Journal of Applied Engineering and Technological Science
Vol 7(2) 2026: 1561-1579

DEEP EMBEDDED CLUSTERING FOR INDONESIAN PROTEIN, FAT, AND ENERGY AVAILABILITY DATA

Zakha Maisat Eka Darmawan[1], Oktavia Citra Resmi Rachmawati[2*], Ashafidz Fauzan Dianta[3], Kholid Fathoni[4], Rizky Yuniar Hakkun[5], Tri Budi Santoso[6], Kevin Ilham Apriandy[7]
Department of Creative Multimedia Technology, The Electronic Engineering Polytechnic Institute of Surabaya, Surabaya, Indonesia[1,3,4,5,6]
Study Program of Computer Technology, Politeknik Internasional Tamansiswa Mojokerto, Indonesia[2,7]
citra.research@gmail.com[2*]

Received: 10 August 2025, Revised: 04 March 2026, Accepted: 04 April 2026
**Corresponding Author*

ABSTRACT
Understanding disparities in regional food availability is crucial for food security policies. Most previous studies on Indonesian food availability use conventional clustering methods. These methods operate directly on the feature space and may miss complex, non-linear relationships in nutritional data. This limitation highlights the need for advanced analytical approaches to uncover deeper patterns. This study analyzes patterns of provincial food availability in Indonesia using Deep Embedded Clustering (DEC). It uses per capita indicators of energy, fat, and protein from both plant and animal sources, as well as the 2023 Food Consumption Pattern (FCP) score. DEC integrates representation learning with clustering. This allows the model to capture latent structures and nonlinear relationships that traditional clustering cannot identify. The analysis began by comparing K-Means and Hierarchical Clustering using the silhouette score to generate pseudo-labels for the DEC model. Hierarchical Clustering with Ward linkage and Euclidean distance achieved the highest silhouette score (0.3958) and was used for pseudo-label generation. Two DEC configurations were implemented, showing improved clustering performance. These achieved silhouette scores of 0.7829 (DEC-1) and 0.6385 (DEC-2). The results reveal four distinct clusters of Indonesian provinces, each with different food availability characteristics. These range from balanced, nutrient-rich regions to provinces with more limited or specific nutritional patterns. The findings show that DEC can capture complex structures in nutritional data. It produces more meaningful clusters than conventional approaches. In practice, the identified clusters provide policymakers, nutrition experts, and the food industry with useful insights for region-specific strategies. These strategies can improve food security and nutritional balance. Theoretically, this study contributes to the use of deep learning-based clustering in food availability analysis. It is especially relevant in national food security research. Future research may extend this approach by integrating time-series data and spatial analysis. This will help understand the temporal and regional dynamics of food availability in Indonesia.
Keywords : *Deep Learning, Auto Encoder, Hierarchical Clustering, Data Analysis, Deep Embedded Clustering*

1. Introduction

Food security has emerged as a global issue, with numerous nations endeavoring to ensure sufficient and nutritious food resources for their populations (Ahmad et al., 2024). The global commitment articulated in the Millennium Development Goals (MDGs) underscores the imperative to diminish poverty and hunger by 2015 (Sutardi et al., 2022). Notwithstanding these efforts, global food security remains an unresolved challenge. The worldwide Food Security Index (GFSI) reported that the average worldwide food security score in 2019 was 62.9 out of 100. In that year, Indonesia attained a score of 62.6, somewhat below the global average and inferior to many ASEAN nations, such as Malaysia (73.8) and Thailand (65.1), indicating that the nation's food security performance requires significant improvement (Rusmawati et al., 2023). Food security in Indonesia is a significant concern due to difficulties in meeting each person's dietary requirements for protein, fat, and energy (Nugroho et al., 2022). Data depicting per capita food availability in Indonesia, as evidenced by the available information. Data.go.id platform

provides critical insights into food consumption trends and potential enhancements to address the population's nutritional requirements (Ramadhan et al., 2025).

With technological progress, data analysis techniques are increasingly becoming complex (Hu & Szymczak, 2023). Machine learning, capable of analyzing extensive datasets and uncovering hidden patterns, offers an advantageous solution to this issue (Darmawan et al., 2025). Clustering algorithms, utilized to categorize data based on attribute similarities, have been implemented in diverse contexts (Murugan et al., 2024), including food management. Traditional clustering techniques, such as K-Means and hierarchical clustering, are predominantly dependent on distance-based similarity metrics and the assumption of linear separability, rendering them susceptible to initialization, noise, and high-dimensional feature spaces (Zhang & Parnell, 2023; Sahria et al., 2026). These constraints diminish their capacity to identify intricate, non-linear correlations present in multidimensional nutritional information, including those related to protein, fat, and per capita energy availability. The use of advanced clustering techniques, such as Deep Embedded Clustering (DEC), which integrates representation learning and clustering within a cohesive deep learning framework (L. Wang et al., 2023), has been infrequently applied to food availability, particularly for numeric data on protein, fat, and energy per capita. The main aim of this work is to investigate and implement the DEC approach for clustering food availability data, yielding fresh perspectives on food security in Indonesia.

While numerous studies have examined clustering in relation to food, few have applied deep learning to Indonesian food availability data. This work will distinguish itself with the application of comprehensive DEC methodologies to analyze protein, fat, and energy data. This approach surpasses earlier studies employing conventional clustering methods such as K-Means and K-Medoids (Julianto et al., 2022) or Fuzzy C-Means (Setiono & Dianto, 2022) by providing a more in-depth analysis and enhanced understanding of the factors affecting food distribution. Prior research predominantly utilized distance-based clustering in low-dimensional representations, such as aggregated food spending or restricted agricultural data, and assessed performance primarily through internal validity metrics. Although these methodologies offer beneficial regional classifications, they presuppose linear separability and rely significantly on predetermined feature spaces, potentially oversimplifying intricate patterns of nutritional availability. Nutritional datasets on protein, fat, and energy availability per capita are inherently multidimensional and potentially non-linear, which limits the efficacy of conventional clustering approaches for identifying latent structural patterns.

This study significantly contributes to both data research and machine learning research. This study elucidates food availability patterns in Indonesia using comprehensive, precise data. This paper presents the use of Deep Embedded Clustering (DEC), a deep learning-based clustering technique, to overcome the shortcomings of traditional distance-based approaches. In contrast to conventional clustering methods that operate directly in the original feature space, DEC integrates representation learning and clustering within a cohesive framework, enabling the model to discern nonlinear latent structures from multidimensional data. Given that indicators of protein, fat, and energy availability may exhibit intricate interdependencies and hidden patterns across different areas and time periods, deep learning-based representation learning is crucial for extracting more distinctive features before clustering. Consequently, the application of DEC is warranted, as it improves clustering quality by identifying intrinsic data structures that may elude detection by superficial or solely distance-based methodologies, especially in extensive, high-dimensional food and nutrition datasets. Consequently, the results of this investigation are anticipated to yield more efficacious recommendations for food security policies and enhance the literature on machine learning methodologies, specifically regarding the use of DEC to intricate data.

This study seeks to establish and implement a Deep Embedded Clustering (DEC) framework to analyze multidimensional data on protein, fat, and energy availability in Indonesia, with the objective of uncovering latent structural patterns that traditional clustering methods fail to capture effectively. This research is innovative for combining deep representation learning and clustering for food security data analysis, advancing beyond traditional distance-based techniques to a nonlinear feature-learning method designed for intricate nutritional datasets. This study offers a comparative analysis of traditional clustering methods versus deep learning-based clustering

regarding national food availability data, thereby enhancing both the methodological and empirical dimensions of the literature on machine learning applications in food security analysis.

This document is organized as follows. The Introduction delineates the research context and rationale. The Literature Review examines K-Means and Hierarchical Clustering, elucidates Deep Embedded Clustering (DEC), analyzes food security data in prior studies, and contrasts traditional clustering methods with deep learning-based clustering techniques. The Research Methods section delineates techniques for data collection, data preparation, data analysis, modeling, and model evaluation. The Results and Discussion section presents the empirical findings and their implications, followed by the Conclusions section, which summarizes the principal contributions and recommendations.

2. Literature Review

Analysis of K-Means and Hierarchical Clustering

K-Means is a widely used partition-based clustering approach in unsupervised learning, known for its computational efficiency and simplicity. The technique divides data into k predetermined clusters by reducing within-cluster variation via recurrent centroid adjustments. The temporal complexity is often O(nkt), making it appropriate for large datasets. K-Means assumes spherical cluster structures and linear separability in the original feature space, which may not accurately reflect the inherent structure of intricate real-world data. The method is sensitive to centroid initialization and requires prior determination of the number of clusters, which may yield unstable or poor clustering outcomes, especially in high-dimensional datasets (Ennaouri & Zellou, 2024).

Hierarchical clustering provides an alternate method by creating hierarchical cluster structures via either agglomerative (bottom-up) or divisive (top-down) techniques. In contrast to K-Means, it does not necessitate a predetermined number of clusters and offers a dendrogram that facilitates multi-tiered analysis of data linkages. Previous comprehensive analyses of hierarchical clustering algorithms indicate that these methods facilitate flexible exploration of cluster structures; however, they often exhibit high computational complexity, ranging from $O(n^2)$ to $O(n^3)$, which limits scalability for large datasets. Moreover, hierarchical clustering is significantly influenced by distance or similarity metrics in the original feature space and is susceptible to the selection of linking criteria (Bui & Phan, 2023).

Although K-Means and Hierarchical Clustering are widely applicable, they predominantly rely on distance-based similarity and operate directly in the original feature space, potentially oversimplifying intricate, non-linear relationships among variables. In multidimensional nutritional statistics encompassing protein, fat, and energy availability, latent structural patterns may not be linearly distinguishable. Thus, conventional clustering techniques may fail to identify intricate interdependencies within the data, underscoring the necessity for more sophisticated representation learning-based clustering methodologies.

Clarification of Deep Embedded Clustering (DEC)

Deep Embedded Clustering (DEC) was initially presented as an unsupervised deep learning framework that concurrently performs representation learning and cluster assignment optimization. In contrast to conventional clustering techniques that operate directly in the original feature space, DEC first uses a stacked autoencoder to derive a low-dimensional latent representation of the input data. Following the pre-training phase, the encoder component is preserved and linked to a clustering layer. The clustering objective is optimized by reducing the Kullback–Leibler (KL) divergence between a soft cluster assignment distribution and a target distribution. This ongoing refinement enables DEC to concurrently update feature representations and cluster centroids, facilitating the model's ability to capture non-linear and complex patterns in high-dimensional data (Lee et al., 2022).

The architecture of DEC generally comprises two primary phases: (1) representation learning via an autoencoder, and (2) clustering optimization within the latent space. The encoder's latent embeddings are initialized via K-Means to establish initial cluster centroids, as demonstrated in previous implementations. Thereafter, soft assignments are calculated using Student’s t-distribution, and the cluster centers are progressively refined by minimizing the KL

divergence. This integrated optimization framework differentiates DEC from traditional pipelines in which dimensionality reduction and clustering are executed independently. By amalgamating these methods, DEC minimizes information loss and enhances cluster compactness and separability (Desai et al., 2023).

Notwithstanding its benefits, numerous studies have recognized the limits of the original DEC paradigm. It largely assumes numerical input features and may exhibit convergence instability due to the moving-target problem during optimization. To tackle these problems, DEC extensions have integrated embedding layers for categorical data and implemented soft-target update algorithms to improve training stability. Empirical assessments using benchmark datasets indicate that DEC-based frameworks frequently outperform conventional clustering methods, including K-Means, Gaussian Mixture Models, and Agglomerative Clustering, in terms of clustering accuracy and normalized mutual information. These findings indicate that DEC is particularly appropriate for intricate, multidimensional datasets where learning hidden structures is crucial.

Clustering of Food Security in Indonesia

Studies on clustering food availability data in Indonesia predominantly employ traditional clustering techniques, including K-Means, K-Medoids, and Fuzzy C-Means. Julianto et al. (2022) conducted a study using K-Means and K-Medoids to cluster Indonesian provinces based on per capita food spending, using secondary data from the Central Statistics Agency (BPS) for 2013–2019. The analysis approach, employing the CRISP-DM technique, encompassed data cleaning, integration, and modeling, with evaluation conducted via the Davies-Bouldin validity index. The findings indicated that K-Means with k=7 yielded the most optimal clustering, as evidenced by a DBI value of 0.341, compared with K-Medoids, which had a value of 0.362. This clustering revealed areas with the highest food expenditure (DKI Jakarta) and the lowest (Central Java, NTT, Southeast Sulawesi, Gorontalo, West Sulawesi), establishing a foundation for strategies to improve food security by augmenting production capacity and food reserves (Julianto et al., 2022).

Setiono and Dianto (2022) examined the availability of rice fields as a metric for food production potential utilizing the Fuzzy C-Means (FCM) approach. This study categorized 34 provinces into three groups: narrow, medium, and wide, utilizing data on land area and rice output from the National Land Agency (BPN) and the Central Statistics Agency (BPS) for the years 2015 to 2019. FCM was selected for its ability to offer variable membership levels within each province within each cluster. The findings indicated that the provinces of Java (West Java, Central Java, and East Java) have the largest rice field acreage and the highest production potential. Conversely, regions with constrained land necessitate policy intervention to avert land conversion and enhance productivity. This study highlights the importance of categorizing regions by food resources to facilitate targeted food security policies (Setiono & Dianto, 2022).

Both results demonstrate that traditional clustering methodologies are proficient in uncovering patterns in the distribution of food availability or supporting resources. The complexity of food data, particularly that encompassing nutritional aspects such as protein, fat, and calories per capita, requires a more advanced methodology. The Deep Embedded Clustering (DEC) approach enhances clustering quality by concurrently combining representation learning using deep neural networks with the clustering process. DEC can identify non-linear patterns and latent correlations among variables, thereby producing more significant clusters for analyzing food availability in Indonesia, which can ultimately inform the development of more precise and evidence-based food security initiatives.

Conventional versus Deep Learning-based Clustering

Conventional clustering techniques, including K-Means and Hierarchical Clustering, are widely used across many fields owing to their simplicity, interpretability, and comparatively low computational cost. K-Means has shown superior partitioning efficacy in customer segmentation experiments, achieving higher silhouette scores than hierarchical clustering in some scenarios. These algorithms generally rely on distance-based similarity metrics (e.g., Euclidean distance) and operate directly in the original feature space. Although they exhibit computational efficiency

and ease of implementation, their efficacy is significantly affected by feature scaling, initialization, and the presumption of linear separability. Furthermore, they may encounter difficulties when managing high-dimensional, sparse, or non-linearly distributed data, since the distance measure may fail to accurately reflect the fundamental relationships among variables (Kumar et al., 2024).

Conversely, deep learning-based clustering methods integrate representation learning with clustering objectives, enabling models to learn latent feature spaces before assigning clusters. Recent research in dynamic network environments indicates that deep learning-based clustering frameworks can substantially outperform traditional clustering methods across performance metrics, including delivery rate, packet drop rate, and latency. Using neural networks, these methodologies may identify intricate, nonlinear patterns and adapt to fluctuations in data distribution. In contrast to conventional techniques that rely solely on geometric proximity, deep models learn hierarchical feature representations, enabling them to reveal hidden structural relationships that may remain undetectable in the original input space (Ali et al., 2024).

Methodologically, the primary distinction resides in the processing of features. Conventional clustering separates feature extraction from clustering, while deep learning-based clustering simultaneously optimizes representation learning and cluster assignment. This integration mitigates information loss caused by manual preprocessing or by dimensionality reduction methods. Nonetheless, deep clustering techniques typically require greater computational resources, meticulous hyperparameter optimization, and larger datasets to achieve stable convergence. Consequently, classical clustering is excellent for low-dimensional or well-structured datasets, whereas deep learning-based clustering is better suited for complex, multidimensional datasets characterized by latent non-linear structures. This comparison provides a theoretical foundation for selecting advanced clustering frameworks for the analysis of complex nutritional availability data.

Previous studies have shown the efficacy of traditional clustering methods, including K-Means, K-Medoids, and Fuzzy C-Means, for examining food expenditure patterns and agricultural resource allocation in Indonesia; however, these techniques predominantly rely on distance-based similarity metrics and operate directly in the original feature space. Previous studies have predominantly focused on economic indicators (e.g., per capita food expenditure) or production-related variables (e.g., rice field area and rice output), neglecting multidimensional indicators of nutritional availability, such as per capita protein, fat, and energy. Furthermore, to our knowledge, no prior research has utilized deep learning-based clustering frameworks—specifically Deep Embedded Clustering (DEC)—to examine Indonesian food availability data. This indicates a significant research gap in integrating representation learning and clustering for complex nutritional datasets. The nutritional availability data may display non-linear relationships and latent structural patterns across regions and time. The lack of deep clustering methodologies in previous Indonesian food security studies underscores the necessity for a more sophisticated methodological framework that can uncover deeper data structures beyond the capabilities of traditional clustering techniques.

3. Research Methods

This study followed several main stages, including data collection, data preparation, data analysis, modeling, and model evaluation. Each stage was designed to ensure optimal clustering results using the DEC approach on food availability data in Indonesia.

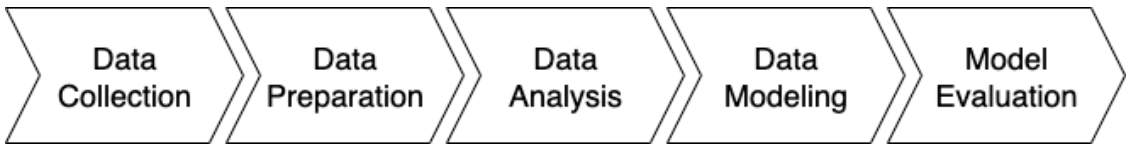


Fig. 1. Research Methodology

Data Collection

Data was collected from the Indonesian government's open data portal via the website data.go.id. The dataset used includes the Energy Availability per Capita (updated in 2024)[1], which shows the amount of food energy available to each individual in kilocalories per day. The sufficiency standard refers to the recommendation of the National Food and Nutrition Council (WNPG) of 2,400 kilocalories per capita per day, and values exceeding this limit indicate food energy sufficiency. Next is the Provincial Consumption Food Pattern Score (PPH) (updated in 2024)[2], which is an index that reflects the diversity and nutritional balance of food consumption among the population, where a higher score indicates a more diverse, nutritionally balanced consumption pattern that is in line with national nutritional recommendations. Another dataset is the Per Capita Protein Availability Figure (updated in 2024)[3], which measures the amount of food protein available to each individual in grams per day, with a WNPG reference of 63 grams per capita per day; values above this threshold indicate sufficient protein for health and growth. Finally, the Per Capita Fat Availability Figure (updated in 2024)[4] shows the amount of dietary fat available per person in grams per day; the higher the value, the greater the contribution of fat to daily energy intake, although balance must still be maintained to avoid health risks. These four datasets cover information at both the national and provincial levels across Indonesia, enabling a comprehensive analysis of food availability and the quality of consumption patterns.

Data Preparation

The dataset used in this study comes from four primary sources on the data.go.id portal, namely the Provincial Consumption Food Pattern Score (PPH), Protein Availability per Capita, Energy Availability per Capita, and Fat Availability per Capita. The PPH dataset contains two columns, namely Year and PPH Score, which represent the diversity and nutritional balance index of food consumption among the population. The protein availability dataset contains information on the availability of plant-based, animal-based, and total protein per capita per day, complete with units and regional descriptions. The energy availability dataset contains data on plant-based energy, animal-based energy, and total energy in kilocalories per capita per day. In contrast, the fat availability dataset covers plant-based fat, animal-based fat, and total fat in grams per capita per day.

Table 1 – Data Preparation.

Type	Number of Rows	Number of Columns
National	7	11
Province	34	10

The initial stage of data processing was carried out by performing an inner join of the four datasets based on province names or national entities. This process ensured that only rows with complete data for all variables were retained for analysis. After the merge, the data is divided into two subsets: national data and provincial data. The national data contains annual aggregations from all provinces, covering the period 2018–2024, with variables for energy availability, fat, and protein (each divided into plant and animal sources) as well as PPH scores. The provincial data covers 34 provinces with the same variables for 2023, but without the year column, as it represents data for the last available year. Rows with missing values were deleted to ensure the integrity of the dataset, so that each observation has complete values for all variables analyzed (Singh & Tiwari, 2025).

Data Analysis

Exploratory data analysis was conducted to understand the relationships between variables (Rachmawati & Darmawan, 2024). Correlations between food availability indicators were analyzed and visualized using heatmaps, while time trends and value distributions were displayed

[1] https://data.go.id/dataset/dataset/angka-ketersediaan-energi-per-kapita-update-tahun-2024
[2] https://data.go.id/dataset/dataset/skor-pola-pangan-harapan-konsumsi-provinsi-update-tahun-2024
[3] https://data.go.id/dataset/dataset/angka-ketersediaan-protein-per-kapita-update-tahun-2024
[4] https://data.go.id/dataset/dataset/angka-ketersediaan-lemak-per-kapita-update-tahun-2024

using line graphs and scatter plots. This stage aimed to identify initial patterns relevant to cluster formation in the modeling stage (Huang et al., 2024). Operationally, national data were read from a curated file, the columns were normalized in naming to ensure consistency, and then a subset containing all nutritional variables (energy, fat, protein—each from plant and animal sources—and PPH scores) was formed. Inter-year trends are then visualized using line graphs for several key metric pairs: plant-based fat versus total fat, plant-based energy versus total energy, and plant-based protein versus total protein. The X-axis is anchored to the year, axis labels and legends are added, and a grid is enabled to facilitate reading changes between periods (Rachmawati et al., 2024). To assess the closeness and potential redundancy between features, a correlation matrix was calculated from the numerical subset, then visualized as a lower triangle heatmap (upper triangle masked) with coefficient annotations—this approach emphasizes variable pairs with high correlations that could potentially affect the modeling process or require attention during the preprocessing stage (Chen et al., 2023). At the provincial level, data is read from the merged and labeled cluster files, then a scatter plot is created for the identical three variable pairs (fat, energy, protein) with the Y-axis on total availability and the X-axis on plant components; marker size is enlarged and a grid is enabled to highlight linear patterns and deviations. This plot serves to check the consistency of the relationship between plant components and total for each indicator, while also observing the presence of potential outliers or heterogeneity between provinces; coloring based on cluster labels can be added at a later stage to assess the visual separability between clusters.

Data Modelling

Before modeling, the data was standardized using Standard Scaler to equalize the scale of all variables so that no attribute dominated the learning process (Kim et al., 2024). The initial phase of modeling involved generating preliminary cluster labels using the Hierarchical Clustering method (Apfel & Liang, 2024), with four clusters derived from the dendrogram produced during the hierarchical grouping procedure. Hierarchical Clustering was chosen to generate pseudo-labels due to its lack of reliance on random centroid initialization and its ability to reveal inherent grouping patterns via a dendrogram representation (Yang & Lin, 2024). This attribute offers a more stable, interpretable initial partition than centroid-based approaches like K-Means, which are susceptible to initialization and may yield inconsistent labels across iterations (Khan et al., 2024). Analysis of the dendrogram revealed the optimal cut level for four clusters, ensuring that the pseudo-labels accurately reflected the data's inherent hierarchical structure (Khaerani et al., 2024). The cluster labels were later used as pseudo-labels in the partially supervised learning phase of the Deep Embedded Clustering (DEC) framework to guide the representation learning during the first optimization stage (S. Wang et al., 2023).

The DEC model is built with an autoencoder architecture that combines dimension reduction and cluster learning simultaneously (Ma et al., 2023). The autoencoder consists of an input layer that receives standardized data, a low-dimensional encoder layer (two neurons) with a ReLU activation function to extract latent representations, and a decoder layer with ELU activation that reconstructs the input data. Additionally, a cluster output layer with SoftMax activation is added to predict cluster membership based on the latent representation. The model is optimized using the Adam optimizer, with two loss functions running simultaneously: mean squared error (MSE) for data reconstruction and sparse categorical Cross-entropy for cluster classification, weighted 1.0 and 0.5, respectively.

To guarantee the robustness and stability of the proposed framework, two DEC configurations were utilized with varying architectural and training parameter settings. The purpose of employing two configurations was to assess the sensitivity of clustering performance to changes in latent space dimensionality and to the weighting of the loss between the reconstruction and clustering objectives. As DEC concurrently optimizes representation learning and cluster assignment, the equilibrium between these two aims can profoundly affect convergence behavior and cluster distinctiveness (Wang et al., 2024). This work aims to evaluate the consistency of the clustering structure across two configurations under varying optimization dynamics, thereby enhancing the reliability of the results. This comparative framework further

illustrates that the observed clustering patterns are not mere artifacts of a particular parameter configuration, but rather indicative of inherent structures within the nutritional availability data.

The training process is run for a maximum of 300 epochs with a small batch size to increase the sensitivity of weight updates (Bussa et al., 2025). Several callbacks are used to improve performance, namely EarlyStopping to stop training when the loss does not improve in three epochs, ReduceLROnPlateau to reduce the learning rate adaptively, and ModelCheckpoint to save the best model weights based on the lowest loss value (Testas, 2024).

Model Evaluation

The evaluation was conducted by monitoring training metrics, including cluster loss, and cluster accuracy, which were visualized in a line graph against the number of epochs. Cluster loss reflects the error rate in predicting cluster labels (pseudo-labels) based on the latent representation generated by the encoder, where a decrease in this value indicates an improvement in cluster separation ability. Meanwhile, cluster accuracy describes the proportion of cluster predictions that match the initial reference labels (Lakshmi et al., 2024). Hence, an increasing accuracy trend indicates the model's ability to maintain or improve the established cluster structure (Azzam et al., 2024). Additionally, clustering performance is measured using the silhouette score (Ilyas & Priscila, 2024). A comparison is made between two scenarios: clustering using Hierarchical Clustering alone and clustering using Deep Embedded Clustering, to assess how much the integration of latent representation learning improves the quality of the generated clusters.

Table 2 – Metric Value Range

Metric	Minimum Value	Maximum Value
Accuracy	0	1
Loss	0	~
Silhouette Score	-1	1

In addition to being used in the selection of the best parameters for Hierarchical Clustering, the silhouette score is also used to compare the quality of clustering results between pure Hierarchical Clustering and the DEC approach. The silhouette score measures how similar objects within a cluster are to other members of the same cluster compared to other clusters, with values ranging from -1 to 1; the closer to 1, the better the separation and compactness of the formed clusters (Thongnim et al., 2023). In the context of this study, the silhouette score value from optimized Hierarchical Clustering is used as an initial reference, then compared with the value produced by DEC. The difference in scores reflects the extent to which the integration of latent representation learning in DEC can improve cluster structure, enhance separation between groups, and reduce overlap between data, compared to traditional hierarchical methods (A. Zheng et al., 2025).

4. Results and Discussions

The improved performance of Deep Embedded Clustering (DEC) compared to K-Means stems from its approach to learning better feature descriptions before grouping data. Unlike K-Means, which works with the original data and relies on distance-based measures of similarity, DEC uses an autoencoder to learn these descriptions during clustering (Li et al., 2024). This means the model changes the input data into a simpler, smaller set of features, making it easier to find detailed patterns among variable (Li et al., 2024) s. In data on food, such as protein, fat, and energy per person, there may be complex, non-linear links among places that simple distance measurements cannot show well. By learning more useful features, DEC can form tighter, more clearly separated groups than K-Means. This helps DEC discover important patterns in detailed food availability data.

This dataset contains annual national data for the period 2018–2024, covering key food availability indicators in Indonesia. Each row represents one year of observation with variables in the form of energy, fat, and protein availability per capita per day, each of which is divided into plant and animal components, as well as the total value. Energy is expressed in kilocalories, while fat and protein are expressed in grams. Additionally, there is a Food Consumption Pattern Score (FCPS) that reflects the level of dietary diversity and nutritional balance, with relatively high

values during this period (86.3–94.1). This data enables analysis of trends in nutritional adequacy and quality at the national level, as well as the relationship between the contribution of plant-based and animal-based foods to total nutritional availability and the PPH score.

Table 3 – National Data.

Year	Availability of Plant-based Energy	Availability of Animal-based Energy	Total Energy Availability	Availability of Plant-based Fat	Availability of Animal-based Fat	Total Fat Availability	Availability of Animal-based Protein	Availability of Plant-based Protein	Total Protein Availability	PPH Score
2018	2726	249	2974	47.09	14.92	62.02	57.22	25.53	82.75	88.4
2019	2595	263	2858	59.70	15.97	75.67	55.65	26.65	82.29	87.9
2020	2624	251	2875	41.79	15.44	57.23	55.34	24.90	80.24	86.3
2021	2856	263	3119	52.39	15.90	68.28	54.59	26.66	81.26	87.2
2022	2907	254	3161	65.14	15.34	80.49	49.60	26.24	75.85	92.9
2023	2820	259	3079	55.51	15.55	71.06	54.95	26.71	81.66	94.1
2024	2602	245	2847	53.96	15.62	69.58	51.22	23.29	72.61	93.5

This correlation heatmap shows the relationship between national food availability variables for the period 2018–2024, where red indicates a positive correlation and blue indicates a negative correlation. Correlation values close to 1 indicate a powerful positive relationship (Khalil et al., 2024), such as between Availability of Plant-based Fat and Total Fat Availability (0.9991), Availability of Plant-based Energy and Total Energy Availability (0.9987), and Availability of Plant-based Protein and Total Protein Availability (0.9307). A high positive relationship is also observed between Availability of Animal-based Energy and Availability of Animal-based Protein (0.8857) as well as between Availability of Animal-based Energy and Availability of Animal-based Fat (0.6595).

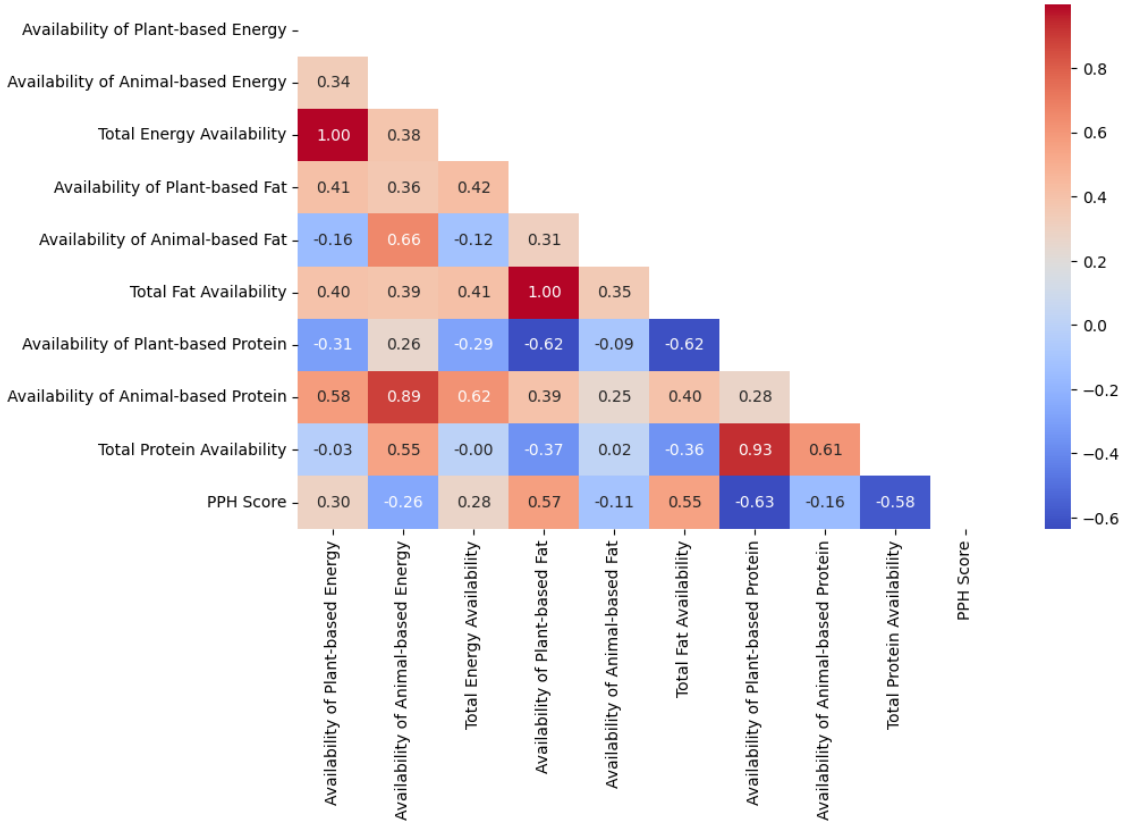


Fig. 2. Correlation Heatmap

On the other hand, there are strong negative correlations, such as between Availability of Plant-based Protein and PPH score (-0.6342) and between Total Protein Availability and PPH score (-0.5832). This indicates that an increase in protein availability, particularly from plant sources, does not always go hand in hand with an increase in dietary diversity. Negative

correlations are also observed between Availability of Plant-based Protein and Availability of Plant-based Fat (-0.6232) as well as between Availability of Plant-based Protein and Total Fat Availability (-0.6182), which may indicate differences in the contributions of plant-based protein sources and fat sources to national consumption patterns. Thus, this heatmap helps identify closely related variables, both positively and negatively, which can influence cluster formation and the interpretation of food security analysis results.

Table 4 – Correlation between Variables.

Variable-1	Variable-2	Correlation
Availability of Plant-based Fat	Total Fat Availability	0.999110
Availability of Plant-based Energy	Total Energy Availability	0.998688
Availability of Plant-based Protein	Total Protein Availability	0.930702
Availability of Animal-based Energy	Availability of Animal-based Protein	0.885701
Availability of Animal-based Energy	Availability of Animal-based Fat	0.659491
Total Protein Availability	PPH Score	-0.583205
Total Fat Availability	Availability of Plant-based Protein	-0.618158
Availability of Plant-based Fat	Availability of Plant-based Protein	-0.623244
Availability of Plant-based Protein	PPH Score	-0.634228

The three graphs show the annual trends between the availability of plant-based foods and the total for fat, energy, and protein in Indonesia for the period 2018–2024. The patterns in the three pairs of variables appear to move in tandem, which is consistent with the high correlation values in the previous analysis, namely 0.9991 for fat, 0.9987 for energy, and 0.9307 for protein. Similar changes follow increases or decreases in the availability of plant-based components in total availability, such as the surge in 2022 that is visible in fat and energy. This indicates that contributions from plant-based sources highly influence variations in total nutrient availability. At the same time, the magnitude of the correlation suggests that the proportion of plant-based components relative to the total remains relatively consistent over time.

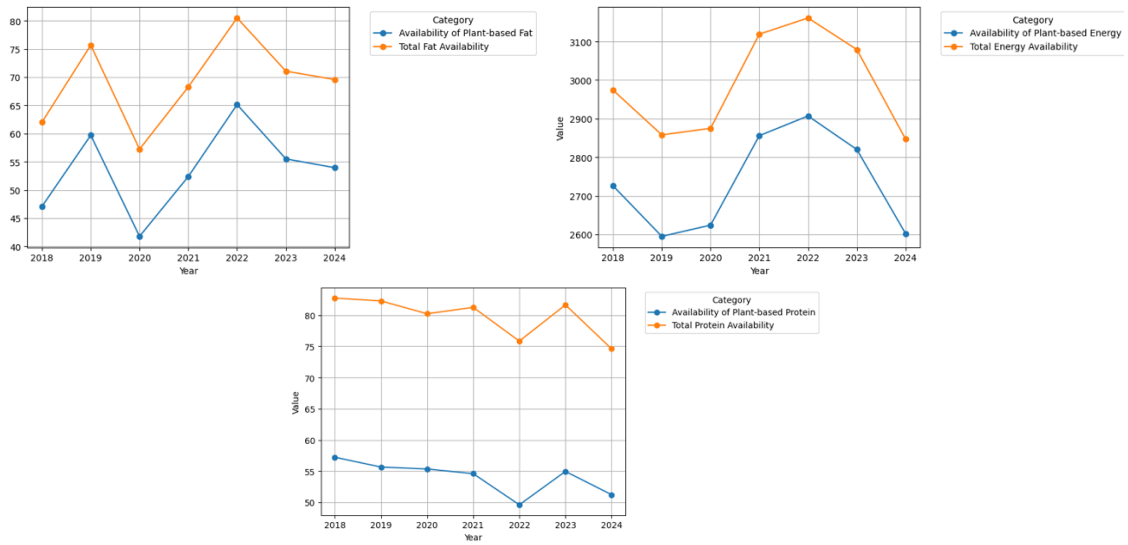


Fig. 3. Line Chart of Positive Correlations

Descriptive statistics for provincial data in 2023 show significant variations between regions in terms of food availability. The average total energy availability reached 3,109 kcal/capita/day, with the main contribution coming from plant sources (2,765 kcal) and the remainder from animal sources (343 kcal). The average fat availability was 68.89 grams/capita/day, consisting of 49.13 grams of plant-based fat and 19.76 grams of animal-based fat. For protein, the average total was 98.51 grams/capita/day, with plant-based protein at 58.41 grams and animal-based protein at 40.10 grams. The data distribution is quite broad, as evidenced by the high standard deviation in some variables, particularly plant-based energy (705.04) and animal-based protein (26.26), indicating disparities between provinces. The minimum and

maximum values also show significant gaps, for example, total energy availability ranges from 1,850 to 4,835 kcal, and total protein from 64.55 to 221.49 grams. This variation serves as an important basis for the application of Deep Embedded Clustering (DEC) to group provinces based on similarities in food availability profiles.

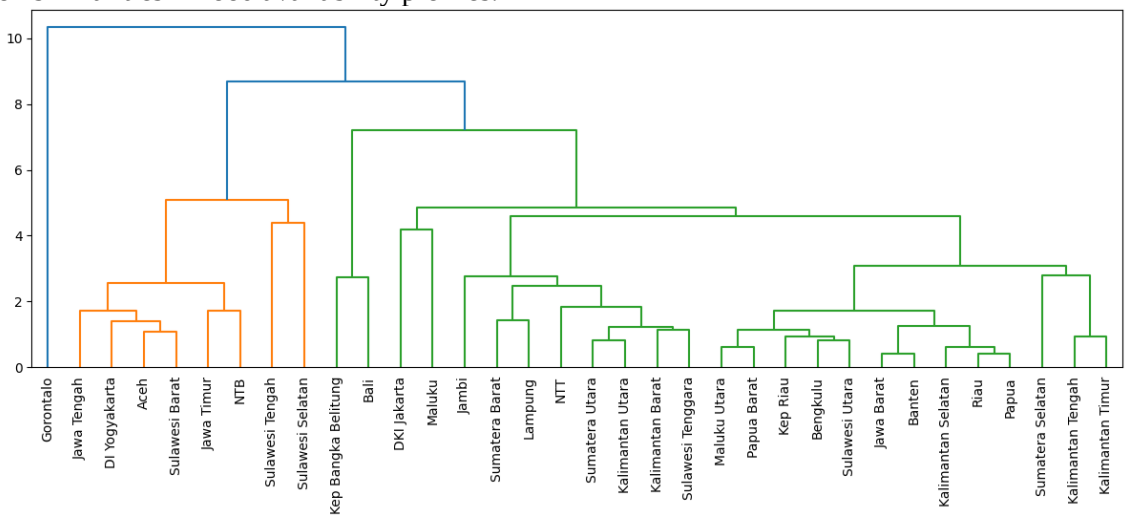


Fig. 4. Hierarchical Clustering Dendrogram

The initial modeling stage was conducted by comparing two traditional clustering methods, namely K-Means and Hierarchical Clustering, using default parameters and the same number of clusters (n_clusters = 4). Standardized data (X_scaled) was used as input for both methods, and the quality of the clustering results was measured using the silhouette score as an indicator of cluster separation. The evaluation results showed that Hierarchical Clustering had a higher silhouette score (0.3958) compared to K-Means (0.3417), indicating that the cluster structure formed was more transparent and had better separation between groups. Based on these results, Hierarchical Clustering was selected as the initial reference for cluster label formation, which will later be used as pseudo-labels in the DEC stage.

Table 5 – Parameter Tuning

Linkage			
	Euclidean	Cosine	Manhattan
Ward	0.3957551078085023	-	-
Single	0.3705437073388827	0.2531547914550028	0.20785585247480093
Average	0.3705437073388827	0.3517315638944432	0.3368982887097225
Complete	0.3957551078085023	0.20227947312472394	0.23260248604993408

After establishing Hierarchical Clustering as the reference clustering method, parameter optimization was performed using GridSearchCV to find the best combination of two important features, namely metric and linkage, in order to obtain the highest silhouette score. This process evaluated various combinations, but it should be noted that the Ward linkage method can only be used with the Euclidean metric due to its calculation properties that minimize variance between clusters. Based on the test results, the combination of Ward and Euclidean was selected because it produced the highest silhouette score among all tested combinations, indicating that these parameters are capable of forming the most transparent cluster structure and achieving optimal separation between cluster members.

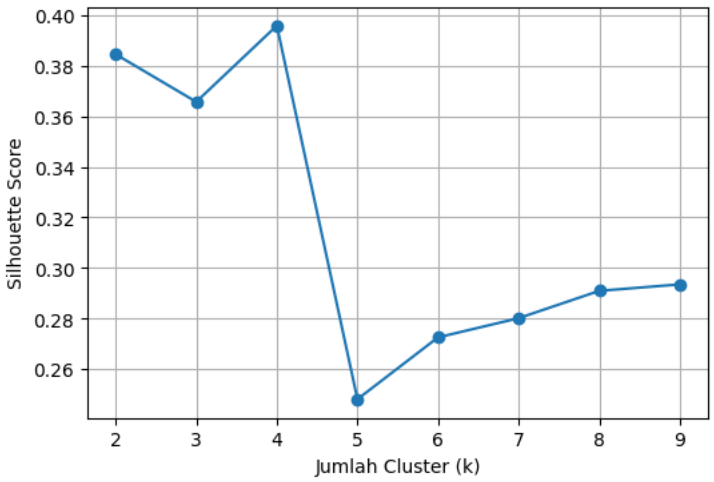


Fig. 5. Elbow Method

To determine the optimal number of clusters (k) in Hierarchical Clustering using the Ward linkage parameter and Euclidean metric, the k-elbow method based on the silhouette score was employed. The graph shows that the highest silhouette score was obtained at k = 4 (0.3958), slightly higher than k = 2 (0.3845) and k = 3 (0.3657). After k = 4, the score drops sharply at k = 5 (0.2482) and only slightly increases for larger k values, but does not exceed the score at k = 4. This indicates that selecting k = 4 provides the best balance between cluster separation and compactness, making it the optimal number of clusters for forming pseudo-labels to be used in the DEC stage.

This plot model (as shown in Figure 6) illustrates the Deep Embedded Clustering (DEC) architecture built on an autoencoder structure with an additional output branch for cluster prediction. The model receives 9-feature inputs (input_dim = 9), which are then projected into a low-dimensional latent representation space (encoding_dim = 2) through the encoder layer. This latent representation branches into two paths: the first leads to the decoder to reconstruct the data back to its original dimension (9 features) with Mean Squared Error (MSE) loss, and the second leads to the cluster layer with the number of neurons corresponding to the number of specified clusters (n_clusters = 4) and a softmax activation function to generate cluster membership probabilities. The model is optimized using the Adam optimizer, with a loss weight of 1.0 for reconstruction and 0.5 for cluster prediction, and monitors the accuracy metric on the cluster output. This architecture enables the learning of latent representations that not only reconstruct data but also form cluster structures aligned with the initial labels from traditional clustering methods.

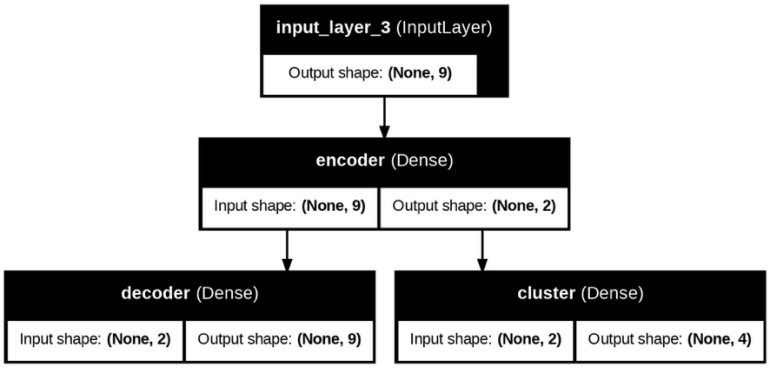


Fig. 6. Model Structure Plot

The parameter information in the model summary shows that the 59 trainable parameters are the weights and biases in the encoder, decoder, and cluster layers that will be updated during the training process to minimize the model loss. The non-trainable parameters have a value of 0,

meaning that no parameters are frozen or kept constant during training. Meanwhile, the 120 optimizer parameters are internal parameters of the optimization algorithm (Adam optimizer in this case) used to control weight updates, such as momentum and gradient average estimation, which are not part of the model weights but are stored and updated during training.

Table 6 – The Params of Model

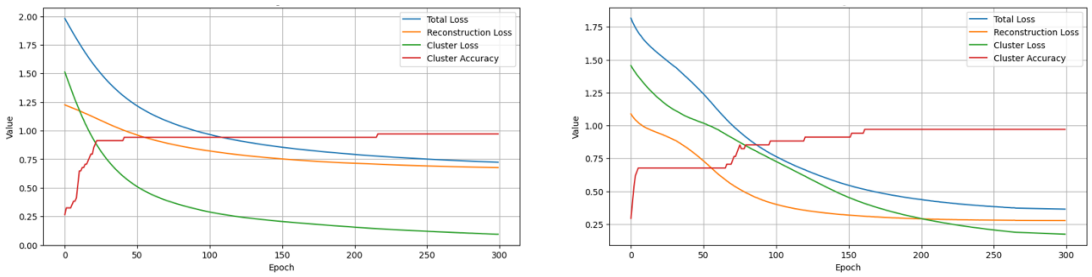
Type of Params	Total
Trainable	59
Non-trainable	0
Optimizer	120

In the application of Deep Embedded Clustering (DEC), two different experiments, labeled DEC-1 and DEC-2, were conducted. Both experiments used an autoencoder architecture with the same cluster branches. However, the difference lay in the training configuration or hyperparameter settings such as batch size, number of epochs, or callback strategies (e.g., EarlyStopping and ReduceLROnPlateau). The objective is to observe the impact of these configuration variations on clustering performance, measured using metrics such as silhouette score, total loss, reconstruction loss, cluster loss, and cluster accuracy.

Table 7 – The Model Experiment

Layer	Activation Function	
	DEC-1	DEC-2
encoder (Dense)	Tanh	ReLu
decoder (Dense)	Sigmoid	ELU
cluster (Dense)	SoftMax	SoftMax

By conducting these two experiments, researchers can compare the results and select the configuration that provides the optimal cluster separation and the best training stability. The activation function significantly impacts the performance of a deep learning model because it determines how nonlinear signals are processed in each neuron, which affects the model's ability to capture complex patterns and avoid problems like vanishing or exploding gradients (Rachmawati et al., 2025).



(a) DEC-1 (b) DEC-2

Fig. 7. Training History Graph

The training results show that both experiments, DEC-1 and DEC-2, successfully achieved high cluster accuracy close to 1.0, with DEC-1 slightly outperforming DEC-2 in terms of average cluster accuracy (0.9202) compared to DEC-2 (0.8734). In terms of loss, DEC-1 has a lower cluster loss (0.3227) than DEC-2 (0.5746), indicating better cluster separation quality. However, DEC-2 outperforms DEC-1 in terms of decoder loss (0.4439 vs. 0.8123) and total loss (0.7312 vs. 0.9736), indicating better latent data reconstruction capabilities. The training graph shows a stable decreasing trend in total loss, reconstruction loss, and cluster loss in both experiments, as well as a consistent increase in cluster accuracy approaching the maximum. Overall, DEC-1 is more optimal for cluster separation, while DEC-2 is superior in representing input data through reconstruction.

Table 8 – The Comparison of Performance Metric Between DEC-1 and DEC-2

Performance Metric	DEC-1	DEC-2
cluster_accuracy	0.920196	0.873431
cluster_loss	0.322736	0.574608
decoder_loss	0.812274	0.443853
loss	0.973642	0.731157

The silhouette score comparison below shows that the application of Deep Embedded Clustering (DEC) significantly improves cluster separation quality compared to pure Hierarchical Clustering. The initial Hierarchical Clustering model had a score of 0.3958, while after being integrated with DEC, its performance improved to 0.7829 on DEC-1 and 0.6385 on DEC-2. The most significant improvement was achieved by DEC-1, which nearly doubled the initial score, indicating that the latent representations from the encoder effectively strengthen the cluster structure and minimize overlap between groups. While DEC-2 also shows significant improvement compared to the initial model, its performance remains below DEC-1, suggesting that the DEC-1 training configuration is more optimal for this data.

Table 9 - The Comparison of Silhouette Score

Model	Silhouette Score
Hierarchical Clustering	0.39575510
Hierarchical Clustering + DEC-1	0.7829114
Hierarchical Clustering + DEC-2	0.63848126

The clustering results produced four groups of provinces with different food availability characteristics. Cluster 0 contains 23 provinces, including most of Sumatra, Kalimantan, and several provinces in Sulawesi, Maluku, and Papua, which are likely to have relatively similar food availability profiles and tend to be at a moderate level in terms of nutritional indicators. Cluster 1 includes eight provinces, such as Aceh, Central Java, East Java, and Yogyakarta, which may have higher food availability or different distribution patterns compared to Cluster 0, given the representation of provinces with large food production bases. Cluster 2 consists only of Gorontalo, indicating a very distinctive food availability profile that differs from other provinces, thus isolating it into a single cluster. Cluster 3 includes the Bangka Belitung Islands and Bali, which, despite their geographical differences, share similarities in food availability indicators that distinguish them from other groups. This cluster structure highlights disparities among provinces, both in terms of quantity and characteristics of food availability.

The clustering results yield several practical consequences for policymakers, nutrition specialists, and the food sector. The observed clusters can inform policymakers in developing regions in developing food security plans, as provinces within the same cluster exhibit analogous patterns of nutritional availability. Regions in Cluster 0 may require policies to enhance food distribution systems and expand access to diverse nutrient sources, whereas provinces in Cluster 1—distinguished by superior food production capacity—might focus on maintaining supply stability and refining value chains for nutrient-dense foods. For nutrition specialists, the cluster patterns can help pinpoint areas where dietary diversity or protein availability require targeted nutritional interventions, such as advocating balanced meals or improving access to animal-derived protein sources. Simultaneously, these clusters offer the food sector valuable insights into regional market attributes, facilitating enhanced planning of food supply chains and distribution strategies, as well as the development of nutrition-focused food products tailored to the specific nutritional requirements of distinct regions. The clustering analysis elucidates structural disparities in food availability and provides practical recommendations to enhance food security and nutritional outcomes across Indonesia.

Table 10 – Province Cluster Result

Cluster	Name	Total
0	Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kep Riau, DKI Jakarta, Jawa Barat, Banten, NTT, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tenggara, Maluku, Maluku Utara, Papua Barat, Papua	23

1	Aceh, Jawa Tengah, Jawa Timur, DI Yogyakarta, NTB, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Barat	8
2	Gorontalo	1
3	Kep Bangka Belitung, Bali	2

Descriptive statistics median per cluster shows striking differences in energy, fat, and protein availability between groups. Cluster 2 stands out with the highest values across all indicators, particularly total energy availability (4,835 kcal), total fat (128.18 g), and total protein (221.49 g), driven by a significant contribution from animal protein (152.55 g). Cluster 1 has high total energy availability (3,867 kcal) and moderate total fat (76.08 g), but relatively low animal protein (24.57 g), indicating a dominance of plant-based sources. Cluster 0 shows the second-lowest nutrient availability, with total energy of 2,674 kcal and total protein of 84.45 g, balanced between plant-based and animal-based sources. Meanwhile, Cluster 3 is unique with relatively low total energy (2,925.5 kcal) but a very high proportion of animal fat (67.58 g out of a total of 90.86 g), indicating a consumption pattern dominated by animal fat despite its total energy and protein intake not being as high as Clusters 2 or 1. These differences illustrate the diversity of food availability patterns across DEC clusters.

Table 11 – Median Cluster Characteristics.

Cluster	Availability of Plant-based Energy	Availability of Animal-based Energy	Total Energy Availability	Availability of Plant-based Fat	Availability of Animal-based Fat	Total Fat Availability	Availability of Animal-based Protein	Availability of Plant-based Protein	Total Protein Availability
0	2386.0	288	2674	38.070	15.600	56.470	48.400	31.270	84.45
1	3574.0	258	3867	66.195	15.105	76.075	76.460	24.570	108.66
2	3944.0	891	4835	104.630	23.550	128.180	68.940	152.550	221.49
3	2076.5	849	2925	23.280	67.580	90.860	49.835	40.085	89.92

The comparatively low protein availability observed in Cluster 0 may be linked to multiple socioeconomic factors that affect food accessibility and consumption behaviors. Areas within this cluster may see diminished household purchasing power, restricted access to diverse food sources, or an increased reliance on staple foods such as rice and other carbohydrate-rich items. In many developing regions, animal-derived protein sources—such as meat, eggs, and dairy—are often more costly and less accessible than plant-based foods, resulting in diminished overall protein availability despite sufficient calorie intake. Moreover, inequities in food delivery systems, regional agricultural output, and market access may exacerbate differences in nutrient availability. These socioeconomic conditions may lead to dietary patterns characterized by adequate calorie intake, while protein intake, especially from animal sources, is constrained. The clustering results not only reveal statistical disparities in nutrient availability but also highlight fundamental structural inequalities in food access and nutritional variety among locations.

Future studies may enhance the current clustering approach by adding time-series forecasting and geospatial mapping tools. This would yield a more thorough understanding of food availability patterns. This study focuses on discovering structural patterns in nutritional availability through clustering. However, adding time-series models, such as ARIMA, LSTM, or other forecasting methodologies, could improve predictions of future trends in protein, fat, and energy availability across locations. Geographic mapping with geographic information systems (GIS) could provide a clearer picture of regional inequities and spatial linkages in food availability. Combining clustering outcomes with temporal and spatial analyses enables policymakers to see not only current nutritional trends but also potential vulnerabilities and concentrated regional food security threats. An integrated analytical framework could support more focused, evidence-driven approaches to improving national food security planning.

5. Conclusion

This study sought to examine food availability patterns throughout Indonesian provinces by utilizing Deep Embedded Clustering (DEC) on multidimensional nutritional variables, including energy, fat, and protein from both plant and animal sources. The modeling procedure commenced with selecting the most appropriate traditional clustering technique, in which

Hierarchical Clustering with Ward linkage and Euclidean distance yielded the highest silhouette score of 0.3958, and was thus employed to provide pseudo-labels for training the DEC framework. The experimental findings indicate that the DEC methodology significantly enhances clustering quality, attaining silhouette scores of 0.7829 in the DEC-1 configuration and 0.6385 in the DEC-2 configuration. The results demonstrate that deep representation learning more efficiently captures nonlinear patterns and latent correlations in nutritional data than classic clustering methods. The observed clusters indicate considerable variation in food supply patterns across Indonesian provinces, ranging from areas with ample and balanced nutrient availability to locations marked by restricted protein intake or unique consumption patterns.

This study advances the literature by demonstrating the efficacy of deep learning-based clustering for analyzing intricate food availability datasets. This research demonstrates that incorporating representation learning via DEC can yield more compact and interpretable clusters, in contrast to prior studies on Indonesian food security that predominantly used conventional clustering methods. The suggested framework demonstrates the successful use of pseudo-labels derived from hierarchical clustering to direct the learning process of a deep clustering model. The findings offer methodological evidence that deep clustering techniques can improve pattern identification in multidimensional nutritional information.

The results also have practical consequences for policymakers, nutrition specialists, and the food sector. By identifying clusters of provinces with analogous food availability patterns, policymakers can formulate more targeted, region-specific food security initiatives, especially in areas with diminished protein availability or imbalanced nutrient composition. For nutrition specialists, clustering patterns can help identify areas that may require dietary diversification initiatives or enhanced access to nutrient-dense food sources. Simultaneously, the food sector may leverage this information to enhance supply chain planning, optimize distribution strategies, and develop nutrition-focused food products tailored to regional dietary requirements.

Notwithstanding these advances, many limits must be recognized. The analysis was performed using a snapshot of existing food data, thereby failing to account for temporal fluctuations in food supply throughout the years. The clustering approach primarily emphasizes indicators of nutritional availability and does not directly integrate socioeconomic variables, such as income levels, food prices, or distribution infrastructure, which may affect regional food consumption patterns. These limitations may limit the ability to comprehensively elucidate the fundamental causes of nutritional differences across areas.

Subsequent research may enhance this study by incorporating Deep Embedded Clustering with multi-year time-series data to elucidate temporal changes in food availability and nutritional dynamics among provinces. Utilizing spatial analytic approaches, including geographic information systems (GIS) mapping, can elucidate spatial dependencies and regional clustering patterns. Moreover, integrating nutritional indicators with socioeconomic and agricultural production variables may facilitate a more thorough comprehension of the factors influencing food security. Such extensions would facilitate the development of a comprehensive analytical framework to support adaptive, evidence-based food security policy planning in Indonesia.

Acknowledgement

The authors wish to extend their heartfelt gratitude to PENS for their invaluable financial assistance, which played a crucial role in the successful completion of this study. This support enabled access to essential resources, tools, and research opportunities, facilitating comprehensive analysis and experimentation in the field of deep learning-based clustering for food availability data. The authors sincerely appreciate the institution's unwavering encouragement, provision of an intellectually stimulating environment, and dedication to fostering research excellence, all of which have significantly contributed to the advancement of this work.

References

Ahmad, A., Liew, A. X. W., Venturini, F., Kalogeras, A., Candiani, A., Di Benedetto, G., Ajibola, S., Cartujo, P., Romero, P., Lykoudi, A., De Grandis, M. M., Xouris, C., Lo Bianco, R., Doddy, I., Elegbede, I., Labate, G. F. D., Del Moral, L. F. G., & Martos, V. (2024). AI can

empower agriculture for global food security: challenges and prospects in developing nations. *Frontiers in Artificial Intelligence*, 7. https://doi.org/10.3389/frai.2024.1328530

Apfel, N., & Liang, X. (2024). Agglomerative hierarchical clustering for selecting valid instrumental variables. *Journal of Applied Econometrics*, 39(7), 1201–1219. https://doi.org/10.1002/jae.3078

Azzam, A. F., Maghrabi, A., El-Naqeeb, E., Aldawood, M., & ElGhawalby, H. (2024). Morphological Accuracy Data Clustering: A novel algorithm for enhanced cluster analysis. *Applied Computational Intelligence and Soft Computing*, 2024(1). https://doi.org/10.1155/2024/3795126

Bui, V. H., & Phan, H. T. (2023). The Computational Complexity of Hierarchical Clustering Algorithms for Community Detection: a review. *Vietnam Journal of Computer Science*, 10(04), 409–431. https://doi.org/10.1142/s2196888823300016

Bussa, S. K., Boppana, N. K., & Deka, B. (2025). A performance-driven evaluation of deep learning for concrete crack detection with varying dataset sizes and training epochs: real-world implications for infrastructure monitoring. *Asian Journal of Civil Engineering*, 26(7), 3063–3081. https://doi.org/10.1007/s42107-025-01361-4

Chen, Y., Li, L., & Li, X. (2023). Correlation analysis of structural characteristics of table tennis players' hitting movements and hitting effects based on data analysis. *Entertainment Computing*, 48, 100610. https://doi.org/10.1016/j.entcom.2023.100610

Darmawan, Z. M. E., Dianta, A. F., Fathoni, K., Rachmawati, O. C. R., & Apriandy, K. I. (2025). Comparison of Machine learning classification Methods for weather Prediction: A Performance analysis. *Jurnal Teknologi Terapan G-Tech*, 9(2), 715–727. https://doi.org/10.70609/gtech.v9i2.6649

Desai, D. D., Dey, J., Satapathy, S. K., Mishra, S., Mohanty, S. N., Mishra, P., & Panda, S. K. (2023). Optimal ambulance positioning for road accidents with deep embedded clustering. *IEEE Access*, 11, 59917–59934. https://doi.org/10.1109/access.2023.3284993

Ennaouri, M., & Zellou, A. (2024). A scoring approach for detecting fake reviews using MRCS similarity metric enhanced by personalized k-means. *Bulletin of Electrical Engineering and Informatics*, 14(1), 587–595. https://doi.org/10.11591/eei.v14i1.8288

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. *Briefings in Bioinformatics*, 24(2). https://doi.org/10.1093/bib/bbad002

Huang, Y., Zeng, P., & Zhong, C. (2024). Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC Bioinformatics*, 25(1). https://doi.org/10.1186/s12859-024-05749-y

Ilyas, F. M., & Priscila, S. S. (2024). An optimized clustering quality analysis in K-Means Cluster using silhouette scores. In *Advances in computational intelligence and robotics book series* (pp. 49–63). https://doi.org/10.4018/979-8-3693-1355-8.ch004

Julianto, I. T., Kurniadi, D., Nashrulloh, M. R., & Mulyani, A. (2022). DATA MINING CLUSTERING FOOD EXPENDITURE IN INDONESIA. *Jurnal Teknik Informatika (Jutif)*, 3(6), 1491–1500. https://doi.org/10.20884/1.jutif.2022.3.6.331

Khaerani, P. I., Musa, Y., Utamy, R. F., & Ishii, Y. (2024). Botanical composition and yields of forages in natural pastures using principal component analysis and cluster dendrogram in South Sulawesi, Indonesia. *OnLine Journal of Biological Sciences*, 24(4), 613–623. https://doi.org/10.3844/ojbsci.2024.613.623

Khalil, N. Z., Kong, N., & Fricke, H. (2024). The influence of GNP on the mechanical and thermomechanical properties of epoxy adhesive: Pearson correlation matrix and heatmap application in data interpretation. *Polymer Composites*, 45(10), 8997–9018. https://doi.org/10.1002/pc.28390

Khan, A. A., Bashir, M. S., Bashir, M. S., Batool, A., Raza, M. S., Bashir, M. A., & Bashir, M. A. (2024). K-Means Centroids initialization based on differentiation between instances attributes. *International Journal of Intelligent Systems*, 2024(1). https://doi.org/10.1155/2024/7086878

Kumar, S., Rani, R., Pippal, S. K., & Agrawal, R. (2024). Customer segmentation in e-commerce: K-means vs hierarchical clustering. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 23(1), 119. https://doi.org/10.12928/telkomnika.v23i1.26384

Lakshmi, H. N., Ramana, T. V., K, L. P., Reddy, L. K. K., & Raju, K. B. (2024). A novel comprehensive investigation for enhancing cluster analysis accuracy through ensemble learning methods. *International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering*, 14(5), 5802. https://doi.org/10.11591/ijece.v14i5.pp5802-5812

Lee, Y., Park, C., & Kang, S. (2022). Deep Embedded clustering framework for mixed data. *IEEE Access*, 11, 33–40. https://doi.org/10.1109/access.2022.3232372

Li, M., Cao, C., Li, C., & Yang, S. (2024). Deep Embedding clustering based onResidual Autoencoder. *Neural Processing Letters*, 56(2). https://doi.org/10.1007/s11063-024-11586-0

Ma, Y., Pei, Y., & Li, C. (2023). A Deep Embedded Clustering Method Based on β-Variational Autoencoder for Single-Cell RNA Sequencing Data. *2023 6th International Conference on Information Communication and Signal Processing (ICICSP)*, 97–102. https://doi.org/10.1109/icicsp59554.2023.10390592

Murugan, T. M., Lenus, C. R., Sridharan, S., & Malligarjun, A. (2024). Life Time Prediction of an Electromagnet Relay using Clustering based Principal Component Analysis with Hybrid Deep Learning Model. *Journal of Applied Engineering and Technological Science (JAETS)*, 6(1), 715–729. https://doi.org/10.37385/jaets.v6i1.5891

Nugroho, H. Y. S. H., Indrawati, D. R., Wahyuningrum, N., Adi, R. N., Supangat, A. B., Indrajaya, Y., Putra, P. B., Cahyono, S. A., Nugroho, A. W., Basuki, T. M., Savitri, E., Yuwati, T. W., Narendra, B. H., Sallata, M. K., Allo, M. K., Bisjoe, A. R., Muin, N., Isnan, W., Ansari, F., . . . Hani, A. (2022). Toward Water, energy, and Food Security in Rural Indonesia: a review. *Water*, 14(10), 1645. https://doi.org/10.3390/w14101645

Rachmawati, O. C. R., Barakbah, A. R., & Karlita, T. (2024). Programming language selection for the development of deep learning library. *JOIV International Journal on Informatics Visualization*, 8(1), 434. https://doi.org/10.62527/joiv.8.1.2437

Rachmawati, O. C. R., Barakbah, A. R., & Karlita, T. (2025). The Comparison of Activation Functions in Feature Extraction Layer using Sharpen Filter. *Journal of Applied Engineering and Technological Science (JAETS)*, 6(2), 1254–1267. https://doi.org/10.37385/jaets.v6i2.5895

Rachmawati, O. C. R., & Darmawan, Z. M. E. (2024). The comparison of deep learning models for Indonesian political hoax news detection. *CommIT (Communication and Information Technology) Journal*, 18(2), 123–135. https://doi.org/10.21512/commit.v18i2.10929

Ramadhan, A., Suhendra, A., & Yohanitas, W. A. (2025). ONE Data Indonesia: A Retrospective Analysis of Data Interoperability in Declaring Regional Planning and Development. *KnE Social Sciences*, 10(16), 152–171. https://doi.org/10.18502/kss.v10i16.19169

Rusmawati, E., Hartono, D., & Aritenang, A. F. (2023). Food security in Indonesia: the role of social capital. *Development Studies Research*, 10(1). https://doi.org/10.1080/21665095.2023.2169732

Setiono, H., & Dianto, T. M. (2022). Analysis of rice field cluster in Indonesia as an evaluation of food production availability using Fuzzy C-Means. *Proceedings of the International Conference on Data Science and Official Statistics*, 2021(1), 326–332. https://doi.org/10.34123/icdsos.v2021i1.245

Sahria, Y., Sudira, P., & Priyanto. (2026). Optimization of image compression using K-means clustering for digital heritage archives. *Advance Sustainable Science, Engineering and Technology (ASSET), 8*(1), 02601020. https://doi.org/10.26877/asset.v8i1.2772

Singh, Y., & Tiwari, M. (2025). A Comprehensive Machine Learning Approach for Early Detection of Diabetes on Imbalanced Data with Missing and Outlier Values. *SN Computer Science*, 6(3). https://doi.org/10.1007/s42979-025-03751-6

Sutardi, N., Apriyana, Y., Rejekiningrum, P., Alifia, A. D., Ramadhani, F., Darwis, V., Setyowati, N., Setyono, D. E. D., Gunawan, N., Malik, A., Abdullah, S., Muslimin, N., Wibawa, W., Triastono, J., Yusuf, N., Arianti, F. D., & Fadwiwati, A. Y. (2022). The transformation of rice crop technology in Indonesia: Innovation and sustainable food security. *Agronomy*, 13(1), 1. https://doi.org/10.3390/agronomy13010001

Testas, A. (2024). *Deep Learning with TensorFlow for Classification*. In Apress eBooks (pp. 431–488). https://doi.org/10.1007/979-8-8688-1017-6_7

Thongnim, P., Charoenwanit, E., & Phukseng, T. (2023). Cluster Quality in Agriculture: Assessing GDP and Harvest Patterns in Asia and Europe with K-Means and Silhouette Scores. *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, 1–5. https://doi.org/10.1109/iementech60402.2023.10423469

Wang, L., Nie, R., Zhang, Z., Gu, W., Wang, S., Wang, A., Zhang, J., & Cai, J. (2023). A deep generative framework with embedded vector arithmetic and classifier for sample generation, label transfer, and clustering of single-cell data. *Cell Reports Methods*, 3(8), 100558. https://doi.org/10.1016/j.crmeth.2023.100558

Wang, S., Beheshti, A., Wang, Y., Lu, J., Sheng, Q. Z., Elbourn, S., & Alinejad-Rokny, H. (2023). Learning distributed representations and deep embedded clustering of texts. *Algorithms*, 16(3), 158. https://doi.org/10.3390/a16030158

Wang, Y., Xiao, H., Zhang, Z., Guo, X., & Liu, Q. (2024). Self-supervised representation learning of metro interior noise based on variational autoencoder and deep embedding clustering. *Computer-Aided Civil and Infrastructure Engineering*, 40(4), 503–522. https://doi.org/10.1111/mice.13336

Yang, J., & Lin, C. (2024). Enhanced Adjacency-Constrained hierarchical clustering using Fine-Grained pseudo labels. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3), 2481–2492. https://doi.org/10.1109/tetci.2024.3367811

Zhang, M., & Parnell, A. (2023). Review of clustering methods for functional data. *ACM Transactions on Knowledge Discovery From Data*, 17(7), 1–34. https://doi.org/10.1145/3581789

Zheng, A., Cai, J., Yang, H., Xun, Y., & Zhao, X. (2025). Triple-Stream contrastive deep embedding clustering via semantic structure. *Mathematics*, 13(22), 3578. https://doi.org/10.3390/math13223578

Zheng, Y., Jia, C., Yu, J., & Li, X. (2023). Deep embedded clustering with distribution consistency preservation for attributed networks. *Pattern Recognition*, 139, 109469. https://doi.org/10.1016/j.patcog.2023.109469