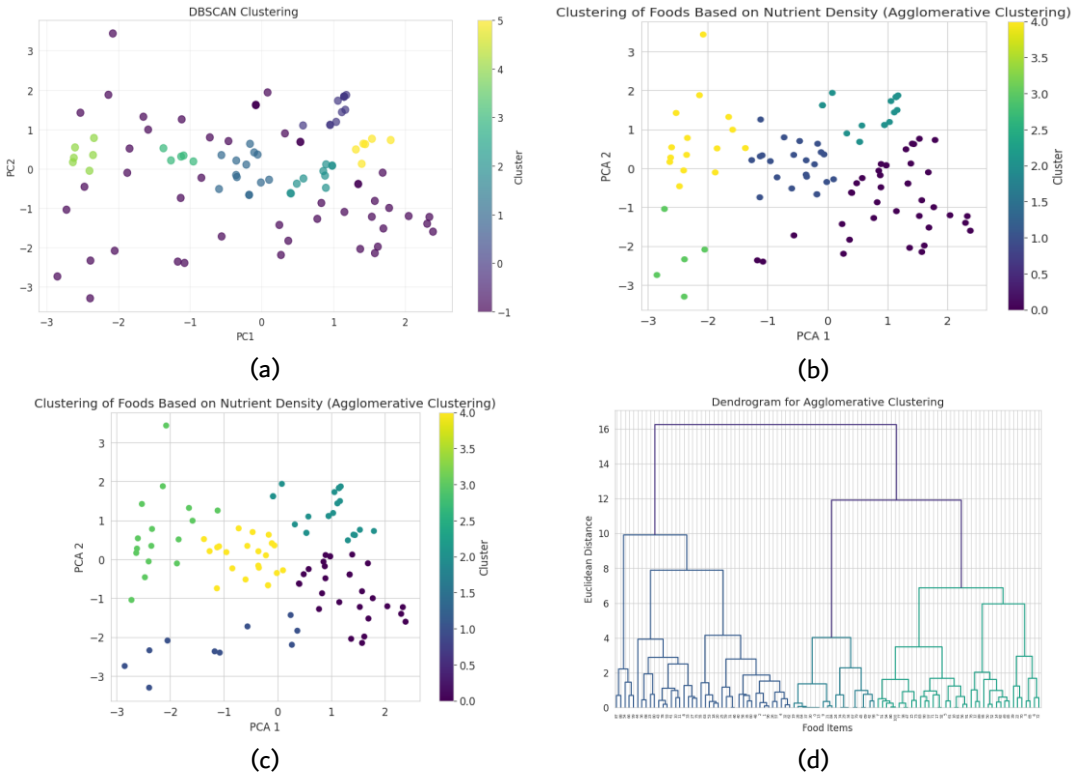

Fig. 2. Visualization and corresponding dendogram. (a) DBSCAN. (b) Agglomerative clustering. (c) Agglomerative clustering with PCA. (d) Dendogram agglomerative clustering.

4. Discussion

The results obtained from the clustering process indicate that the food items are grouped based on their nutritional content, as observed from the results of the DBSCAN algorithm and the Agglomerative Clustering algorithm. However, the performance of the algorithms differs based on various significant parameters. In the case of the DBSCAN algorithm, the moderate

performance of the algorithm was observed based on the Silhouette Score of 0.32, which indicates that the clusters are cohesive, but at the same time, the clusters are overlapping. This overlapping may be due to the fact that the DBSCAN algorithm cannot effectively identify clusters based on the variable density of the clusters, as the clusters are distributed in a complex manner. In addition, the fact that the number of clusters does not have to be defined in the DBSCAN algorithm may be considered a drawback of the algorithm, as the clusters may not be defined in a proper manner based on the structure of the data. However, the algorithm was effective in identifying outliers, as the outliers are considered noise in the algorithm.

On the other hand, this study used Agglomerative Clustering, especially when combined with PCA for dimensionality reduction, more differentiated clusters were formed. This is further confirmed by the Silhouette Score of 0.41 for the Agglomerative Clustering combined with PCA, which shows that more differentiated clustering is formed in contrast with DBSCAN. This may be due to the nature of Agglomerative Clustering, where data points are grouped based on their similarity. In addition, PCA dimensionality reduction helped in creating more differentiated clusters by simplifying the data and highlighting more important nutritional information. Not only were more differentiated clusters formed by using Agglomerative Clustering combined with PCA, but more differentiated clusters were also formed in terms of their separation from one another, as confirmed by the higher value of the Calinski-Harabasz Index of 85.97. This shows that more differentiated and separate clusters are formed when using Agglomerative Clustering combined with PCA for clustering food items based on their nutritional content.

The results show that PCA assisted in the clustering of the food items based on nutrient density, as depicted in the observed clustering characteristics. Upon the application of PCA, the formed clusters showed a high degree of correspondence with the expected food categories, for example, high-protein-containing foods and high-fat-containing foods, as opposed to the situation where PCA was not used, in which the observed clustering showed a high degree of overlap. This implies that the PCA reduced the irrelevant features of the data, thus enabling the clustering function to focus on the main nutritional features of the food items. Therefore, the PCA function improved the quality of the clusters produced by the Agglomerative Clustering function.

The implications of the present study's results are significant in relation to the upcoming studies on clustering algorithms concerning nutritional data analysis. One of the most important conclusions of the present study was the benefit of utilizing dimension reduction techniques, like principal component analysis (PCA), to improve the clustering algorithm's performance, especially concerning high-dimensional datasets. In fact, agglomerative clustering, along with PCA, presented better clustering performance, as shown by the presence of better-defined clusters and higher evaluation metrics. Future studies may consider other dimension reduction techniques, like t-SNE or UMAP, that may present better results concerning the clustering algorithm's performance due to their ability to better represent non-linear relationships between features.

Another area that can be explored in the future involves the use of hyperparameter optimization for clustering algorithms. In the current study, the selection of the hyperparameters for the DBSCAN and the Agglomerative Clustering algorithm was based on conventional heuristics; however, more advanced methods, such as grid search or randomized search, can also be used to

find the optimal values for the hyperparameters, including the values for the parameters epsilon (ε), min_samples, and n_clusters. Hyperparameter optimization can lead to better clustering results, especially when dealing with heterogeneous data that may require the use of finely tuned hyperparameters to accurately reflect the data.

Additionally, another area that can be explored in the future involves the interpretation of the clusters from a nutritional perspective, including the assessment of the link between the nutritional content of the food items within the clusters and the health impacts of the food items within the clusters. While the current study was focused on the clustering of food items based on the nutritional features, future studies can involve the assessment of the link between the clusters and the health impacts, including the food classification and health impacts, and the health and dietary recommendations, which can lead to deeper insights into the use of clustering in the context of health and nutrition.

Although the results of the study are promising, several limitations of the study are considered in depth. First, the data set used in the study had only 101 food items, which may not be enough to represent the wide diversity of global food intake. This may potentially affect the results of the clustering because some of the food items may be underrepresented. Therefore, to ensure the accuracy of the results, a more diverse data set representing a wide range of food items should be considered in the future.

The second limitation of the study may be the nutritional data used, which may not represent the wide range of factors that may affect a food's profile. For example, the mode of food preparation may differ from one region to another, as may the amount of food consumed per person. In other instances, the nutritional value of the food consumed may differ due to the ingredients used in the food's preparation. In the study, the nutritional data of the food was based on the average nutritional values per 100 grams of the respective food items. Therefore, to ensure a more accurate representation of the food profile, a more diverse data set representing a wide range of food profiles, including the mode of food preparation, amount of food consumed, and nutritional value of the food consumed, should be considered in the future.

Lastly, while PCA is a very powerful technique for dimensionality reduction, it also has a limitation. This limitation is that PCA can only be effectively used when the data has a linear relationship. This implies that when the data has complex relationships that are non-linear, PCA may not be able to capture all the relationships. To achieve this, future research could focus on using other advanced techniques such as t-distributed stochastic neighbor embedding and autoencoders. These techniques can be used for dimensionality reduction. This could result in a more effective clustering outcome.

Despite inherent limitations, the empirical findings demonstrate substantial utility across dietary assessment, food product formulation, nutritional epidemiology, and public health communication. Within dietary assessment frameworks, the Nutritional Dietary Index (NDI) functions as a scalable metric of food quality easily integrated into consumer-facing applications, offering an intuitive 0-to-100 continuum that circumvents the complex interpretive burdens characteristic of traditional nutrient profiling systems. Computational automation via K-means clustering further enhances this process by systematically labeling food groups, thereby generating granular, data-driven insights superior to conventional classification methods. Moreover, this

cluster analysis serves as a diagnostic tool in food product development to detect nutritional lacunae, wherein the conspicuous absence of a high-protein, high-fiber, and moderate-energy-density cluster signals a definitive market opportunity for functional food innovation. In the domain of nutritional epidemiology, these empirically derived cohorts establish an objective alternative for exposure classification, enabling rigorous investigations into the longitudinal associations between dietary patterns and chronic disease pathogenesis without the constraints of a priori food categorization. Ultimately, the stark delineations between high-NDI and low-NDI food profiles optimize public health communication by facilitating streamlined, actionable dietary guidelines, proving that behavioral shifts can be successfully incentivized without necessitating comprehensive micronutrient literacy among target populations.

5. Conclusion

Based on the findings of the current research, it was established that the Agglomerative Clustering technique with PCA outperformed the DBSCAN and the Agglomerative Clustering technique without PCA in all cases. In particular, the technique that included PCA produced the best results in the clustering of the data, as shown in the high Silhouette Score of 0.41 and the high Calinski-Harabasz Index of 85.97. In fact, the clusters produced in the current research using the technique that included PCA were much more distinguishable than those produced in the other two techniques. Much of the success of the technique in the current research can be attributed to the PCA technique, which reduced the dimensionality of the data. By reducing the complexity of the data, the technique was much more successful in analyzing the nutritional features of the food items.

In isolation, the performance of the DBSCAN algorithm is moderate, as evidenced by the Silhouette Score of 0.32. This indicates that there is some level of overlap between the formed clusters. Although the performance of the DBSCAN algorithm is moderate, identifying the outliers as noise points is reasonable. However, the algorithm faces problems in creating clear boundaries between the clusters, especially considering the complex and changing densities of the data. Agglomerative Clustering, on the other hand, performed similarly to the DBSCAN algorithm, having a Silhouette Score of 0.32 and a Calinski-Harabasz Index of 68.21, making it slightly lower than the performance of the PCA-based version of the Agglomerative Clustering algorithm.

In summary, the findings of this experiment show that the Agglomerative Clustering algorithm using PCA is the best approach to cluster the food items based on nutritional values. Aside from the performance, the PCA version of the Agglomerative Clustering algorithm is more interpretable, making the findings more useful for real-world applications, such as personalized nutrition planning and food item classification. Future studies on this topic may include using more efficient algorithms, tuning the hyperparameters, and using more data to validate the findings.

References

[1] G. De Bhowmick, B. Guieysse, D. W. Everett, M. G. Reis, and C. Thum, "Novel source of microalgal lipids for infant formula," *Trends Food Sci. Technol.*, vol. 135, pp. 1–13, May 2023, doi: 10.1016/j.tifs.2023.03.012.

[2] H. Shin and H. Seo, "Nutrient profile-based food categorization and group-wise missing data imputation for commercial food composition database," *J. Food Compos. Anal.*, vol. 150, p. 108828, Feb. 2026, doi: 10.1016/j.jfca.2025.108828.

[3] W. Quan, J. Zhou, J. Wang, J. Huang, and L. Du, "Machine Learning-Driven Precision Nutrition: A Paradigm Evolution in Dietary

Assessment and Intervention," *Nutrients*, vol. 18, no. 1, p. 45, Dec. 2025, doi: 10.3390/nu18010045.

[4] Y. Balakrishna, S. Manda, H. Mwambi, and A. van Graan, "Determining classes of food items for health requirements and nutrition guidelines using Gaussian mixture models," *Front. Nutr.*, vol. 10, Oct. 2023, doi: 10.3389/fnut.2023.1186221.

[5] M. M. Medina-Vadora *et al.*, "A Clustering Study of Dietary Patterns and Physical Activity among Workers of the Uruguayan State Electrical Company," *Nutrients*, vol. 16, no. 2, p. 304, Jan. 2024, doi: 10.3390/nu16020304.

[6] E. A. F. da Silva Torres, M. L. Garbelotti, and J. M. Moita Neto, "The application of hierarchical clusters analysis to the study of the composition of foods," *Food Chem.*, vol. 99, no. 3, pp. 622–629, 2006, doi: 10.1016/j.foodchem.2005.08.032.

[7] Y. Balakrishna, S. Manda, H. Mwambi, and A. van Graan, "Statistical Methods for the Analysis of Food Composition Databases: A Review," *Nutrients*, vol. 14, no. 11, p. 2193, May 2022, doi: 10.3390/nu14112193.

[8] D. Tsolakidis, L. P. Gymnopoulos, and K. Dimitropoulos, "Artificial Intelligence and Machine Learning Technologies for Personalized Nutrition: A Review," *Informatics*, vol. 11, no. 3, p. 62, Aug. 2024, doi: 10.3390/informatics11030062.

[9] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective," *Trends Food Sci. Technol.*, vol. 72, pp. 83–90, Feb. 2018, doi: 10.1016/j.tifs.2017.12.006.

[10] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.

[11] F. Nurulhikmah and D. N. E. Abdi, "Classification of Foods Based on Nutritional Content Using K-Means and DBSCAN Clustering Methods," *Teknika*, vol. 13, no. 3, pp. 481–486, Oct. 2024, doi: 10.34148/teknika.v13i3.1067.

[12] T. Dinh *et al.*, "Data clustering: a fundamental method in data science and management," *Data Sci. Manag.*, Aug. 2025, doi: 10.1016/j.dsm.2025.08.001.

[13] H. S. Al Jauhar, S. Solimun, and R. Fitriani, "Application Of DBScan For Clustering Society Based On Waste Management Behavior," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 19, no. 2, pp. 961–972, Apr. 2025, doi: 10.30598/barekengvol19iss2pp961-972.

[14] R. Watanabe, H. Ashida, M. Kobayashi-Miura, A. Yokota, and J. Yodoi, "Effect of chronic administration with human thioredoxin-1 transplastomic lettuce on diabetic mice," *Food Sci. Nutr.*, vol. 9, no. 8, pp. 4232–4242, Aug. 2021, doi: 10.1002/fsn3.2391.

[15] D. S. Ludwig, "The Ketogenic Diet: Evidence for Optimism but High-Quality Research Needed," *J. Nutr.*, vol. 150, no. 6, pp. 1354–1359, Jun. 2020, doi: 10.1093/jn/nxz308.