ICDSOS
The 3rd International Conference on Data Science and Official Statistics
November 27 - 28, 2025

"From Data to Impact: Leveraging Innovation in Data Science and Official Statistics for Data-Driven Policy Making and Global Competitiveness"

Regional Clustering of Food Insecurity to Support the Attainment of SDG 2: Zero Hunger through Machine Learning Approaches

S Nuradilla[1,*], W Saputra[1], and M Rizal[1]

[1] Study Program of Statistics and Data Science, Department of Statistics, IPB University, Dramaga, Bogor, Indonesia

*Corresponding author's email: siti.nuradilla@apps.ipb.ac.id

Abstract. Food security remains a persistent development challenge in Indonesia, with regional disparities posing significant barriers to achieving equitable access to nutritious and sufficient food. This study aims to classify and cluster districts and cities in Indonesia based on their food security vulnerability levels, thereby supporting the attainment of SDG 2: Zero Hunger. We employed a machine learning approach using a dataset of 514 regions and nine food security indicators sourced from national databases. The classification phase compared three algorithms, Random Forest, XGBoost, and LightGBM, under multiple data preprocessing scenarios, including outlier handling (IQR and Isolation Forest) and class balancing (SMOTE). LightGBM with IQR preprocessing delivered the best performance, achieving an accuracy and F1-score of 0.984. For clustering, DBSCAN and HDBSCAN were applied using the six most important features identified by the classifier. DBSCAN showed slightly better performance based on Silhouette Score (0.5639), resulting in three regional groupings: food-secure, highly vulnerable, and outlier regions. The analysis revealed that socio-economic factors and access to basic infrastructure remain critical determinants of food insecurity. The results underscore the importance of data-driven approaches in policy formulation and highlight the value of machine learning in producing more targeted, efficient, and adaptive food security interventions in Indonesia.

Keyword: Clustering, DBSCAN, Food Security, LightGBM, Machine Learning

1. Introduction

Food security is one of the fundamental pillars of sustainable development. Ensuring strong food security guarantees the availability of sufficient, safe, and nutritious food for the entire population. However, Indonesia continues to face significant challenges in achieving equitable food security across all regions. According to the Global Food Security Index (GFSI) in 2022, Indonesia ranked 69th out of 113 countries, with a score of 59.2, indicating a relatively low level of national food security compared to other ASEAN countries such as Singapore (77.4) and Malaysia (70.1) [1].

The issue of food security in Indonesia is not only evident at the national level but also reveals considerable disparities among regions. Data from Badan Pangan Nasional (2022) shows that some districts/cities fall into the category of highly food-insecure, while others are categorized as highly food-secure. This disparity leads to unequal access to nutritious food, exacerbates poverty, and worsens public health issues such as the high prevalence of stunting in several provinces.

Food security has become a global priority through the commitment to the Sustainable Development Goals (SDGs), particularly Goal 2: Zero Hunger, which aims to end hunger, achieve food security, improve nutrition, and promote sustainable agriculture [3]. At the national level, food security is also a strategic priority in Indonesia's National Medium-Term Development Plan 2020–2024. Therefore, innovative, data-driven strategies are urgently needed to accelerate the identification of food-insecure regions and to enhance national food security.

Traditional methods of analyzing and mapping food security are often static and slow, lacking the capacity to capture the complex, multivariate, and dynamic nature of inter-regional data. With the advancement of technology, data mining and machine learning approaches offer promising solutions for more efficient, accurate, and adaptive classification of regions based on their vulnerability levels [4]. These approaches enable the analysis of large-scale datasets and the detection of patterns that are not easily captured by conventional statistical methods.

Extensive research has been conducted on the application of machine learning methods through classification techniques to determine food vulnerability levels. Study [5] used K-Nearest Neighbor (KNN) to classify food security in Central Java. Another study used Ordinal Probit regression on panel data to model the food security index in Indonesia [6]. Unfortunately, this research lacks flexibility in capturing complex non-linear relationships between indicators, and its performance often degrades when the data is highly heterogeneous or contains outliers. The limitations of conventional machine learning methods can be overcome through tree ensemble-based models such as Random Forest and Gradient Boosting (including XGBoost and LightGBM). These models have the advantage of capturing non-linear relationships and interactions between variables without requiring strict data distribution assumptions [7], [8]. Random Forest offers high performance on complex and heterogeneous data, as it reduces the risk of overfitting and improves generalization. Meanwhile, Gradient Boosting excels at gradually improving predictive performance, minimizing the error of previous models. Ensemble learning and gradient boosting models help minimize bias and volatility for optimal results [9]. Meanwhile, most previous food security classification studies have not addressed the characteristics of regional groups based on similar indicators. Understanding regional cluster patterns can provide additional insights for more effective strategic interventions.

Clustering techniques are needed to identify the characteristics of food-vulnerable areas, allowing for more targeted interventions tailored to the specific circumstances of the area. Several studies have explored regional groupings based on indicators in the agriculture and food security sectors. Study [10] combined K-Means with hierarchical clustering (HAC) to map the food security framework of the rice sector in Indonesian provinces based on rice consumption, production, and prices. On the other hand, the study [11] compared K-Means and K-Medoids in clustering food crop production in West Sumatra. Although K-Means is a popular clustering method, this algorithm has the disadvantage of determining the number of clusters (k) upfront, which is sometimes unclear in complex data contexts. This algorithm also assumes spherical or homogeneous clusters, which is unrealistic for irregular spatial patterns. K-Means is also sensitive to outliers and lacks a mechanism for handling extreme noise. In the context of inter-regional food security data that is heterogeneous, non-normally distributed, and the possible presence of extreme data, these limitations can compromise the validity of clustering.

Addressing the gaps in previous research, this study uses a two-stage approach: conducting a comparative analysis to classify districts/cities experiencing food insecurity using Random Forest, XGBoost, and LightGBM. These algorithms are more robust in handling non-linear and complex data and are able to provide information on the importance of each variable (feature importance). Second, the classification results with the best model will select the most influential indicators to then be used as the basis for clustering using DBSCAN and HDBSCAN. These clustering methods were chosen because they do not require determining the number of initial clusters, are robust to outliers, and are able to handle irregular cluster shapes, making them very suitable for data between districts/cities that are heterogeneous and contain noise [12]. This study highlights nine key indicators of food security, including the percentage of the poor population, access to clean water, stunting prevalence, and life expectancy at birth. This study aims to build a machine learning-based food vulnerability prediction model and examine regional characteristics to formulate strategies to support the achievement of SDGs-2 "Zero Hunger."

This research is expected to make a significant academic contribution to the development of data-driven models for food security analysis and provide policymakers with practical insights for designing more targeted food intervention strategies. Strengthening the early warning system for food-insecure areas will help Indonesia accelerate its progress toward the SDGs' Zero Hunger target while enhancing national resilience against future food crises.

2. Research Method

This study was conducted through several main stages, including data collection, data preprocessing, classification modeling, clustering analysis, and model evaluation. The purpose of these stages is to develop predictive models and identify patterns of food security vulnerability in each region. Each step in the methodology is described in detail in the following subsections.

2.1. Data Collection

This study utilizes data obtained from two primary sources: the Indonesian Data Portal (data.go.id) and Statistics Indonesia (bps.go.id). These sources provide official and open-access data that are relevant for analyzing food security classification and clustering in Indonesia. The dataset encompasses 514 regencies and cities throughout Indonesia, spanning six years (2018–2023) and yielding a total of 3,084 observations. The collected data includes various social, economic, and health indicators related to food security, as presented in table 1 below:

Table 1. Indicators Employed in the Study.

No.	Variable	Indicator	Description
1.	X1	NCPR (Normative Consumption per Capita Ratio)	Ratio of nominal per capita consumption to normative standard
2.	X2	Poverty Rate (%)	Percentage of population living below the poverty line
3.	X3	Food Expenditure (%)	Proportion of household expenditure spent on food
4.	X4	Without Electricity (%)	Percentage of population without access to electricity
5.	X5	Without Clean Water (%)	Percentage of population without access to clean water
6.	X6	Female Mean Years of Schooling	Average years of schooling for females
7.	X7	Health Worker Ratio	Number of health workers per capita
8.	X8	Life Expectancy	Average estimated life expectancy

No.	Variable	Indicator	Description
9.	X9	Prevalence of Undernourishment (%)	Percentage of undernourished population
10.	X10	Population Size	Total population per administrative region
11.	X11	Undernourished Population	Proportion of population suffering from undernourishment
12.	X12	Stunting (%)	Percentage of children under five experiencing stunting
13.	Y	Food Security Status	Classification of regional food security status

2.2. Data Preprocessing

Before conducting the analysis, the collected data was cleaned and prepared through a data preprocessing stage. This phase included data cleaning (handling missing values, dealing with outliers, and removing duplicate entries) and data normalization. Outlier handling in this phase involved two main techniques: the Interquartile Range (IQR) method and the Isolation Forest algorithm.

The IQR method identifies outliers based on values falling outside the lower and upper quartiles [13]. In contrast, Isolation Forest is an ensemble-based algorithm that detects outliers by isolating anomalies using random tree structures, observations that are isolated more quickly are considered outliers [14].

2.3. Classification Modeling

Classification modeling is employed to predict the food security status of each region based on the indicators presented in Table 1. This study compares the performance of three decision tree-based classification models: Random Forest, XGBoost, and LightGBM.

2.3.1 Random Forest

Random Forest improves model accuracy by reducing the correlation between individual trees [15]. The trees are constructed independently to ensure minimal dependence on one another. In classification tasks, the final decision of a Random Forest model is made using a majority voting mechanism [16].

Several hyperparameters in Random Forest directly influence model performance [17]. These include max depth, which controls the maximum depth of each tree, and max features, which specifies the number of features considered at each split. Additionally, the minimum sample split determines the minimum number of samples required to split an internal node. In contrast, min samples leaf sets the minimum number of samples needed to be at a leaf node to maintain model simplicity. Finally, n_estimators refers to the number of trees constructed in the forest.

2.3.2 XGBoost

XGBoost (Extreme Gradient Boosting) combines multiple decision trees through a gradient boosting approach, iteratively correcting the errors of previous trees to enhance prediction accuracy [18]. It is particularly effective due to its regularization capabilities, which help control overfitting, and its ability to handle missing values automatically, making it well-suited for managing large datasets efficiently [19]. In addition, XGBoost employs a shrinkage mechanism to regulate the contribution of each newly added tree during the learning process, thereby enhancing model stability.

Tunable hyperparameters in XGBoost include colsample_bytree (the proportion of features used per tree), gamma (the minimum loss reduction required to make a split), max_depth (the maximum tree depth to capture complex patterns while preventing overfitting), learning_rate (the step size used for weight updates), and n_estimators (the number of trees in the ensemble) [20].

2.3.3 LightGBM

LightGBM (Light Gradient Boosting Machine) is an ensemble-based algorithm that implements boosting by combining multiple weak learners to form a strong predictive model [21]. One of the main advantages of LightGBM over traditional boosting algorithms is its use of histogram-based decision tree learning and a leaf-wise growth strategy, which significantly accelerates the training process while reducing memory consumption. A study by [8] demonstrated that LightGBM can outperform XGBoost in both speed and accuracy, particularly on large-scale datasets.

Several hyperparameters in LightGBM can be adjusted to optimize model performance, including num_leaves (the maximum number of leaves per tree), max_depth (the maximum depth of the tree), learning_rate (the step size for updating weights), n_estimators (the number of trees built), and min_data_in_leaf (the minimum number of data points in a leaf). Additionally, feature_fraction and bagging_fraction serve as regularization parameters to help prevent overfitting.

2.4 Clustering Modeling

As an advanced modeling approach to gain deeper insights, clustering analysis is conducted to group regions based on the characteristics of their vulnerability. The clustering results are then interpreted to uncover specific patterns among the identified regional groups.

2.4.1 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm capable of detecting arbitrarily shaped clusters and identifying noise or outliers [12]. The core principle of DBSCAN relies on two main parameters: ε (epsilon), which defines the maximum distance between two points to be considered density-reachable, and minPts, the minimum number of neighboring points required for a point to be classified as a core point. The distance between data points is typically calculated using the Euclidean Distance function, which is formulated as follows:

$$d_{ij} = \sqrt{\sum_{a=1}^{p}\left(x_{ia} - x_{ja}\right)^2} \text{ for } i = 1, \dots, n; j = 1, \dots, n \tag{1}$$

in the formula above, x_{ia} represents the a-th variable of the i-th object.

The advantage of DBSCAN lies in its ability to detect clusters of arbitrary shapes, its computational efficiency compared to other algorithms such as CLARANS, and the fact that it does not require predefining the number of clusters, as is necessary with K-Means [22]. However, the effectiveness of DBSCAN depends heavily on the appropriate selection of ε and minPts values. Improper parameter settings can significantly affect the accuracy and coverage of the resulting clusters.

2.4.2 HDBSCAN

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm developed as an improvement over DBSCAN, designed to detect more complex clustering patterns. Unlike DBSCAN, which relies on a distance-based parameter (epsilon), HDBSCAN requires only a single primary parameter, minimum points (minPts), to determine the minimum number of points that define a dense region [23].

The algorithm uses mutual reachability distance as a metric to compute the distance between points, incorporating the distance to the K-Nearest Neighbors (KNN). It allows HDBSCAN to identify cluster structures with varying densities effectively. Through a hierarchical approach, HDBSCAN constructs a cluster tree and subsequently condenses this structure to produce the most stable, flat clusters.

2.5. Model Evaluation

The performance of the classification models in this study is evaluated using several metrics that assess accuracy, balance, and predictive precision, particularly in the context of imbalanced class distributions. The primary metric employed is the Weighted F1 Score, which measures the harmonic mean of Precision and Recall while assigning weights based on the proportion of each class. The F1 Score serves as a critical indicator, as it balances false positives and false negatives, especially important in scenarios where one class is dominant [24]. The F1 Score is formulated as follows:

$$F1\ Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall} \tag{2}$$

The formulas for Precision and Recall, respectively, are defined as follows:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN} \tag{3}$$

where TP denotes True Positives, FP is False Positives, and FN is False Negatives. A high F1 Score indicates that the model consistently detects positive samples with minimal error. Additionally, Precision (Weighted) and Recall (Weighted) evaluate the proportion of correct predictions while taking into account imbalanced class distributions. Precision focuses on the correctness of positive predictions, whereas Recall emphasizes the model's ability to detect all actual positive samples.

Meanwhile, Balanced Accuracy measures the average accuracy across all classes, thereby addressing the bias commonly found in standard accuracy metrics when dealing with imbalanced datasets [25]. Balanced Accuracy is formulated as follows:

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{4}$$

with TN denoting True Negatives.

Clustering performance is evaluated using three commonly adopted metrics: Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, which are widely used to assess clustering quality [26]. Silhouette Score measures how well each object lies within its assigned cluster and how distinct it is from other clusters. Davies-Bouldin Score evaluates the similarity between clusters, where lower values indicate better-separated and more distinct clusters. Calinski-Harabasz Score assesses the ratio of between-cluster dispersion to within-cluster dispersion, with higher values indicating more well-defined cluster structures.

3. Result and Discussion

This section presents the findings obtained from the application of machine learning techniques to the food security dataset across Indonesian districts/cities. The results are systematically organized, beginning with exploratory data analysis to understand the dataset characteristics, followed by the performance evaluation of classification and clustering models. The implications of these findings are also discussed in terms of their relevance to food security policy and SDGs 2 targets.

3.5. Data Exploration

Exploratory data analysis (EDA) was conducted to evaluate the quality and characteristics of the dataset, with a particular focus on detecting outliers and assessing class distribution. Outliers and imbalanced data can lead to biased model results and reduce the reliability of the evaluation process. Figure 1 presents the boxplot for each variable, showing extreme values that may hinder model generalization.

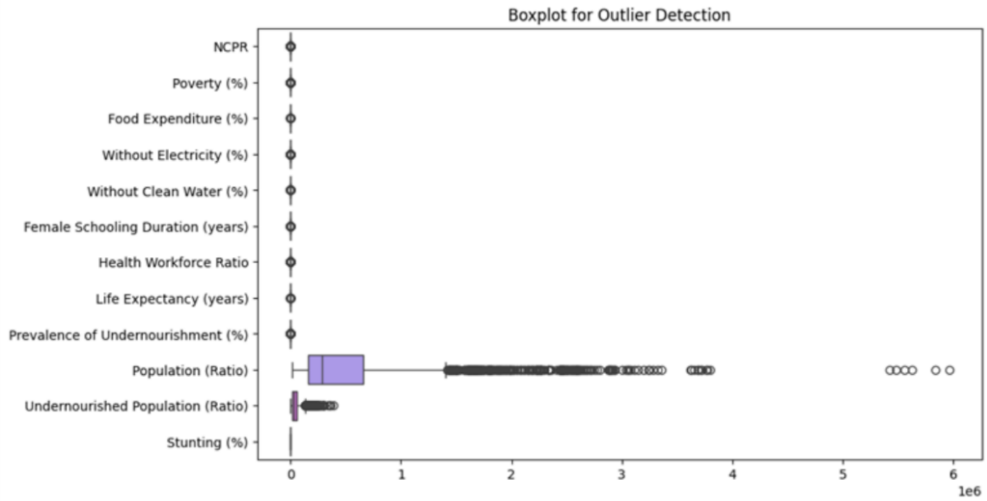

Figure 1. Boxplot Outlier Detection.

Figure 1 shows the boxplots of all variables using the Interquartile Range (IQR) method. It can be observed that the variables Population (Ratio) and Undernourished Population (Ratio) have outlier values that may affect the evaluation process. To provide additional information, the number of outliers was compared between the IQR and Isolation Forest methods, as shown in Figure 2.

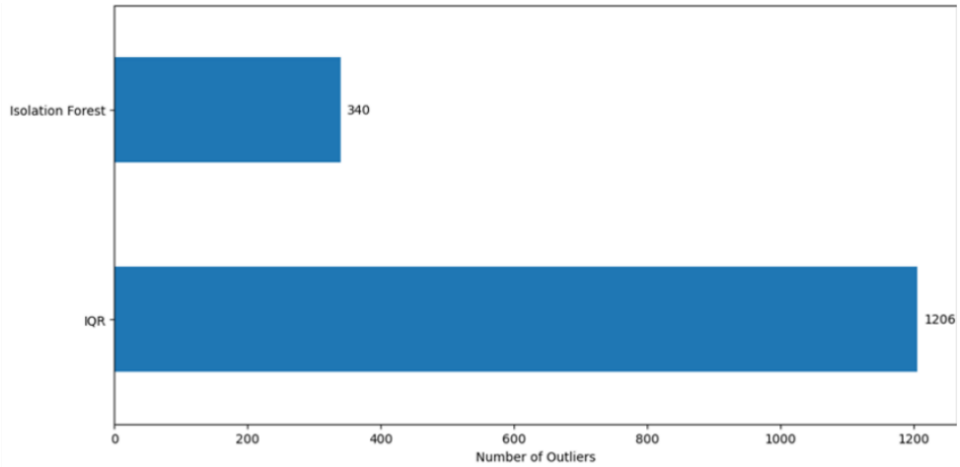

Figure 2. Comparison of the Number of Outliers.

Figure 2 compares the number of outliers detected by the IQR method as a statistical approach and the Isolation Forest algorithm as an algorithmic approach. The IQR method detected 1,206 outliers, while the Isolation Forest method identified 340. Outliers detected using the IQR method were primarily associated with regions experiencing poor food security conditions, such as high poverty and stunting rates, as well as limited access to clean water and electricity. However, several regions exhibited positive extreme values, indicating exceptionally good conditions. These outliers overall reflect the structural disparities across Indonesian regions. This difference occurs because the IQR method operates univariately, flagging extreme values for each indicator independently, whereas Isolation Forest adopts a multivariate approach that identifies only those observations with extreme patterns across multiple indicators simultaneously. This distinction highlights the differing sensitivities of the two methods in detecting outliers. Although the extreme data points were removed during the model training stage, the substantive interpretation of these extreme regions was retained in the discussion to ensure that relevant policy insights were not overlooked.

After addressing data irregularities through the detection and removal of outliers, the next preprocessing step focused on handling class imbalance in the target variable. Figure 3 illustrates the distribution of Food Security Status before and after applying the SMOTE technique.

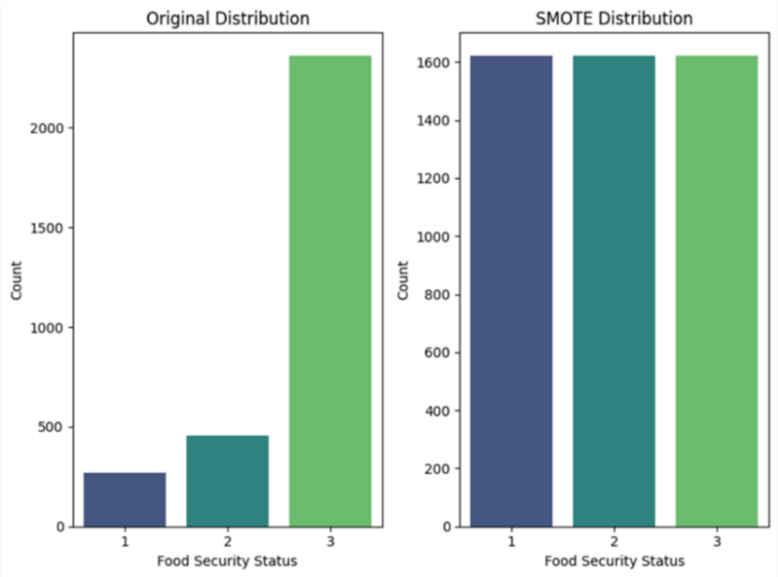

Figure 3. Food Security Status Distribution.

Figure 3 illustrates the class distribution of the target variable Food Security Status before and after applying the SMOTE technique. In the original dataset (left), the distribution was imbalanced, with class 3 having a considerably higher number of observations compared to classes 1 and 2. SMOTE was then applied to generate synthetic samples for the minority classes by interpolating between existing observations. As shown in the correct chart, the dataset became more balanced, with each class represented by approximately 1,600 observations.

3.6. Comparison of Machine Learning Results in Classifying Food-Insecure Regions

To evaluate the performance of various machine learning algorithms in Food Security Status classification, several experiments were conducted using different combinations of outlier handling and class imbalance treatment techniques. The models compared in this study include Random Forest (RF), XGBoost (XGB), and LightGBM (LGBM), with outlier detection handled via either the Interquartile Range (IQR) or Isolation Forest (IF), and class imbalance addressed using the SMOTE oversampling technique. Table 2 summarizes the classification performance of each model configuration based on accuracy, precision, recall, and F1-score.

Table 2. Classification Performance of Machine Learning Models for Food Security Status.

.Technique	Accuracy	Precision	Recall	F1 Score
LGBM IQR - No Handling	0.984043	0.983906	0.984043	0.983788
XGB IQR SMOTE	0.980496	0.980334	0.980496	0.980358
LGBM IQR - SMOTE	0.976950	0.977421	0.976950	0.977132
XGB IQR - No Handling	0.975177	0.975060	0.975177	0.974676
RF IQR - SMOTE	0.971631	0.971977	0.971631	0.971780
RF IQR - No Handling	0.971631	0.971162	0.971631	0.970979
XGB Outlier - No Handling	0.967603	0.967057	0.967603	0.966493
LGBM IF - SMOTE	0.966019	0.965878	0.966019	0.965867
LGBM Outlier - No Handling	0.966523	0.965963	0.966523	0.965316
LGBM IF Outlier - No Handling	0.964806	0.964030	0.964806	0.963921

As shown in table 2, a comparison of classification models demonstrates the superiority of decision tree-based ensemble algorithms, particularly LGBM, in handling heterogeneous socio-economic data. LGBM consistently achieved the highest accuracy and F1-score when combined with IQR-based outlier handling and class balancing using SMOTE. This superior performance stems from LGBM's ability to efficiently model non-linear relationships and capture complex interactions between features [8]. In contrast, Random Forest, while known to be robust to overfitting, exhibited slightly lower performance. These results align with previous research [7], which showed that boosting algorithms often outperform bagging-based models when feature heterogeneity and interaction effects between factors are strong.

The XGB model combined with IQR and SMOTE also demonstrated strong performance, highlighting the potential benefits of class balancing in certain settings. However, LGBM's histogram-based optimization and leaf-wise growth strategy provided better computational efficiency and predictive accuracy. From a data perspective, socioeconomic food security indicators, such as poverty rates, access to clean water, electricity, and life expectancy, tend to exhibit skewed and non-linear patterns. LGBM's ability to provide adaptive weights during the iterative boosting process allows it to better capture these variations than traditional Decision Tree and Random Forest algorithms. Furthermore, IQR-based preprocessing methods have proven more stable than algorithmic outlier detection methods like Isolation Forest, likely because anomalies in this dataset are more extreme in univariate terms than complex multivariate noise.

3.7. Cluster Analysis of Food Insecurity Regions

This section presents the results of the clustering analysis conducted to identify groups of regions with similar food security characteristics. The study utilized the top six features identified as having the highest importance based on the tuned LightGBM classification model. Density-based clustering approaches, specifically DBSCAN and HDBSCAN, were applied to uncover natural groupings within the data.

3.3.1 Selection of Key Features for Clustering

Before the clustering process, feature selection was performed to determine the most influential variables based on the feature importance scores from the LightGBM model. As shown in figure 4, the six top-ranked features selected for clustering analysis are as follows: NCPR (Normative Consumption per Capita Ratio), Poverty Rate (%), Without Clean Water (%), Life Expectancy (years), Population Size, and Without Electricity (%).

These features represent critical dimensions in determining the food security level of a region. Their selection ensures that the resulting clusters are based on the most relevant and impactful characteristics.

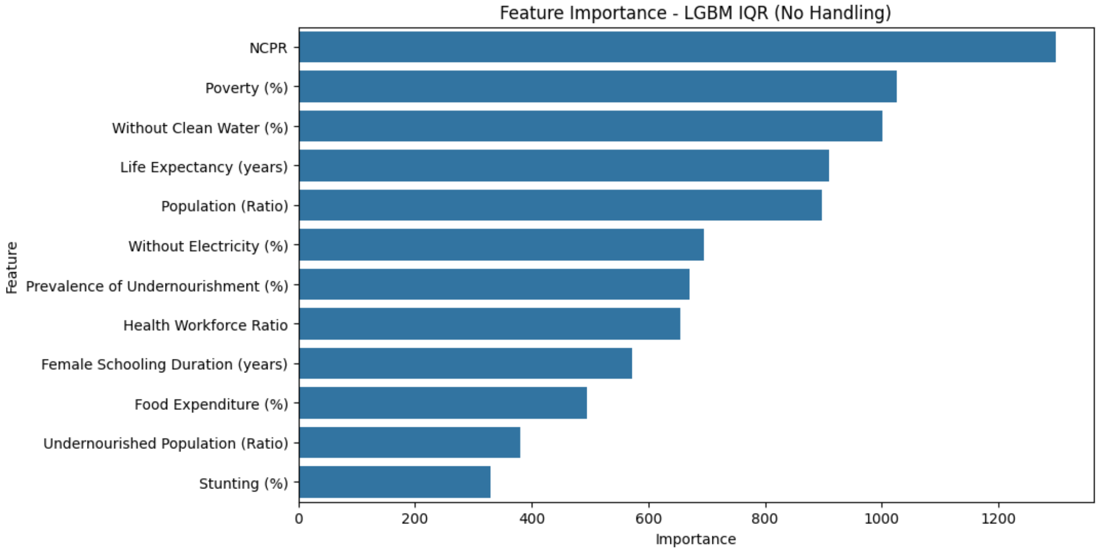

Figure 4. Feature Importance of the LightGBM Model.

3.3.2 Clustering Model Tuning and Evaluation

The 2023 data, filtered based on the six selected key features and standardized using the StandardScaler, were subsequently used as input for the clustering models. Parameter tuning was performed for each model (DBSCAN and HDBSCAN) to optimize the cluster structure based on the Silhouette Score.

Table 3. Evaluation Metrics for Clustering Models.

Model	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
DBSCAN	0.5639	2.1290	74.2002
HDBSCAN	0.5517	1.8127	70.8905

As shown in table 3, both DBSCAN and HDBSCAN demonstrated adequate performance in identifying regional clusters based on food security indicators. For DBSCAN, the optimal parameter configuration was obtained at an epsilon (ε) value of 1.90 with a fixed min_samples of 5, resulting in a Silhouette Score of 0.5639, indicating adequate cluster separation. Meanwhile, HDBSCAN with min_cluster_size = 5 produced a slightly lower Silhouette Score (0.5517) but a slightly better Davies–Bouldin value, indicating higher internal cluster cohesion.

Overall, the comparison of the two models demonstrates complementary strengths. DBSCAN has an advantage in cluster separability, making it more suitable for datasets with clear regional boundaries. Conversely, HDBSCAN exhibits stronger internal consistency and excels on data with a hierarchical structure or layered density patterns. Given that this study's dataset displays distinct inter-regional differences, rather than a stratified density pattern, DBSCAN was deemed the most appropriate model for the final cluster analysis. This decision aligns with the findings of [27], which confirmed DBSCAN's superiority in detecting clusters of arbitrary shape and in situations with high inter-cluster density contrast. Therefore, further analysis and policy interpretation in this study are based on the DBSCAN cluster results using the six key features identified in the classification stage.

3.3.3 Cluster Visualization

To visualize the structure of the resulting clusters, the UMAP (Uniform Manifold Approximation and Projection) technique was employed to reduce the dimensionality of the data to two components. This visualization enables the representation of clusters in a two-dimensional space, as illustrated in figure 5 and figure 6.

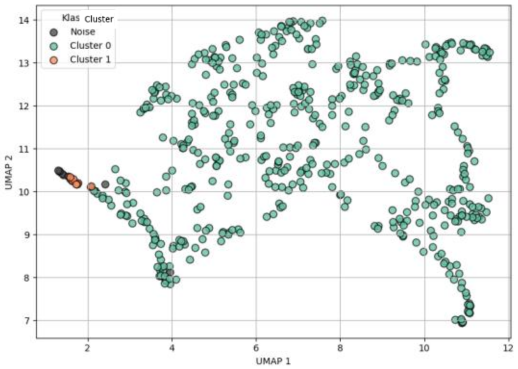

Figure 5. Visualization of Food Vulnerable Area Clusters using DBSCAN.

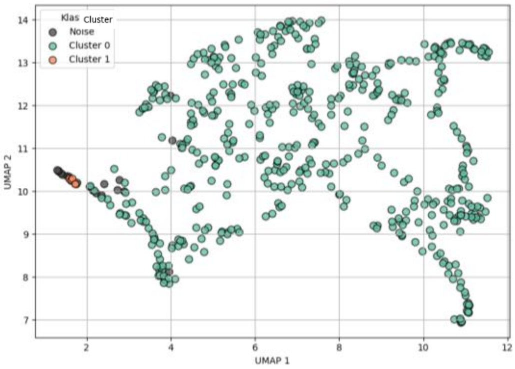

Figure 6. Visualization of Food Vulnerable Area Clusters using HDBSCAN.

In this visualization, each point represents a single region, and different colors indicate cluster membership. Gray-colored points represent data identified as noise or outliers by the algorithm, referring to regions that do not fit into the main densely formed clusters.

3.3.4 Characteristics of Food-Insecure Regions Using DBSCAN

A deeper analysis was conducted by examining the average statistics of each feature within the clusters identified by DBSCAN (Top 6 Features). The results are summarized in **table 4.**

Table 4. Average Feature Statistics per Cluster (DBSCAN – Top 6 Features)

DBSCAN Claster	NCPR	Poverty (%)	Without Clean Water (%)	Life Expectancy (Years)	Population Size (Ratio)	Without Electricity (%)
-1 (Noise)	3.6775	26.7375	64.8083	63.8167	1013888	33.16
0	1.1645	10.8801	28.8202	70.1588	538181	1.1327
1	5	35.5533	86.8267	65.4911	134059	27.0889

Based on table 4, the post-cluster analysis shows clear differences in socio-economic and infrastructure characteristics among the three clusters formed (Cluster 0, Cluster 1, and Noise). Each cluster has distinct policy implications and can serve as a basis for developing more targeted intervention strategies. A complete interpretation of each cluster is provided below.

- **Cluster 0 (Food Security)**: This cluster is the largest, encompassing 493 regions. Regions within this cluster show low average values for negative indicators, such as poverty rate (10.88%), percentage of households without access to clean water (28.82%), and percentage of households without electricity (1.13%). Conversely, positive indicators such as life expectancy (70.16 years) are relatively high. These characteristics indicate that regions within this cluster are generally urban and have stable economies with diverse livelihood structures, as well as good food security.

- **Cluster 1 (Highly Food Insecure)**: This cluster consists of nine regions facing severe limitations on various welfare indicators, such as poverty rates above 35%, low access to clean water (86%), and limited access to electricity (27%). Life expectancy (65.49 years) is also lower than that of Cluster 0. This condition reflects regions with extreme levels of food vulnerability due to limited basic infrastructure and low household socio-economic capacity.
- **Cluster -1 (Noise or Outliers Regions):** A total of 12 regions were identified as noise. These regions exhibit diverse and extreme characteristics, preventing them from being densely grouped into any cluster. The average indicators for this cluster fall between Clusters 0 and 1 but tend to be worse than Cluster 0, for instance, poverty (26.74%), lack of clean water (64.81%), and life expectancy (63.82 years). Some areas in this cluster have moderate poverty levels despite poor infrastructure access, while others show the opposite.

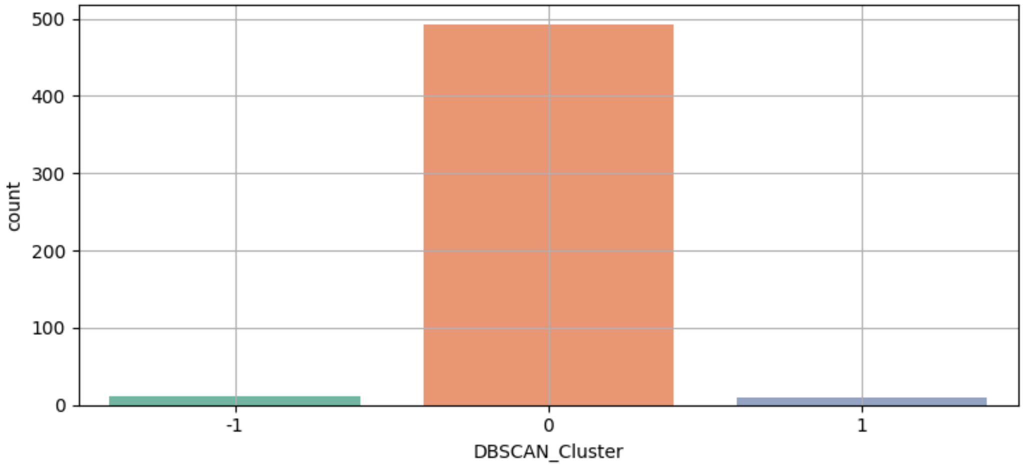

Figure 7. Distribution of Number of Members per DBSCAN Cluster.

3.3.5 Policy Implications

The clustering results provide a more nuanced understanding of the spectrum of food insecurity across Indonesia. Each cluster represents a distinct risk profile, enabling the formulation of more targeted and efficient policy interventions. For regions in Cluster 0 (Food Secure), policy focus can shift from crisis response to strengthening long-term food security foundations, promoting local economic development, and enhancing sustainable quality of life. In line with study [28] which states that economic and social infrastructure (including access to electricity, transportation, education) as well as household economic conditions significantly influence food security. On the other hand, study [29] also shows that income is a determinant of a region's food security. Policies for regions included in this cluster should focus on strengthening long-term food security through efforts such as sustainable agricultural intensification, strengthening local food systems, and innovation in the food supply chain to maintain production and distribution stability. In addition, it is important to encourage food diversification based on local resources and the development of urban farming to strengthen access to nutritious food and reduce dependence on certain commodities [30] [31]. This approach can also be strengthened through digitalization of food distribution and logistics systems, implementation of climate risk early warning systems, and strengthening regional food institutions to be more adaptive to economic and environmental dynamics [32]. These steps are in line with the direction of the national food security policy which emphasizes the transformation of a sustainable and inclusive food system through increasing productivity, distribution efficiency, and strengthening food institutions at the regional level [33].

In contrast, regions in Cluster 1 (Highly Food Insecure) require urgent and comprehensive interventions, with a focus on improving access to basic infrastructure such as clean water and electricity. Additionally, more intensive programs for poverty alleviation and public health improvement are needed to address the high prevalence of malnutrition. According to the Food and Agriculture Organization (FAO), the dimensions of food vulnerability are not only influenced by food availability, but also by economic access factors and basic infrastructure that support sustainable food distribution and consumption [34]. Meanwhile, the United Nations Development Programme (UNDP) emphasizes that multidimensional poverty, which includes deprivation in aspects of education, health, and living standards, exacerbates the risk of food insecurity at the household level [35]. Policies for regions within this cluster need to prioritize budget allocations directed at meeting basic needs, particularly rural infrastructure development (clean water, sanitation, and electricity), nutrition-based social protection programs, and food subsidies for highly vulnerable households. This approach aligns with the 2020-2024 National Food and Nutrition Security Strategic Policy, which emphasizes the importance of integrating food security programs with social protection in food-insecure areas [33]. Furthermore, identifying areas within Cluster 1 can serve as the basis for more targeted fiscal planning, for example through the special allocation fund for basic infrastructure development and improving community nutrition.

Regions identified as noise (-1) demand in-depth, case-by-case analysis to uncover the specific underlying factors contributing to their vulnerability. These uniquely characterized areas require local and contextual interventions, in accordance with the principles of place-based policy design [36]. Local governments in these areas need to be empowered to conduct more in-depth micro-analysis to design interventions tailored to the specific socio-economic conditions of their areas. This understanding is crucial for designing highly tailored and context-specific interventions.

4. Conclusion

This study develops a hybrid machine learning framework that integrates classification and clustering techniques to identify regions based on their food security conditions across Indonesia. The LightGBM model, combined with IQR-based outlier handling and SMOTE balancing, achieved the highest predictive performance (Accuracy and F1-score = 0.984). Feature importance analysis revealed six dominant indicators that shape food vulnerability, namely NCPR (Normative Consumption per Capita Ratio), Poverty Rate (%), Without Clean Water (%), Life Expectancy (years), Population Size, and Without Electricity (%). These indicators were then used for clustering using DBSCAN, which classified regions into three distinct groups: Cluster 0 (Food Security), Cluster 1 (Highly Food Insecure), and Cluster −1 (Outlier or Noise). Post-clustering analysis revealed clear socio-economic and infrastructure disparities among these clusters. Cluster 0 (Food Security) regions, which are generally urban and economically stable, must shift their policy focus from short-term responses to strengthening long-term food system resilience. Strategies such as promoting sustainable agriculture, encouraging local food diversification, and enhancing innovation in food distribution systems can help maintain stability and resilience to future disruptions. In contrast, Cluster 1 (Highly Food Insecure) regions exhibit severe deficits in basic infrastructure and socio-economic capacity. These regions require urgent, coordinated, and multisectoral interventions aimed at improving access to clean water, electricity, and nutrition, while addressing structural poverty. Identification of these clusters provides valuable guidance for more targeted resource allocation, ensuring that investment and assistance are prioritized for the most vulnerable regions. Meanwhile, Cluster −1 (Outlier Regions) exhibit unique and diverse characteristics that cannot be generalized to the broader typology of regions. These regions require localized, context-specific strategies, guided by micro-level analysis to understand their specific challenges and opportunities. Strengthening local data systems and integrating regional findings into a national monitoring framework will improve policy coordination and ensure that decision-making is based on accurate, on-the-ground insights. Overall, this hybrid framework demonstrates that combining classification and clustering not only improves analytical precision but also enhances policy relevance. By translating technical findings into actionable insights, such as prioritizing infrastructure development and improving data integration for monitoring, this study contributes to data-driven policymaking for equitable and resilient food security. Aligning these findings with national development priorities and the Sustainable Development Goals (SDG 2: Zero Hunger and SDG 10: Reduced Inequality) ensures that data-driven governance supports Indonesia's long-term vision for an inclusive and sustainable food system. While the results of this study demonstrate promising performance, several limitations should be acknowledged. First, the analysis uses cross-sectional data, which cannot capture the temporal dynamics of food security over time. Future research could use panel or time-series data to study regional change and resilience longitudinally. Second, although the hybrid framework used successfully integrates classification and clustering methods, its performance still depends on the representativeness and quality of the data, especially in remote areas that still experience data gaps.

References

[1] Economist Intelligence Unit, *Global Food Security Index 2022: Building Resilience in the Face of Crisis.* London: The Economist Group, 2022.

[2] Badan Pangan Nasional, Statistik Ketahanan Pangan Indonesia 2022. Jakarta: BPN, 2022.

[3] United Nations, “Transforming our world: The 2030 Agenda for Sustainable Development.” Accessed: July 26, 2025. [Online]. Available: https://sdgs.un.org/goals

[4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, 2012.

[5] L. N. Aziza, R. Y. Astuti, B. A. Maulana, and N. Hidayati, “Penerapan algoritma K-Nearest Neighbor untuk klasifikasi ketahanan pangan di Provinsi Jawa Tengah: Application of The K-Nearest neighbor algorithm for food security classification in Central Java Province,” *MALCOM*, vol. 4, no. 2, pp. 404–412, Feb. 2024, doi: 10.57152/malcom.v4i2.1201.

[6] A. L. Azhar, S. Suliyanto, N. Chamidah, E. Ana, and D. Amelia, “Pemodelan Indeks Ketahanan Pangan di Indonesia berdasarkan pendekatan regresi logistik ordinal data panel efek acak,” *JKN*, vol. 29, no. 2, p. 166, Sept. 2023, doi: 10.22146/jkn.86511.

[7] Z. Rahmatinejad et al., “A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department,” *Sci Rep*, vol. 14, no. 1, p. 3406, Feb. 2024, doi: 10.1038/s41598-024-54038-4.

[8] G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree”, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*,p 3146–3154, 2017.

[9] N. Bujiku Sende, S. Saha, L. Ruganzu, and S. Kar, “Prediction of multidimensional poverty status with machine learning classification at household level: Empirical evidence from Tanzania,” *IEEE Access*, vol. 13, pp. 23461–23471, 2025, doi: 10.1109/ACCESS.2025.3537807.

[10] R. F. Sinaga, M. A. Prabukusumo, and J. Manurung, “Comparison of K-means clustering with hierarchical agglomerative clustering for the analysis of food security of rice sector in Indonesia,” *J. Intell. Decis. Support Syst*., vol. 8, no. 1, 2025.

[11] Syahraini and Y. Rizal, “Perbandingan K-Means dan K-Medoids Clustering dalam pengelompokan kabupaten/kota berdasarkan produksi tanaman pangan di Provinsi Sumatera Barat,” vol. 8, 2024.

[12] M. Tanzil Furqon and L. Muflikhah, “Clustering the potential risk of tsunami using density-based spatial clustering of spplication with noise (DBSCAN),” *JEEST*, vol. 3, no. 1, pp. 1–8, July 2016, doi: 10.21776/ub.jeest.2016.003.01.1.

[13] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, “An outliers detection and elimination framework in classification task of data mining,” Decision Analytics Journal, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.

[14] N. Z. Fanani, A. G. Sooai, K. Khamid, and F. Y. Rahmanawati, “Two stages outlier removal as pre-processing digitizer data on fine motor skills (FMS) classification using covariance estimator and isolation forest,” *IJIES*, vol. 14, no. 4, pp. 571–582, Aug. 2021, doi: 10.22266/ijies2021.0831.50.

[15] F. Sohil, M. U. Sohali, and J. Shabbir, “An introduction to statistical learning with applications in R: by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7,” *Statistical Theory and Related Fields*, vol. 6, no. 1, pp. 87–87, Jan. 2022, doi: 10.1080/24754269.2021.1980261.

[16] I. D. Mienye and Y. Sun, “A survey of ensemble learning: Concepts, algorithms, applications, and prospects,” *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[17] P. Probst, M. Wright, and A.-L. Boulesteix, “Hyperparameters and tuning strategies for random forest,” *WIREs Data Min & Knowl*, vol. 9, no. 3, p. e1301, May 2019, doi: 10.1002/widm.1301.

[18] Z. A. Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, “eXtreme gradient boosting algorithm with machine learning: a Review,” *ACAD J NAWROZ UNIV*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.

[19] S. Sharma and N. Joshi, “A fusion approach to detect sarcasm using NLTK models BERT and XG Boost,” *JIOS*, vol. 45, no. 4, pp. 981–990, 2024, doi: 10.47974/JIOS-1621.

[20] Y. Jiang, G. Tong, H. Yin, and N. Xiong, “A Pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters,” *IEEE Access*, vol. 7, pp. 118310–118321, 2019, doi: 10.1109/ACCESS.2019.2936454.

[21] C. Lokker et al., “Boosting efficiency in a clinical literature surveillance system with LightGBM,” *PLOS Digit Health*, vol. 3, no. 9, p. e0000299, Sept. 2024, doi: 10.1371/journal.pdig.0000299.

[22] A. Saputra and R. Yusuf, “Perbandingan algoritma DBSCAN dan K-MEANS dalam segmentasi pelanggan pengguna transportasi publik transjakarta menggunakan metode RFM: Comparison of the DBSCAN and K-MEANS algorithms in segmenting customers using public transportation of Transjakarta using the RFM method,” *MALCOM*, vol. 4, no. 4, pp. 1346–1361, July 2024, doi: 10.57152/malcom.v4i4.1516.

[23] G. Stewart and M. Al-Khassaweneh, “An implementation of the HDBSCAN* clustering algorithm,” *Applied Sciences*, vol. 12, no. 5, p. 2405, Feb. 2022, doi: 10.3390/app12052405.

[24] G. Naidu, T. Zuva, and E. M. Sibanda, “A Review of evaluation metrics in machine learning algorithms,” in *Artificial Intelligence Application in Networks and Systems*, R. Silhavy and P. Silhavy, Eds., Cham: Springer International Publishing, 2023, pp. 15–25. doi: 10.1007/978-3-031-35314-7_2.

[25] A. M. Carrington et al., “Deep ROC Analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, doi: 10.1109/TPAMI.2022.3145392.

[26] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, “Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index evaluation on K-Means algorithm for classifying flood-affected areas in Jakarta,” *JAIC*, vol. 7, no. 1, pp. 89–97, July 2023, doi: 10.30871/jaic.v7i1.4947.

[27] E.-R. Ardelean, R. L. Portase, R. Potolea, and M. Dînșoreanu, “A path-based distance computation for non-convexity with applications in clustering,” *Knowl Inf Syst*, vol. 67, no. 2, pp. 1415–1453, Feb. 2025, doi: 10.1007/s10115-024-02275-4.

[28] D. W. Sari, P. C. A. Yudha, and W. Restikasari, “The Effect of economic and social infrastructure on household food security in Indonesia,” *JEI*, vol. 8, no. 2, pp. 191–201, Dec. 2019, doi: 10.52813/jei.v8i2.4.

[29] A. Akbar, R. Darma, I. M. Fahmid, and A. Irawan, “Determinants of household food security during the COVID-19 pandemic in Indonesia,” *Sustainability*, vol. 15, no. 5, p. 4131, Feb. 2023, doi: 10.3390/su15054131.

[30] M. P. Hutagaol and R. Sinaga, “Pengaruh pendapatan dan harga pangan terhadap diversifikasi pangan di Pulau Jawa,” *Sci J Reflect.*, vol. 5, no. 3, 2022.

[31] A. Abdillah, I. Widianingsih, R. A. Buchari, and H. Nurasa, “Implications of urban farming on urban resilience in Indonesia: Systematic literature review and research identification,” *Cogent Food & Agriculture*, vol. 9, no. 1, p. 2216484, Dec. 2023, doi: 10.1080/23311932.2023.2216484.

[32] D. R. Hakim, A. Rahmiwati, R. Flora, and Novrikasari, “Menjelajahi dinamika pangan di era perubahan iklim terhadap dampak di Indonesia dan proyeksi masa depan: A systematic review,” *RRJ*, vol. 7, no. 3, pp. 1703–1720, Mar. 2025, doi: 10.38035/rrj.v7i3.1411.

[33] Badan Ketahanan Pangan, *Rencana Strategis Ketahanan Pangan Tahun 2020-2024*. BKP, 2020.

[34] FAO, *The State of Food Security and Nutritioon in The World: Transforming Food Systems for Food Security, Improved Nutrition and Affordable Healthy Diets for All*. 2021.

[35] UNDP, *Global Multidimensional Poverty Index 2023*. 2023.

[36] A. Morisson and M. Doussineau, “Regional innovation governance and place-based policies: design, implementation and implications,” *Regional Studies, Regional Science*, vol. 6, no. 1, pp. 101–116, Jan. 2019, doi: 10.1080/21681376.2019.1578257.