Scientific Journal of Informatics
Vol. 12, No. 3, August 2025
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Performance Analysis of Machine Learning Models using RFE Feature Selection and Bayesian Optimization in Imbalanced Data Classification with Shap-Based Explanations

Nurzatil Aqmar[1*], Hari Wijayanto[2], Farit Mochamad Afendi[3]
[1,2,3]Department of Statistics and Data Science, IPB University, Indonesia

Abstract.
Purpose: This research aims to evaluates the performance of Random Forest (RF) and Light Gradient Boosting Machine (LightGBM) models integrated with Recursive Feature Elimination (RFE) for feature selection, Bayesian Optimization (BO) for hyperparameter tuning, and three imbalanced data handling techniques Random Undersampling (RUS), Random Oversampling (ROS), and SMOTENC. Identifying key determinants of household food insecurity in Papua using SHAP for transparent feature interpretation.
Methods: The research used 2022 SUSENAS data from Papua Province. Exploring data composition and variable characteristics, and aggregating individual data into household data. Data were split using random sampling (80% training, 20% testing). Eighteen experimental scenarios were created by combining feature selection or no feature selection, three imbalance handling methods, and default or hyperparameter tuning. RF and LightGBM were evaluated over 50 iterations using accuracy, sensitivity, specificity, and G-Mean, with SHAP applied to the best-performing models for interpretability.
Result: LightGBM achieved the highest accuracy and stability, particularly when combined with SMOTENC and RFE+BO. RF showed better performance in maintaining G-Mean when paired with RUS, with the highest G-Mean (0.756) obtained by RF + BO + RUS. Three-way ANOVA proved that model type, imbalance handling, feature selection, and their interaction significantly affected the G-Mean value. SHAP analysis shows that health, financial, and educational limitations can increase the risk of food insecurity.
Novelty: This research offers a new integration between feature selection, hyperparameter tuning, and imbalanced data handling within an interpretable machine learning framework, thereby providing a robust solution for food vulnerability classification on imbalanced datasets.

Keywords: Random forest, Light gradient boosting machine, Recursive feature elimination, Bayesian optimization, Shapley additive explanations
Received July 2025 / **Revised** September 2025 / **Accepted** October 2025

INTRODUCTION

Classification is a fundamental machine learning task in which data are grouped into predefined categories. Class imbalance can have a detrimental effect on how well classification algorithms work. Because classifiers have a tendency to prefer the dominant class, this difficulty makes it difficult for them to forecast the minority class accurately. This problem can be solved by resampling techniques as SMOTENC, Random Oversampling (ROS), and Random Undersampling (RUS). RUS is beneficial for extremely unbalanced data [1], ROS is appropriate for moderate imbalance [2], and SMOTENC achieves high performance using SVM [3].

Despite these challenges, machine learning remains capable of building accurate predictive models and accelerating data analysis. Random Forest (RF) and LightGBM are two widely recognized classification algorithms in machine learning. RF, an ensemble method based on decision trees and utilizing a bagging approach, is particularly effective in reducing overfitting and delivering robust results [4] . The LightGBM model is an efficient boosting algorithm for large datasets, utilizing leaf-wise tree growth, GOSS, and EFB strategies to accelerate training [5]. LightGBM has outperformed RF and XGBoost in classifying the Human Development Index in Indonesia [6].

*Corresponding author.
Email addresses: nurzatil99aqmar@apps.ipb.ac.id (Aqmar)*, hari@apps.ipb.ac.id (Wijayanto), fmafendi@apps.ipb.ac.id (Afendi)
DOI: 10.15294/sji.v12i3.31459

RF excels in prediction stability, while LightGBM efficiently handles large-scale features. The performance of both depends heavily on feature selection, so the presence of irrelevant or redundant variables can reduce model performance, increase the risk of overfitting, and make it difficult to interpret the results. Irrelevant features can reduce accuracy and increase complexity, so feature selection is necessary to improve model interpretability and efficiency [7]. Machine learning-based feature selection methods are generally classified into three types: filter, embedded, and wrapper methods [8]. The wrapper method tightly integrates feature selection with model training by directly evaluating the impact of various feature subsets on model performance to identify the optimal feature combination [9]. Recursive Feature Elimination (RFE) is a variable selection method that adopts a wrapper approach first developed by [10]. This technique works by gradually removing less relevant variables and building models based on the remaining variable subsets. By leveraging the performance of machine learning algorithms, RFE identifies the most significant combinations of variables in predicting the target variable [11]. The use of RFE requires consideration of two factors: the number of variables retained and the machine learning algorithm.

In addition to features, hyperparameter selection also affects model performance. Methods such as babysitting, grid search, and random search have limitations, especially in complex models [12]. Bayesian Optimization is a more efficient alternative because it estimates the optimal configuration based on previous results [13]. Through their study, [14] found that feature selection methods with hyperparameters tuned using Bayesian optimization often produce better recall rates, and further transcriptomic data analysis showed that feature selection guided by Bayesian optimization can improve the accuracy of disease risk prediction models. Previous research conducted by [15] showed that the combination of Random Forest, SMOTE, RFE, and PSO successfully increased the accuracy of student study period predictions from 55% to 81%, as well as improving other metrics such as precision, recall, and f1-score. This study only uses one classification method and does not compare several balancing techniques, so a comparison of combinations is needed to obtain the optimal method.

The complexity of black-box machine learning algorithms makes model interpretation difficult [16]. One way to interpret models is through feature importance, such as the SHAP method, which can provide global and local explanations of feature contributions [12]. Previous research by [17] used CatBoost + SHAP to determine which variables most influenced food insecurity in households. Variables such as the head of household's education, type of flooring/walls, source of drinking water, sanitation, and family members who smoke were among the significant ones. For example, research on food insecurity has been conducted in previous studies [18] , which shows that low education and social capital influence food insecurity in 134 countries. Indonesia is committed to 17 SDG targets, one of which is measured using the Food Insecurity Experience Scale (FIES) [19]. BPS data shows that 4.85% of the population experienced moderate to severe food insecurity in 2022. The 2022 Food Security Index (IKP) data recorded that Papua Province was categorized as "Vulnerable" with many households experiencing limited access to food [20].

This study focuses explicitly its analysis on one region, Papua Province, which was selected due to its high level of food vulnerability in Indonesia and high class imbalance, with the dominance of the majority class potentially reducing classification performance in the minority class. The study aims to evaluate the performance of Random Forest and LightGBM models combined with the Recursive Feature Elimination (RFE) feature selection method, hyperparameter tuning using Bayesian Optimization (BO), and class imbalance handling using RUS, ROS, and SMOTENC. Food insecurity data from Papua Province will be used to build and compare various combinations of models and techniques. The best model will be evaluated using the G-Mean metric, which is sensitive to class imbalance, and will examine important variables in predicting food-insecurity households in Papua using Shapley Additive Explanation (SHAP) to provide a more transparent explanation of each feature's contribution to the model's prediction and to explain the most influential factors in determining food-insecurity households.

METHODS

The data used will undergo a preparation process before modeling, which includes exploring the composition and characteristics of each variable, aggregating individual variables into household variables, sorting and excluding observations with the codes “Refused to Answer” or “Don’t Know” from the eight FIES questions, and determining the value of the response variable (Y) based on the presence or absence of a “Yes” answer to these questions. Certain numeric variables (X3, X4, X18) are converted to numeric types so the algorithm can process them. The response variable (Y) is converted to a binary factor with the labels "0 = Food Security" and "1 = Food Insecurity ". Categorical variables are converted to factors with numeric labels. For example, the head of household's education level (X1) is coded from 1 = No schooling to 8 = Master's/Doctorate. Binary variables such as asset ownership or housing conditions (X6–X14, X21) are coded as 0 = No and 1 = Yes. Variables with more categories (e.g., roof type, flooring, water source, fuel) are assigned numerical labels according to the number of categories.

Data will be explored by describing samples and presenting the prevalence of food insecurity. Randomly partition the data into two parts: 80% for training data and 20% for test data. Perform resampling on minority classes in the training data using the RUS undersampling technique and oversampling techniques ROS and SMOTE NC. Performing feature selection using the RFE method on the balanced training data. This method selects the most relevant features for food insecurity classification by gradually eliminating less informative features. Selection is performed using 5-fold cross-validation, with the optimal number of features selected based on the highest classification performance. Performing model optimization using Bayesian methods to obtain the best combination of hyperparameters. Bayesian optimization uses 20 initial seeds to create a replacement model, which is then optimized over 10 iterations. This process is performed using stratified 5-fold cross-validation. One fold is used as test data, and the rest for training [21]. The hyperparameters that were adjusted are listed in Table 1.

Table 1. Random forest and LightGBM model hyperparameter domains

Algorithm	Hyperparameter	Domain	Default
Random Forest	*Trees*	100 – 300	500
	Maxnodes	20 – 60	Null
	Mtry	3 – 15	$\sqrt{p}$
	nodesize	1 – 20	1
LighGBM	*max_depth*	100 – 500	500
	learning_rate	8 – 15	6
	num_leaves	20 – 100	31
	n_estimator	100 – 500	100
	max_bin	5 – 50	255

Comparing the metric averages of the Bayesian optimization results, the best model is selected from the models with the highest G-Mean average values. Perform Random Forest and LightGBM classification modeling on the training data based on the best hyperparameter combination and evaluate the test data 50 times. The classification model was evaluated using a confusion matrix, which compares actual labels with predicted labels, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) metric. Model performance was evaluated through Accuracy, Sensitivity, Specificity, and G-Mean. The result of accuracy, sensitivity, specificity, and G-Mean results in a boxplot. Next, calculate the SHAP feature importance (SHAP FI) value as the importance score of variables in classification modelling. The variable with the highest SHAP score will be the most important.

Data Analysis Procedure

The research was conducted through three main stages: data preprocessing, modelling, and model evaluation. The preprocessing stage included data preparation and exploration, followed by data splitting into training and testing sets. In the modelling stage, various resampling methods and machine learning algorithms were applied along with feature selection and hyperparameter tuning. Finally, the model was evaluated using performance metrics and feature importance analysis. The complete workflow is illustrated in Figure 1.

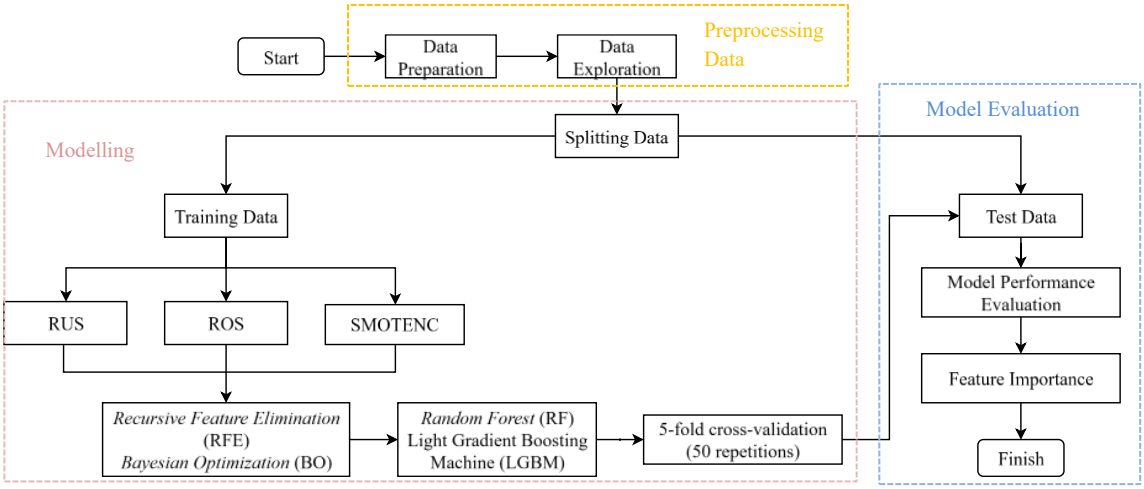


Figure 1. Systematic workflow of the data analysis process

This study uses the empirical data of 14,676 households from the March 2023 National Socioeconomic Survey (SUSENAS) in Papua Province. During the data exploration stage, 384 households were excluded because they had at least one "Don't know" or "No answer" response to the eight FIES questions. The total number of households covered by the study was 14,292. Based on household responses to the eight FIES questions, which form the basis for determining the response variable class Y (household vulnerability status), there were 2,557 households with food insecurity status and 11,375 with non-food insecurity status, as shown in the Figure 2. The explanatory variables used refer to previous research conducted by Dharmawan [22]. A detailed description of the variables employed in this study is presented in Table 2.

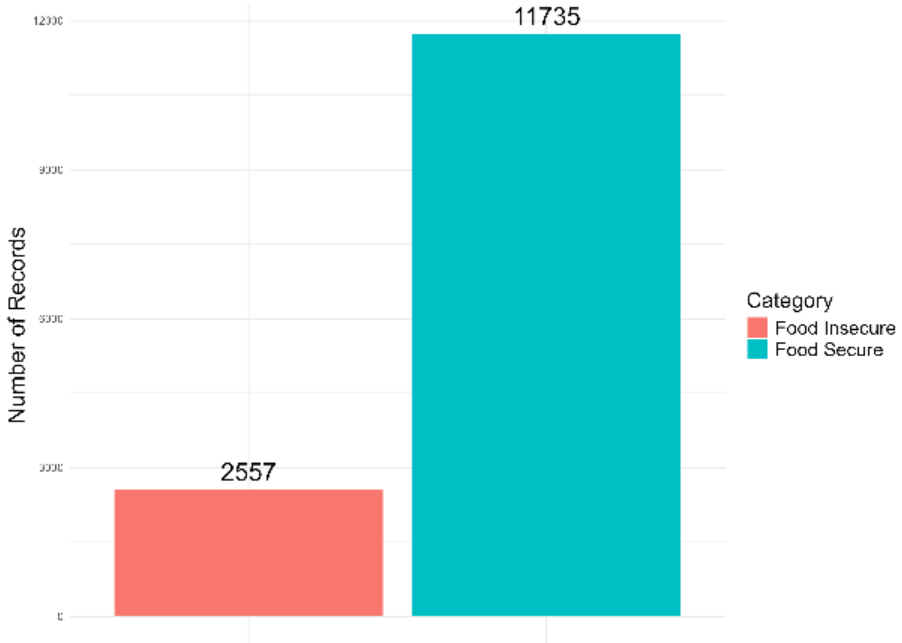


Figure 2. Proportion of households based on food insecurity status

Table 2. Data description

Variable	Description	Data Type
Household food insecurity status (Y)	1= Food Insecurity 0= No Food Insecurity	Nominal
Household Head Education (X1)	1 = No Education; 2=Elementary School; 3= Junior High School; 4= Senior High School; 5=Diploma 1-Diploma 3 6=Diploma 4/Bachelor's Degree; 7=Professional Degree; 8=Master's Degree/Doctorate Degree	Nominal
Vulnerable Households (X2)	1= vulnerable; 0= resistant	Nominal
Number of illiterate household members (X3)	Number of household members who cannot read and write	Continuous
Number of household members who save money (X4)	Number of household members who save part of their income	Continuous
Sources of Income for Transfer Recipients (X5)	1= Household members; 2=Remittances; 3=Investments; 4=Retirees	Nominal
Asset ownership (X6)	1= Yes, 0 = No	Nominal
Internet usage (X7)	1= Yes, 0 = No	Nominal
Sick but Not Hospitalized (X8)	1= Yes, 0 = No	Nominal
Non-cash Food Assistance (BPNT) Resipient (9)	1= Yes, 0 = No	Nominal
Keikutsertaan PKH (X10)	1= Yes, 0 = No	Nominal
Keikutsertaan KKS (X11)	1= Yes, 0 = No	Nominal
Recipients of Assistance from Local Governments (X12)	1= Yes, 0 = No	Nominal
Ownership of BPJS/PBI (X13)	1= Yes, 0 = No	Nominal
Participation in the Indonesia Pintar Program PIP/KIP (X14)	1= Yes, 0 = No	Nominal
Roof Types (X15)	1=Concrete; 2=Tiles; 3=Zinc; 4=Asbestos; 5=Bamboo; 6=Wood; 7= Straw/palm fiber/leaves/rattan; 8= Others	Nominal
Wall Types (X16)	1=Wall; 2=Bamboo; 3=Wood/board; 4=Bamboo mesh; 5=Wooden pole; 6=Bamboo; 7=Others	Nominal
Floor Types (X17)	1= Marble/granite; 2= Ceramic; 3= Parquet/vinyl; 4= Tiles/terrazzo; 5= Wood/boards; 6= Cement; 7= Bamboo; 8= Soil; 9= Others	Nominal
Floor areas (X18)	Total floor area of the household's dwelling (in square meters)	Continuous
Light Source (X19)	1= PLN electricity with meter; 2= PLN electricity without meter; 3= Non-PLN electricity; 4= Not electricity	Nominal
Cooking Fuel (X20)	1= Electricity ; 2= LPG 5.5 kg; 3= LPG 12 kg ; 4= LPG 3 kg; 5= Biogas; 6= Kerosene; 7= Briquettes; 8= Charcoal; 9= Firewood; 10= Does not cook at home; 11= others	Nominal
Adequate Sanitation (X21)	1= Yes, 0 = No	Nominal
Drinking Water Source (X22)	1= Branded bottled water; 2= Refillable water; 3= Leding; 4= Drilled well; 5= Protected well; 6= Unprotected well; 7= Protected spring; 8= Unprotected spring; 9= Surface water;10= Rainwater; 11= Others	Nominal

Imbalanced data

Data imbalance is one of the factors that can reduce the performance of machine learning algorithms [23].This situation arises when the distribution of observations between classes is very uneven. The majority class refers to the class with the largest number of observations, while the class with the smallest number of observations is referred to as the minority class [24].Generally, the level of imbalance in data is divided into three categories: Mild, which has a class proportion of 20%–40% of the dataset; Moderate, which has a class proportion of 1%–20% of the dataset; and Extreme, which has a class proportion of <1% of the dataset [12].

Random undersampling (RUS)

Random under-sampling (RUS) is a technique that reduces the number of samples in the majority class by randomly removing instances until a balance with the minority class is achieved. While this approach may alter the data distribution and reduce the representativeness of the majority class—potentially increasing the risk of misclassification RUS has demonstrated competitive performance compared to other undersampling and data cleaning techniques [25]. An illustration of the RUS mechanism is presented in the following Figure 3 [2].

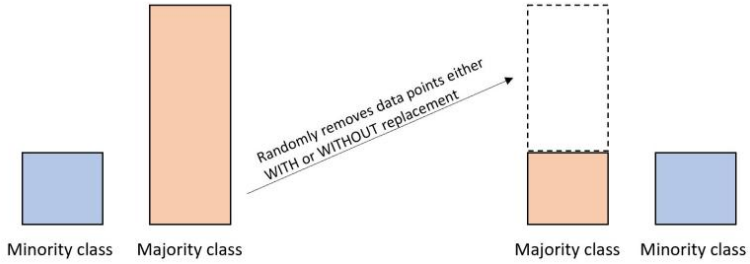


Figure 3. Illustration of how RUS works

Random oversampling (ROS)

Random over-sampling (ROS) works by randomly replicating instances of the minority class with replacement until class distribution becomes balanced. This means that minority class samples (e.g., students who dropped out) are duplicated at random until they equal the number of majority class samples (e.g., students who remained enrolled) [26]. ROS serves as a foundational technique for more advanced oversampling methods, such as SMOTE. By increasing the number of minority class samples, ROS helps mitigate class imbalance and enhances the predictive accuracy of classification models. An illustration of the ROS mechanism is presented in the following Figure 4 [2]:

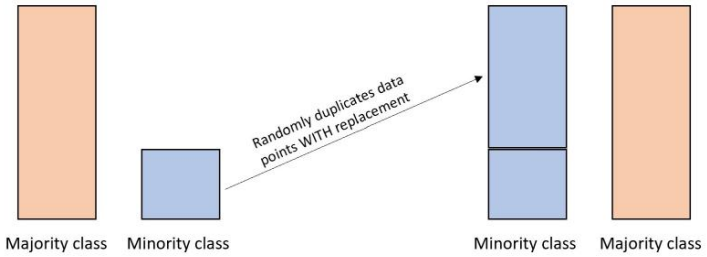


Figure 4. Illustration of how ROS works

SMOTE-NC (synthetic minority oversampling technique-nominal continuous)

SMOTE-NC is an extension of SMOTE for handling imbalanced data in numerical and categorical variables combinations. This method improves model performance through a trade-off between precision and sensitivity, enabling it to detect more actual minority classes [3]. To handle datasets containing a combination of nominal and continuous features, Chawla proposed a method called Synthetic Minority Oversampling Technique for Nominal and Continuous features (SMOTE-NC). Figure 5 demonstrates how SMOTENC work. An explanation of the SMOTE-NC algorithm is presented below [27]:

1. Median Calculation: This involves computing the median of the standard deviations across all consecutive features within the minority class. When nominal attributes differ between a sample and its nearest neighbors, this median value is incorporated into the Euclidean distance computation. It serves as a penalty factor, quantifying nominal feature discrepancies in terms comparable to the typical variation observed in continuous features.
2. Nearest Neighbor Calculation: The Euclidean distance is computed within the continuous feature space to identify the nearest neighbor of a given minority class instance. The previously determined median of the standard deviations is added to the distance calculation for each nominal feature that differs between the instance and its potential neighbor. This adjustment ensures that nominal differences are appropriately reflected in the overall distance measure.

Generating the synthetic instance: The continuous attributes of the newly created synthetic minority sample are generated using the same SMOTE technique described previously. For nominal features, the value assigned is the one that appears most frequently among the k nearest neighbors.

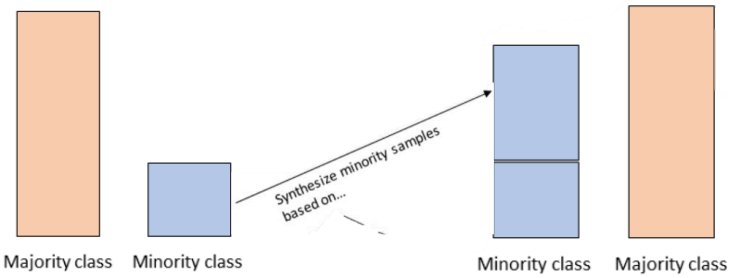


Figure 5. Illustration of how SMOTENC works

Random forest is a classification model algorithm developed from Regression Tree (CART) and the Classification method. Random forest consists of a combination of independent classification trees. Random Forest is formed from many trees that use random data samples. Before the trees are formed, a random feature selection stage is carried out. The basic concept of random forest is to apply the bootstrap aggregating (bagging) method [4]. The randomization process in the random forest algorithm is not only applied to data sampling, but also to predictor variable selection. This results in a set of decision trees that vary in structure and size, and have low correlation between trees, thereby reducing the overall prediction error rate [4]. An illustration of the Random Forest algorithm is presented in the following Figure 6.

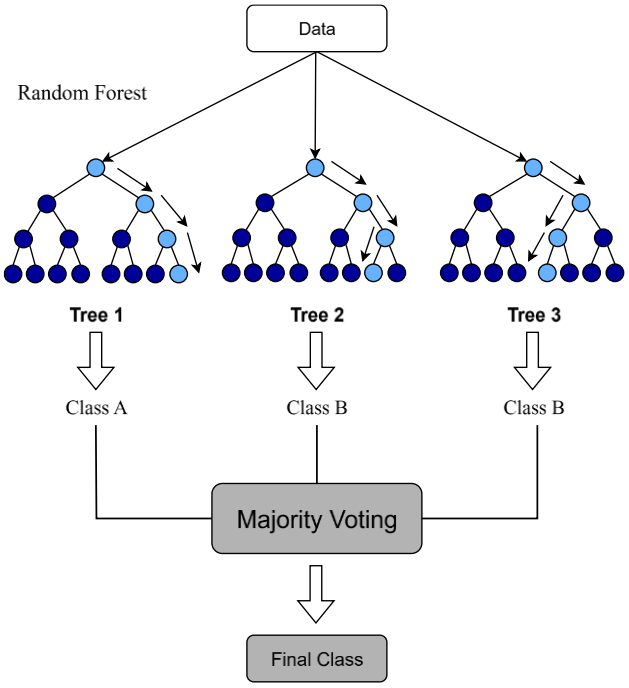


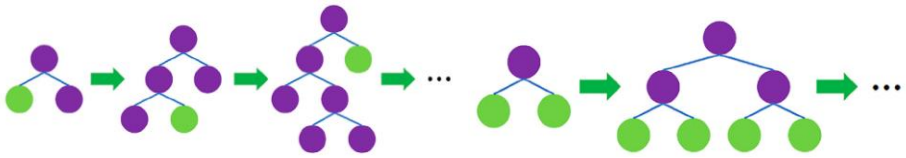
Figure 6. Illustration of Random Forest Algorithm

1. Bootstrap stage. Bootstrap is sampling accompanied by replacement. At this stage, take n random samples from the training data.
2. Random feature subset stage. Build a tree based on the previous bootstrap data. At each split, randomly select m explanatory variables where $m < p$. Next, perform the best split.
3. Repeat steps 1 to 2 k times to obtain k trees.
4. Perform aggregation of the prediction results of the k trees and use majority vote (taking the most votes) to determine the prediction result.

Light gradient boosting machine (LGBM)

Light Gradient Boosting Machine (LightGBM) merupakan algoritma yang dikembangkan oleh Microsoft Research Asia dengan memanfaatkan kerangka kerja Gradient Boosting Decision Tree (GBDT) [5]. This algorithm is designed to enhance computational efficiency, enabling faster and more effective predictions when working with large-scale datasets [27]. LightGBM has several advantages over other GBDT methods,

such as faster training, higher efficiency, lower memory usage, and good accuracy. LightGBM can also handle large-scale datasets with support for parallel learning and GPUs [27]. As a fast and distributed Gradient Boosting framework, this algorithm can be used for various machine learning tasks, such as classification [28]. It is assumed that the raw dataset consists of examples $N = \{1,2, \dots, n\}$, and a LightGBM model comprising $T = \{1,2, \dots t\}$ trees is generated. After t iterations, the final prediction is obtained by combining the results from the $((1 - t)\ th$ dan tth iterations. In XGBoost, tree growth follows a level-wise pattern, splitting all nodes at the same level even if the splitting gain value is small. This often wastes computational resources. In contrast, LightGBM uses a leaf-wise pattern, selecting the node with the largest gain to split. This strategy makes the model more controlled, prevents overfitting through tree depth limits, and is more efficient. As a result, LightGBM can achieve higher accuracy than other boosting algorithms. The iterative process is illustrated in the following Figure 7 as follows [28].



Leaf-wise tree growth Level-wise tree growth

Figure 7. Leaf-wise vs. level-wise tree growth

Recursive feature elimination (RFE)

In this study, a simulation is carried out to generate a data series. Recursive Feature Elimination (RFE) is a widely used technique that enhances data quality by selecting the most relevant features. It provides an effective approach to identify key variables before incorporating them into a machine learning model. RFE was introduced by Guyon [10] and initially applied to cancer classification using Support Vector Machines (SVM). The method begins by utilizing all features to construct the SVM model, then ranks each feature based on its contribution to the model. Irrelevant features with minimal impact are subsequently eliminated from the feature set.

The Recursive Feature Elimination (RFE) feature selection method aims to select the best combination of attributes while performing elimination based on the ranking of feature importance generated by the model recursively up to the desired number of attributes. According to [10], the results of feature selection using Recursive Feature Elimination (RFE) can speed up the training process and improve classification capabilities.

Bayesian optimization

Bayesian optimization is an iterative method aimed at finding the optimal hyperparameters for machine learning models [29]. It is a surrogate model-based optimization strategy that focuses on optimizing functions that are costly to evaluate. This approach selects the next set of hyperparameters by leveraging the outcomes of previous evaluations [30], enabling it to efficiently approach the optimal configuration within a limited number of iterations. The key procedures involved in Bayesian optimization are outlined as follows [31]:

1. Choosing n initial points (observed points) to assess the previously unrecognized objective function (true function).
2. Employing the initial data to train the surrogate model, namely the probabilistic model (GP mean), will progressively converge to the true objective function.
3. The acquisition function determines the subsequent point for evaluation. The probability of improvement (PI), expected improvement (EI), and GP upper confidence bound (GP-UCB) are three acquisition functions that are often utilized. The PI measures the probability that the objective function value at the subsequent evaluation point will surpass the best value previously attained. EI determines the anticipated degree of improvement that will be achieved at a point close to the ideal value. GP-UCB uses the Gaussian process prediction value to determine the following site to investigate.
4. Add the new evaluation point to the surrogate model.
5. Until the maximum number of iterations is achieved, keep choosing the next point, assessing the objective function, and updating the surrogate model.

Shapley additive explanations (SHAP)

Shapley Additive Explanations (SHAP) is a technique used to explain individual predictions based on calculations derived from the Shapley value concept in game theory [12].The purpose of SHAP is to calculate the contribution of each variable so that it can explain each individual's prediction. SHAP can explain global and local predictions by calculating Shapley values developed by Lloyd Shapley (1953), a method in games to determine rewards for players proportionally and reasonably by considering the players' contributions to the collected value. Players collaborate in a coalition and receive certain benefits from the collaboration. The average marginal influence of all possible combination of variable values. Players are analogized as predictor variables, and rewards are estimates. Shapley values are derived from this game concept. To calculate Shapley values, all possible combinations of players or predictor variables must be evaluated with and without predictor variable j. The Shapley value for player j is obtained using equation (1).

$$\phi_j = \sum_{s \subseteq M(j)} \frac{|S|!(M-(|S|-1)!}{M!} \left(v(S \cup \{j\}) - v(S)\right) \quad (1)$$

Where,

ϕ_j : Shapley value, representing the contribution of the j feature,
S : A coalition is a subset of features that omits feature j,
M : Total number of feature,
$v(S \cup \{j\})$: The prediction obtained from the coalitionSwith the inclusion of feature j,
$v(S)$: The prediction obtained from the coalition Swithout feature j

A high positive ϕ_{ij} indicates a synergistic interaction, while a negative value suggests redundancy. The full matrix of SHAP interaction values across all feature pairs forms a symmetric $M \times M$ matrix, offering a comprehensive view of inter-feature dependencies within the model.

Confusion matrix

A confusion matrix is an easy and effective technique for measuring the performance of a classification system. The primary purpose of a confusion matrix is to assess the performance or accuracy of a classification system in classifying test data. Model performance was evaluated through Accuracy, Sensitivity, Specificity, and G-Mean. Accuracy does not serve as a reliable metric in cases of imbalanced data, as it merely reflects the overall classification performance and can provide a distorted view. In contrast, sensitivity and specificity are crucial for evaluating how well a model predicts positive and negative classes, respectively. These measures offer a more comprehensive assessment of a classifier's effectiveness across both the majority and minority classes [32]. The G-mean is a useful metric for handling imbalanced datasets, where one class contains significantly fewer samples than the other. The formulas for calculating these metrics are presented in equations (2) to (5). Formulas 2 to 4 based on the information in Table 3.

Table 3. Confusion matrix for binary classification

Actual	Predicted	
	Positive (1)	Negative (0)
Positive (1)	TP (True Positive)	FN (False Negative)
Negative (0)	FP (False Positive)	TN (True Negative)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{FN+TP} \quad (3)$$

$$\text{Spesificity} = \frac{TN}{TN+FP} \quad (4)$$

$$\text{G-Mean} = \sqrt{Sensitivity \times Spesificity} \quad (5)$$

RESULTS AND DISCUSSIONS

The 2022 National Socio-Economic Survey (Susenas) was conducted across various provinces in Indonesia. For this study, data from Papua Province was used, consisting of 14,292 observations, with households as the unit of analysis. However, the exploratory data analysis did not include all households, as some had at least one response of "don't know" or "refused to answer" to the eight FIES (Food Insecurity Experience Scale) questions. The household responses to the FIES questions were used to determine the category of the response variable Y (household food insecurity status).

Figure 8 shows the proportion of food-security and food-insecurity households across two provinces, highlighting a disparity in Papua Province. In Papua, 82.1% of households were classified as food-security, while 17.9% were considered food-insecurity. The significantly different class composition of the response variable Y reflects a typical case of imbalanced data. The dataset used in this study consists of 22 predictor variables and one categorical response variable.

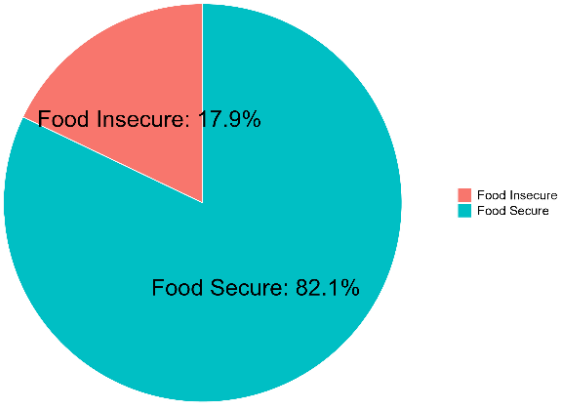


Figure 8. Food insecurity in Papua Province

Model development

The classification model in this study uses two types of machine learning algorithms with different characteristics, namely Random Forest (RF) and Light Gradient Boosting (LGBM). The modeling of these two classification models will be combined with three treatments: data imbalance handling using three techniques, feature reduction using Recursive Feature Elimination (RFE), and hyperparameter tuning using Bayesian Optimization (BO), as well as a combination of RFE and BO. Eighteen experimental scenarios were generated by combining two ensemble models (Random Forest and LightGBM), three imbalance handling techniques (RUS, ROS, SMOTENC), and three treatments (RFE, BO, RFE+BO). For Random Forest, this resulted in nine scenarios (e.g., RF+RUS+RFE, RF+ROS+BO, RF+SMOTENC+RFE+BO), and for LightGBM, another nine scenarios were formed (e.g., LGBM+RUS+RFE, LGBM+ROS+BO, LGBM+SMOTENC+RFE+BO). Each scenario was evaluated across 50 repetitions of stratified random sampling, and performance was compared using Accuracy, Sensitivity, Specificity, and G-Mean. The combination of machine learning algorithms and the three treatments will result in eighteen classification models, as shown in Table 4.

Table 4. Model combination

No	Model	Unbalanced Data Handling	Treatment	Combination
1	Random Forest (RF)	RUS	RFE	RF+RUS+RFE
2		ROS	RFE	RF+ROS+RFE
3		SMOTENC	RFE	RF+SMOTENC+RFE
4		RUS	BO	RF+RUS+BO
5		ROS	BO	RF+ROS+BO
6		SMOTENC	BO	RF+SMOTENC+BO
7		RUS	RFE+BO	RF+RUS+RFE+BO
8		ROS	RFE+BO	RF+ROS+RFE+BO
9		SMOTENC	RFE+BO	RF+SMOTENC+RFE+BO
10	LightGBM (LGBM)	RUS	RFE	LGBM+RUS+RFE
11		ROS	RFE	LGBM+ROS+RFE
12		SMOTENC	RFE	LGBM+SMOTENC+RFE
13		RUS	BO	LGBM+RUS+BO
14		ROS	BO	LGBM+ROS+BO
15		SMOTENC	BO	LGBM+SMOTENC+BO
16		RUS	RFE+BO	LGBM+RUS+RFE+BO
17		ROS	RFE+BO	LGBM+ROS+RFE+BO
18		SMOTENC	RFE+BO	LGBM+SMOTENC+RFE+BO

Model evaluation

This study produced models combined with feature selection using RFE, Bayesian Optimization, and imbalanced data handling techniques. Model evaluation used a confusion matrix, specifically accuracy, sensitivity, specificity, and G-Mean metrics, on both training and testing datasets to assess performance. The results were then visualized and grouped based on the models and treatments applied.

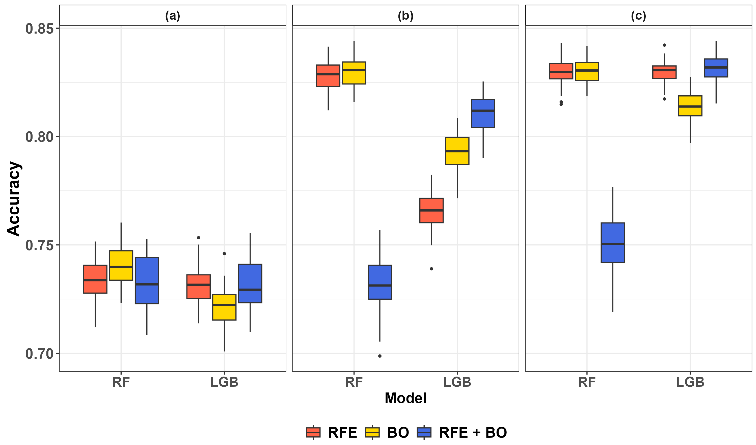


Figure 9. Accuracy performance of 50 repetitions for (a) RUS, (b) ROS, and (c) SMOTENC

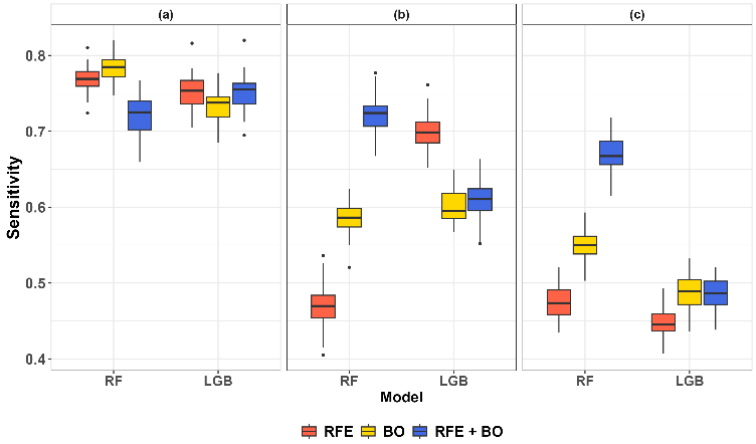


Figure 10. Sensitivity performance of 50 repetitions for (a) RUS, (b) ROS, and (c) SMOTENC

The accuracy performance evaluation in Figure 9 showed that under the RUS (a) balancing method, accuracy tended to be low and unstable, especially for the Random Forest (RF) model. The Bayesian Optimization (BO) technique produced more stable results than RFE alone, but the combination of RFE and BO often led to decreased performance in RF. Conversely, when using ROS (b) and SMOTENC (c), model performance improved significantly, particularly for the LightGBM (LGB) model. The combination of RFE + BO applied to LGB with SMOTENC yielded the highest and most stable accuracy, demonstrating strong synergy between feature selection, tuning, and balancing methods. Overall, LGB outperformed RF, showing greater stability toward tuning and feature selection, and achieving optimal performance when combined with SMOTENC and RFE + BO. It appears that the boosting mechanism in LGB is more adaptive to unbalanced data, while bagging-based RF tends to be biased towards the majority class. The Bayesian Optimization (BO) technique has also proven to be more effective than RFE, because parameter tuning allows the model to adapt to the data, whereas RFE sometimes removes important features.

The sensitivity results in Figure 10 show that of the three combinations, the highest sensitivity is found in the model combined with RUS, while ROS and SMOTENC significantly reduce sensitivity, especially in LGB. The RFE + BO combination in RF yields better results than other methods for all unbalanced data treatments. Additionally, RUS produces the most stable sensitivity, particularly in the RFE + BO combination with RF and LGB. The results show that sensitivity decreases slightly in ROS and SMOTENC,

but RFE + BO combined with RF remains consistent in accurately predicting food insecurity. Overall, using RUS as a balancing method and RF as a model algorithm combined with the RFE + BO combination is the most effective approach to improving sensitivity in Papua Province. In this metric, RF actually outperforms LGBM, especially with the combination of RFE+BO and RUS. RUS makes the data distribution more balanced, allowing RF to focus more on recognizing minority classes. Conversely, LGBM tends to experience a decrease in sensitivity on ROS and SMOTENC due to the risk of overfitting and synthetic patterns that are not fully representative.

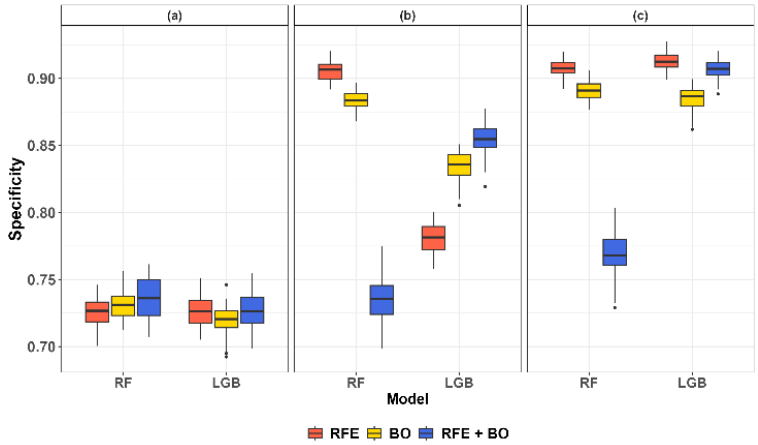


Figure 11. Specificity performance of 50 repetitions for (a) RUS, (b) ROS, and (c) SMOTENC

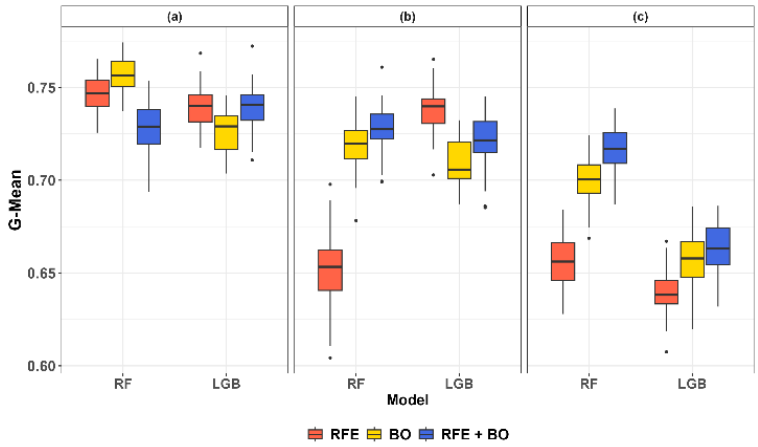


Figure 12. G-Mean performance of 50 repetitions for (a) RUS, (b) ROS, and (c) SMOTENC

Figure 11 shows that specificity is superior, especially in combination with RFE and BO, both RF and LGB. The combination model of LGB with RFE+BO and SMOTENC handling appears to be superior to other combination models. The classification method applied with RUS imbalanced data handling performs poorly predicting food insecurity. The combination of ROS and SMOTENC achieves high specificity performance when applied with the combination of RFE+BO and the LGBM model. LGBM outperforms RF because it is more consistent in recognizing the majority class. Although all three balancing methods improve sensitivity, LGBM can maintain high specificity, especially with SMOTENC. RF often loses stability; when its sensitivity increases with RUS, its specificity decreases.

Based on the G-Mean evaluation results in Figure 12, the Random Forest (RF) model consistently shows better performance than LightGBM (LGB) in maintaining a balance between sensitivity and specificity when combined with the RUS undersampling method. The RUS data balancing technique produces the highest and most stable G-Mean values in treatments with RFE, with BO, and in the RFE+BO combination. Meanwhile, using ROS provides fairly competitive performance, but tends to be slightly below RUS. On the other hand, the SMOTENC method generally produces lower G-Mean values, indicating an imbalance between sensitivity and specificity. The model combines RFE feature selection and hyperparameter tuning

with the RF classification model, and all data handling methods appear to be the most consistent and stable for improving G-Mean performance. It can therefore be concluded that the RFE+BO combination model can produce better performance when using Random Forest and LightGBM classification models with three techniques for handling imbalanced data. Meanwhile, the model using only RFE feature selection achieves good G-Mean performance when combined with the Random Forest classification model and RUS data handling. The model with BO hyperparameter tuning performs better when combined with the Random Forest classification model and RUS and ROS handling. As a balance metric, G-Mean confirms the superiority of LGBM. The combination of RFE+BO with SMOTENC produces the highest G-Mean, because sensitivity increases without sacrificing specificity. RFE or BO does not significantly help RF, so G-Mean tends to be lower and less stable. Thus, LGBM outperforms RF in maintaining performance balance between classes.

Based on the results of the three-way ANOVA analysis, it was found that the model factor, unbalanced data handling, and feature selection or hyperparameter optimization treatments each had a significant effect on the G-Mean value ($p < 0.001$). In addition, all interactions between factors, both two-way and three-way, also showed interactions. This indicates that the influence of a factor cannot be viewed in isolation, as it depends on its combination with other factors. Thus, model performance is not only determined by the type of model used but is also greatly influenced by the data balancing method and feature processing applied.

Performance evaluation of the classification model based on the G-Mean values shown in Table 5 for the top 10 results based on G-Mean indicates that the RF classification model combined with Bayesian optimization and RUS handling outperforms other methods in terms of average performance across the three treatments. The best G-Mean value obtained was 0.756, with a range of values from 0.737 to 0.774. SMOTENC outperformed on data with both numerical and categorical features.

Table 5. Best Method Combinations based on G-Mean

No.	Model	Imbalance	Treatment	Min	Max	Mean	SD
1	RF	RUS	BO	0.737	0.774	0.756	0.009
2	RF	RUS	RFE	0.726	0.752	0.747	0.009
3	LGB	RUS	RFE	0.717	0.753	0.740	0.010
4	LGB	ROS	RFE	0.703	0.782	0.740	0.011
5	LGB	RUS	RFE+BO	0.711	0.755	0.735	0.012
6	LGB	RUS	BO	0.704	0.746	0.729	0.012
7	RF	RUS	RFE+BO	0.694	0.753	0.729	0.014
8	RF	ROS	RFE+BO	0.699	0.761	0.728	0.012
9	RF	ROS	BO	0.678	0.745	0.720	0.013
10	RF	SMOTENC	RFE+BO	0.687	0.739	0.717	0.011

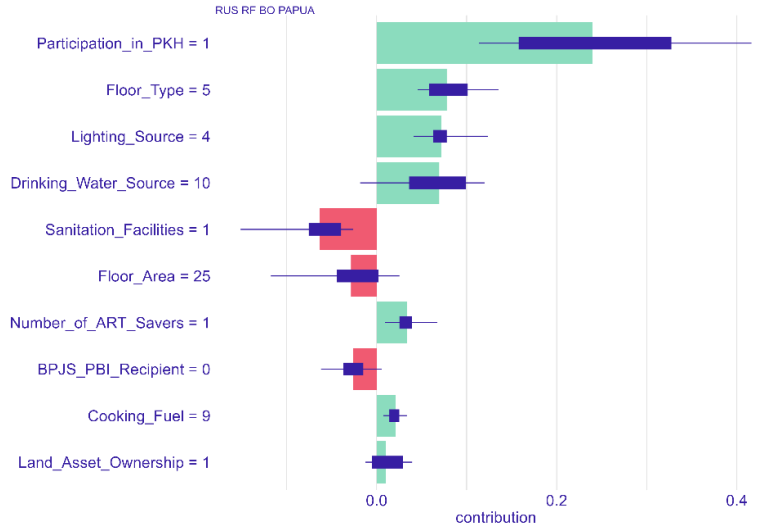


Figure 13. SHAP Plot: most significant contributions in the model

Figure 13 shows that recipients of PKH social assistance, wooden floors, non-electric lighting, and rainwater as a source of drinking water are the four main predictors that increase the probability of a

household being classified as food insecurity. This is indicated by the SHAP (Shapley values) on the positive side (green to the right), which contributes to the predicted increase in food insecurity risk. Conversely, predictors such as households with adequate sanitation, a minimum house size of 25 square meters, and not being a BPJS PBI recipient have a negative contribution, meaning that these three predictors reduce the probability of a household being classified as food insecurity. The implications of the SHAP results show that social assistance programs such as PKH are on target, namely, vulnerable households. However, their effectiveness needs to be strengthened to reduce food insecurity tangibly. In addition, household condition indicators such as floor type, lighting source, and access to clean water emphasize the importance of basic infrastructure development interventions. Adequate sanitation, house size, and BPJS PBI membership also show that housing quality and social health protection play an important role in reducing the risk of food insecurity. Thus, policies to alleviate food insecurity in Papua should focus on food assistance and improving basic infrastructure and integration with social protection programs.

Similar research [33] shows that combining LightGBM with Bayesian Optimization and feature selection can significantly improve classification performance. Unlike those studies, which only highlight the advantages of LGBM, the results of this study show that Random Forest with the undersampling (RUS) method remains more stable in conditions of extreme imbalance, while LGBM achieves the best performance when combined with RFE+BO and SMOTENC. This difference confirms that the effectiveness of a method is greatly influenced by the characteristics of the data and the class balancing technique used.

CONCLUSION

This study shows that the performance of classification models is greatly influenced by the combination of algorithms, data balancing methods, and feature processing techniques. LightGBM excels in accuracy and stability, especially with SMOTENC and the RFE + BO combination, while Random Forest is better at maintaining sensitivity-specificity balance (G-Mean), especially with RUS. ANOVA results confirm that the model factor, data balancing method, and treatment (RFE, BO, RFE+BO) along with their interactions significantly influence model performance. Based on G-Mean, the best combination is RF + BO + RUS.

This study integrates a model interpretability approach using SHAP (SHapley Additive exPlanations) to identify the main factors of household food insecurity in Papua. The results of the analysis show that receiving PKH social assistance, having wooden floors, non-electric lighting, and using rainwater as a source of drinking water increase the likelihood of a household being classified as food insecurity, while proper sanitation, a house area of at least 25 m², and not being a BPJS PBI recipient reduce the risk. These findings confirm that food insecurity is influenced by a combination of physical household conditions, social aspects, and access to basic services. Therefore, strategies to alleviate food insecurity need to be comprehensive, involving the strengthening of social assistance, the development of basic infrastructure, and integration with social protection.

This study proves that the combination of Recursive Feature Elimination (RFE) and Bayesian Optimization (BO) in LightGBM and Random Forest models can improve the performance of classifying imbalanced data in the context of food insecurity. This methodological approach is still rarely applied in the literature, making it a new finding that reinforces the simultaneous use of feature selection and hyperparameter optimization techniques. Applying the combination of RFE and BO in ensemble models (LightGBM and Random Forest) confirms that integrating these two techniques can produce a more robust model than either technique alone, which has not been widely explored in related studies. In addition, this study can assist in formulating food intervention policies in areas with high levels of vulnerability.

This study has several limitations that need to be considered. The data used only covers one province, so the results may not represent the national conditions. The limited number of Bayesian Optimization iterations may not have found the most optimal hyperparameter configuration. In addition, the study only applied three data balancing techniques (RUS, ROS, and SMOTENC) and compared two main algorithms, namely Random Forest and LightGBM. Given these limitations, future research should use a more diverse data set, including different levels of imbalance, so that model performance can be tested more comprehensively. Deep learning approaches and GAN-based resampling techniques can also be explored to improve model quality.

REFERENCES

[1] T. Hasanin and T. M. Khoshgoftaar, "The effects of random undersampling with simulated class imbalance for big data," *Proc. - 2018 IEEE 19th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2018*, pp. 70–79, 2018, doi: 10.1109/IRI.2018.00018.

[2] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[3] D. T. Utari, "Integration of Svm and Smote-Nc for Classification of Heart Failure Patients," *Barekeng*, vol. 17, no. 4, pp. 2263–2272, 2023, doi: 10.30598/barekengvol17iss4pp2263-2272.

[4] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[5] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, no. December, pp. 3147–3155, 2017, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[6] Y. M. Indah, R. Aristawidya, and A. Fitrianto, "Comparison of Random Forest , XGBoost and LightGBM Methods on the Human Development Index Classification Comparison of Random Forest , XGBoost and LightGBM Methods on the Human Development Index Classification," 2025.

[7] O. Okun, "Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations," *United States Amerika (USA)IGI Glob. snippet*, 2011.

[8] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowl. Inf. Syst.*, vol. 66, no. 3, pp. 1575–1637, Mar. 2024, doi: 10.1007/s10115-023-02010-5.

[9] Y. Li, J. Bian, and R. Song, "Video-based deception detection using wrapper-based feature selection," in *2024 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Jun. 2024, pp. 1–5. doi: 10.1109/CIVEMSA58715.2024.10586610.

[10] I. Guyon, J. Weston, and S. Barnhill, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, pp. 389– 422, 2002.

[11] M. Artur, "Review the performance of the Bernoulli Naïve Bayes Classifier in Intrusion Detection Systems using Recursive Feature Elimination with Cross-validated selection of the best number of features," *Procedia Comput. Sci.*, vol. 190, no. 2019, pp. 564–570, 2021, doi: 10.1016/j.procs.2021.06.066.

[12] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," *Comput. Biol. Med.*, vol. 149, no. September, p. 106043, 2022, doi: 10.1016/j.compbiomed.2022.106043.

[13] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *J. Electron. Sci. Technol.*, vol. 17, no. 1, pp. 26–40, 2019, doi: 10.11989/JEST.1674-862X.80904120.

[14] K. Yang, L. Liu, and Y. Wen, "The impact of Bayesian optimization on feature selection," *Sci. Rep.*, vol. 14, no. 1, pp. 1–11, 2024, doi: 10.1038/s41598-024-54515-w.

[15] R. Y. Krisnabayu, M. K. Dr. Drs. Achmad Ridok, and S. T. . M. T. . P. D. Agung Setia Budi, "Prediksi Masa Studi Mahasiswa Menggunakan Random Forest Dengan Seleksi Fitur RF-RFE (Recursive Feature Elimination) dan Particle Swarm Optimization," Universitas Brawijaya, 2022.

[16] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. November, pp. 4766–4775, 2017, doi: 10.48550/arXiv.1705.07874.

[17] M. SUBIANTO, I. Y. ULYA, E. RAMADHANI, B. SARTONO, and A. F. HADI, "Application of SHAP on CatBoost classification for identification of variabels characterizing food insecurity occurrences in Aceh Province households," *J. Nat.*, vol. 23, no. 3, pp. 230–244, Oct. 2023, doi: 10.24815/jn.v23i3.33548.

[18] M. D. Smith, M. P. Rabbitt, and A. Coleman- Jensen, "Who are the World's Food Insecure? New Evidence from the Food and Agriculture Organization's Food Insecurity Experience Scale," *World Dev.*, vol. 93, no. January 2017, pp. 402–412, 2017, doi: 10.1016/j.worlddev.2017.01.006.

[19] C. Cafiero, S. Viviani, and M. Nord, "Food security measurement in a global context: The food insecurity experience scale," *Meas. J. Int. Meas. Confed.*, vol. 116, no. November, pp. 146–152, 2018, doi: 10.1016/j.measurement.2017.10.065.

[20] Badan Pangan Nasional, "Indeks Ketahanan Pangan," *Indeks Ketahanan Pangan*, vol. 58, no. 12, pp. 7250–7257, 2022.
[21] J. G. Moreno-Torres, J. A. Saez, and F. Herrera, "Study on the impact of partition-induced dataset shift on k-fold cross-validation," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 23, no. 8, pp. 1304–1312, 2012, doi: 10.1109/TNNLS.2012.2199516.
[22] H. Dharmawan, "Perbandingan Ukuran Kepentingan Peubah dari Berbagai Algoritme Pembelajaran Mesin untuk Kondisi Data Tidak Seimbang dengan Berbagai Jenis Perlakuan," 2023.
[23] M. Koziarski, "Potential Anchoring for imbalanced data classification," *Pattern Recognit.*, vol. 120, 2021, doi: 10.1016/j.patcog.2021.108114.
[24] M. Błaszczyk and J. Jedrzejowicz, "Framework for Imbalanced Data Classification. Di dalam: 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems," in *Framework for Imbalanced Data Classification. Di dalam: 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 2021, vol. Volume ke-, p. hlm 3477– 3486.
[25] S. Ahmed, A. Mahbub, F. Rayhan, R. Jani, S. Shatabda, and D. M. Farid, "Hybrid Methods for Class Imbalance Learning Employing Bagging with Sampling Techniques," *2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2017*, no. December, pp. 1–5, 2018, doi: 10.1109/CSITSS.2017.8447799.
[26] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 92–122, 2014, doi: 10.1007/s10618-012-0295-5.
[27] P. Jain, M. T. Islam, and A. S. Alshammari, "Comparative analysis of machine learning techniques for metamaterial absorber performance in terahertz applications," *Alexandria Eng. J.*, vol. 103, no. August 2023, pp. 51–59, 2024, doi: 10.1016/j.aej.2024.05.111.
[28] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, 2021, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8467876/
[29] E. C. Garrido-Merchán and D. Hernández-Lobato, "Dealing with categorical and integer-valued variables in Bayesian Optimization with Gaussian processes," *Neurocomputing*, vol. 380, pp. 20–35, 2020, doi: 10.1016/j.neucom.2019.11.004.
[30] M. Aghaabbasi, M. Ali, M. Jasinski, Z. Leonowicz, and T. Novak, "On Hyperparameter Optimization of Machine Learning Methods Using a Bayesian Optimization Algorithm to Predict Work Travel Mode Choice," *IEEE Access*, vol. 11, no. January, pp. 19762–19774, 2023, doi: 10.1109/ACCESS.2023.3247448.
[31] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," *Adv. Neural Inf. Process. Syst.*, vol. 4, pp. 2951–2959, 2012.
[32] V. P. K. Turlapati and M. R. Prusty, "Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19," *Intell. Med.*, vol. 3–4, no. July, p. 100023, 2020, doi: 10.1016/j.ibmed.2020.100023.
[33] J. Zhou, X. Tong, J. Zhou, R. Liu, and S. Bai, "Feature selection of power system data for frequency prediction with BO-LightGBM-Boruta," *Energy AI*, vol. 21, no. August, p. 100581, 2025, doi: 10.1016/j.egyai.2025.100581.