BAREKENG: Journal of Mathematics and Its Applications
June 2026 Volume 20 Issue 2 Page 1111–1124
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20no2pp1111-1124

MIXED-EFFECTS MODELS WITH GENERALIZED RANDOM FOREST: IMPROVED FOOD INSECURITY ANALYSIS

Herlin Fransiska[1*], Agus Mohamad Soleh[2], Khairil Anwar Notodiputro[3], Erfiani[4]

[1,2,3,4]*Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Meranti Wing 22 Level 4 Kampus IPB Darmaga, Bogor, Jawa Barat, 16680 Indonesia

*Corresponding author's e-mail: * **17.hfransiska@apps.ipb.ac.id***

Article Info

Article History:

Received: 1st May 2025
Revised: 23rd June 2025
Accepted: 17th August 2025
Available online: 18th January 2026

Keywords:

Food insecurity;
Generalized random forest;
Mixed-effects models;
Prediction.

ABSTRACT

Food insecurity is a complex issue that requires a deep understanding of its influencing factors. Accurate predictions are crucial for effective interventions. Machine learning is well-suited to the large and complex data available in the big data era. However, machine learning generally does not accommodate hierarchical or clustered data structures, making them challenging for machine learning modeling. One model that accommodates hierarchical data structures is the mixed-effects model. This study introduces a novel approach to predict food insecurity by integrating mixed-effects models and a generalized random forest. Mixed-effects models capture variations in hierarchical or clustered data, such as differences between regions, and the generalized random forest, as extended and developed from the traditional random forest, is integrated to model fixed effects and improve prediction accuracy. The empirical data used were the food insecurity data from 2021 in West Java, Indonesia. The results show that mixed-effects models with a generalized random forest significantly improve the prediction accuracy compared to other models. The average performance measure shows GMEGRF is a good model and has a balanced accuracy value of 0.6789709, which is the highest result compared to other methods. This methodological advancement offers a new robust model for understanding and potentially mitigating food insecurity, ultimately informing efforts towards SDG 2 (Zero Hunger).

How to cite this article:

H. Fransiska, A. M. Soleh, K. A. Notodiputro and Erfiani., "MIXED-EFFECTS MODELS WITH GENERALIZED RANDOM FOREST: IMPROVED FOOD INSECURITY ANALYSIS", *BAREKENG: J. Math. & App.*, vol. 20, no. 2, pp. 1111-1124, Jun, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

Food insecurity is a global issue owing to its potential to be a widespread problem affecting individuals across lifespans in terms of health and well-being [1]. This means that a lack of access to sufficient nutritious food can have serious consequences for people of all ages, from babies and children to adults and seniors. Addressing this problem requires a thorough understanding. Food insecurity frequently varies across geographical regions and socioeconomic groups. Statistics can play a role in addressing the problem of food insecurity by developing statistical models to predict future trends in food insecurity. Accurate prediction of food insecurity is essential for the design and implementation of effective interventions and policies aimed at mitigating its impact and achieving food security.

Big data related to food security, encompassing demographic, economic, environmental, and social indicators, make machine learning approaches particularly well-suited for analyzing and predicting this phenomenon. Several studies on machine learning models to predict food vulnerability include G. Nica-Avram *et al.* [2], A. H. Villacis *et al.* [3], C. Gao *el al.* [4], S. Gholami *et al.* [5] and J. J. L. Westerveld *et al.* [6]. Machine learning algorithms are a valuable tool that enhances model goodness of fit, uncovers meaningful and valid hidden patterns in data, detects nonlinear and non-additive effects, offers insights into data trends, methodology, and theory, and advances scientific research [7]. However, food insecurity data often exhibit hierarchical or clustered structures. For instance, data may be collected at the household level within villages, districts, or provinces. These structures introduce dependencies between observations within the same cluster, violating the independence assumption of many standard machine learning algorithms. Several studies on mixed models: A. Hajjem *et al.*[8], A. Hajjem *et al.* [9], Hajjem *et al.* [10], J. Hu and S. Szymczak [11], P. Krennmair and T. Schmid [12], M. Pellagatti *et al.* [13] , R. J. Sela and J. S. Simonoff [14] , J. L. Speiser *et al.* [15], L. Fontana *et al.* [16] , and D. Kusumaningrum *et al.* [17]. The hierarchical data approach, as highlighted by P. C. Chen *et al.*, offers a good performance evaluation method by integrating expert judgment and data-driven techniques [18]. This approach is highly valuable as it provides more accurate and relevant insights, which ultimately contribute to the improvement of food security modeling.

Machine learning models are generally good at finding patterns in complex datasets and using these patterns to make predictions. Random Forest is a popular and powerful machine-learning algorithm introduced by L. Breiman (2001) [19]. In machine learning, RF assumes that observations are obtained independently from a population. If data are hierarchical (nested, like students within classrooms within schools) or clustered in structure (grouped, like per capita income from various villages within a regency/district). It does not inherently understand these groupings and can treat them as independent data points, which can lead to biased inference owing to the underestimation of standard errors in linear models [20] and the identification of false subgroups and inaccurate variable selection [14], [21]. GRF, an extension of the traditional random forest, offers increased flexibility in modeling the relationships between features and the target variable [22]. The GRF algorithm employs forest-based local estimation and splitting to maximize heterogeneity for optimal split selection and utilizes a gradient tree algorithm to optimize an approximate criterion [22], [23], [24]. Several studies on GRF: E. Zhou and D. Lee 2024 [23] and H. Fransiska et al. [24]. The GRF model tends to be stable.

Mixed-effects models are statistical techniques specifically designed to handle hierarchical and clustered data. Generalized Mixed-Effects Random Forest (GMERF) is one such example. This statement suggests that GMERF combines the strengths of mixed-effects models and random forests. Although it offers comparable performance to mixed models using linear methods, GMERF exhibits greater robustness [13].

This study proposes a novel approach to enhance the analysis and prediction of food insecurity by integrating mixed-effects models with a generalized random forest (GRF) algorithm. This approach aims to capture both the hierarchical structure of the data through mixed-effects models and the complex, nonlinear relationships through the GRF. Specifically, the mixed-effects component accounts for variations across regions or other relevant clusters, whereas the GRF models the fixed effects and improves overall prediction accuracy.

2. RESEARCH METHODS

This section details the methods used to predict food insecurity, focusing on machine learning techniques suitable for hierarchical or clustered data. We primarily employed the generalized mixed-effects

random forest (GMERF) and introduced a novel modification, the Generalized Mixed-Effects Generalized Random Forest (GMEGRF), aimed at improving prediction accuracy. For comparison, we also included standard Random Forest (RF) and Generalized Random Forest (GRF) models. All models were trained on high-quality food insecurity data and validated using cross-validation techniques to ensure robust and accurate predictions. This research framework is designed in such a way as to achieve the main objective, namely to obtain food vulnerability classification prediction results that have high accuracy and are relevant to existing data conditions.

2.1 Frame work

The framework is designed with components intended for reusability in various contexts. These components are organized within a structured and well-defined workflow. This workflow ensures that each step in the prediction process is conducted systematically and consistently, leading to more reliable results and a reduction in the potential for errors. For effective research, data preprocessing was also performed, including data cleaning and reducing variables with small variances. Data preprocessing is a crucial step in preparing raw data for use in the model [25]. This framework utilizes ML algorithms, both single algorithms and mixed models, to construct the most efficient prediction model. To ensure optimal model performance, cross-validation techniques were employed. Cross-validation is a critical technique for evaluating the performance of an ML model in assessing how well the model will perform on unseen data [26]. A workflow diagram of the study is shown in Fig. 1.

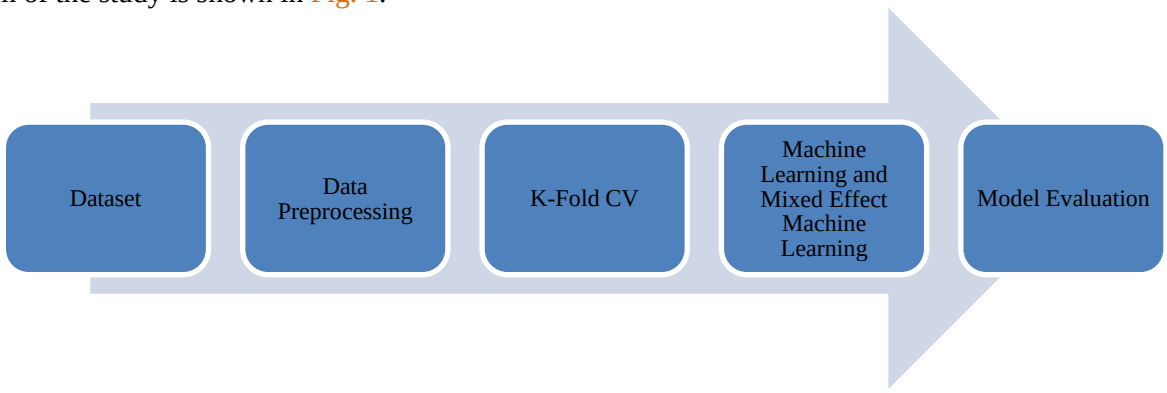

Figure 1. The Study Workflow Diagram
(Source: Smartart in word)

Fig. 1 shows the five stages of our research methodology. Stage 1 consists of a dataset of food insecurity collected through a survey in West Java, Indonesia, in 2021 Stage 2 is data preprocessing, which is a crucial step to ensure the quality of the data used for modeling. In this stage, we cleaned the dataset to ensure that the dataset is devoid of incorrect or erroneous data and ready for the next stage of analysis [27].

Additionally, variables with small variances are reduced because they provide limited information and can be removed to simplify the model and enhance computational efficiency. Near-zero variance variables either have a single, distinct value, or most of the data falls into one group [28]. Following this, feature scaling was carried out for numeric variables, transforming them using Z-score standardization. It was chosen over other methods because outliers are found in numerical variables, and this method is sensitive to outliers. This improves the computational efficiency of the machine learning models and promotes better prediction performance by preventing features with larger values from dominating the learning process. Stage 3: K-Fold Cross-Validation. K-fold cross-validation is a model evaluation technique that partitions a dataset into k equal-sized subsets known as folds. In each iteration, one-fold was designated as the testing data, whereas the remaining k-folds served as the training data. This process was repeated k times, with each fold taking a turn as the testing data [26].

Stage 4 centers on the development and evaluation of predictive models utilizing both machine learning algorithms and a mixed-effects modeling approach. Specifically, this stage explores the application of single machine learning algorithms, Random Forest and Generalized Random Forest, along with their mixed-effects counterparts, Generalized Mixed-Effects Random Forest (GMERF) and Generalized Mixed-Effects Generalized Random Forest (GMEGRF), for predictive model construction. The incorporation of mixed-effects models suggests the presence of a hierarchical or clustered structure within the data, which the researchers aim to address within their modeling framework. In stage 5 model evaluation, the performance of the chosen machine learning algorithms was evaluated using four popular values to evaluate classification

tasks: accuracy, sensitivity, specificity, and balanced accuracy. These values provide a comprehensive picture of the overall prediction accuracy, ability to identify positive and negative classes, and balance accuracy prediction, and their calculations are based on a confusion matrix [29], [30]. The confusion matrix serves as a pivotal evaluation tool for classification, summarizing the performance of a model through four critical scenarios: true positive (correct positive prediction), false positive (incorrect positive prediction), false negative (incorrect negative prediction), and true negative (correct negative prediction). It offers a profound understanding of the specific types of errors a model commits and forms the foundation for calculating more informative evaluation metrics [31], [32].

2.2 Dataset Detail

This study examines the state of household food insecurity in West Java, Indonesia, in 2021. A household was classified as food insecure if it answered "yes" to any of the eight questions in the Food Insecurity Experience Scale (FIES) survey. The data source for this study was the National Socio-Economic Survey (2021), which includes 23 fixed-effect predictor variables, one random-effect predictor variable, and one response variable: the classification of household food insecurity experience (Y) (Table 1). The dataset comprises 25,890 observations (households).

Table 1. Details of the Features Present in the Dataset

Variable	What it does?
Y	The Classification of Household Food Insecurity Experience
X_1	Number Of Household Members
X_2	Gender Of Head of Household
X_3	Age Of Head of Household
X_4	Literacy Status of Head of Household
X_5	Highest Education Level of Head of Household
X_6	Status of Bank Savings Accounts Owned
X_7	Health Insurance Contribution Assistance Recipient Status
X_8	Health Insurance Ownership Status
X_9	Smoking Status of Head of Household
X_{10}	Home Ownership Status
X_{11}	House Size
X_{12}	Type of House Wall Material
X_{13}	Type of House Floor Material
X_{14}	Adequacy of Home Sanitation
X_{15}	Feasibility of Drinking Water Source
X_{16}	People's Business Credit Recipient Status
X_{17}	Bank/Cooperative Loan Recipient Status
X_{18}	Village-Owned Enterprise Benefit Recipient Status
X_{19}	Value of House and/or Land Assets
X_{20}	Prosperous Family Card Recipient Status
X_{21}	Family Hope Program Recipient Status
X_{22}	Non-Cash Food Assistance Recipient Status
X_{23}	Other Routine Assistance Recipient Status
R	Subdistrict

2.3 Methods

The research methods utilized included Random Forest (RF), Generalized Random Forest (GRF), Generalized Mixed-Effects Random Forest (GMERF), and Generalized Mixed-Effects Generalized Random Forest (GMEGRF). Random Forest is a machine learning technique that constructs an ensemble of decision trees to generate robust and accurate predictive models. This methodology involves the iterative development of multiple decision trees, each trained on random subsets of data and features, followed by the aggregation of their respective predictions. It operates through a sequence of distinct procedural steps: initially, bootstrap sampling generates multiple data subsets from the original dataset using randomized sampling with replacement/bootstrap sample; subsequently, decision tree (Dt) construction involves the development of individual decision trees for each of these subsets; each tree undergoes tree growth, expanding to its maximum potential without pruning, ensuring high variance; the algorithm then performs prediction, synthesizing the outputs of all constituent trees to formulate a final predictive result (majority vote); and finally, out-of-bag (OOB) error estimation evaluates the model's performance by utilizing data points excluded from the training

of individual trees, providing a robust measure of generalization. The illustration schematic diagram of the RF algorithm can be seen in Fig. 2 [33].

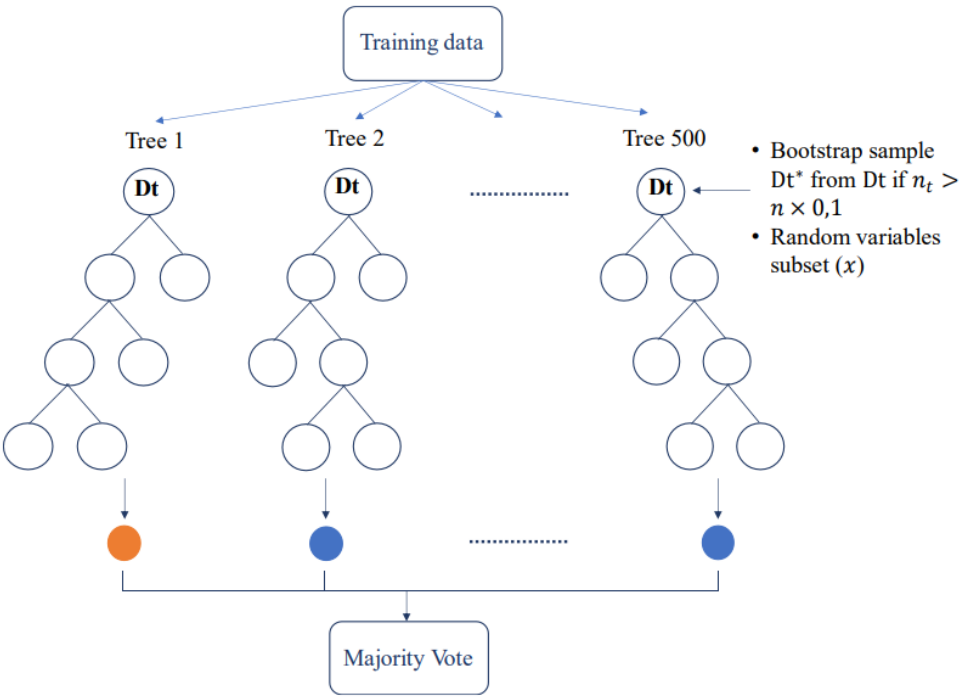

Figure 2. Illustration Schematic of RF Algorithm
(Source: A. Khairunnisa *et al.*, 2024)

The Generalized Random Forest (GRF) adapts the traditional RF framework through modified procedural steps tailored to specific statistical estimation tasks; initially, data may be partitioned for "honest" estimation, followed by tree construction where splitting criteria are adapted to the estimation task, such as heterogeneous treatment effects, utilizing information from local moment equations; subsequently, leaf node value assignments are determined based on the estimation goal, potentially involving calculations of treatment effects or other relevant quantities; finally, predictions from individual trees are aggregated, and methods are employed to calculate confidence intervals and perform statistical inference, thereby enabling the estimation of a broader range of statistical quantities beyond simple conditional means [22], [24].

Generalized mixed-effects random forest (GMERF) models represent an important advancement in machine learning techniques, particularly for datasets with complex hierarchical structures or correlated random effects. The GMERF algorithm aims to build an accurate predictive model for data with hierarchical or clustered structures by considering both random and fixed effects. The steps are as follows: first, identify the data structure and then initialize random effects, where random effects, which represent variations between groups, are initialized. This can be either a given initial value or initialized as zero. These random effects are iteratively updated during the algorithm process. Then, fit initial GLM Model, where an initial Generalized Linear Model (GLM) is fitted to obtain an initial estimate of the predictions. This model uses all covariates (predictor variables) but does not yet account for random effects. The algorithm then enters an iterative process in which the model is continuously updated until it reaches convergence (stable results) or a predetermined maximum number of iterations. Within the iterative sequence, it calculates the target value, fits random forest, fits the GLMM, and performs convergence checking. After the iterations are complete, the algorithm produces the final GMERF model, which includes the fitted GLMM, fitted RF model, and the final estimation of the random effects. In essence, The GMERF algorithm iteratively combines information from RF and GLMM to update the estimation of random effects and improve prediction accuracy. This process allows the model to capture nonlinear relationships in the data and consider hierarchical structures, resulting in a more accurate and comprehensive model [13].

Given $y_{ij} = \left(y_{i1}, y_{i2}, \ldots, y_{in_j}\right)$ with observation units $i, i = 1,2,\ldots,n_j$, in groups $j, j = 1,2,\ldots,J$. Y_{ij} is assumed to follow an exponential family distribution using Eq. (1), as follows:

$$f_i(y_{ij}|b_j) = exp\left\{\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} + c(y_{ij},\phi)\right\}, \quad (1)$$

where b_j is a random component, a and c are specific functions, η_{ij} is a natural parameter, and ϕ is a dispersion parameter. The mean and variance of y_{ij} are respectively $E(y_{ij}|b_j) = a'(\eta_{ij}) = \mu_{ij}$ and $Var(Y_{ij}|b_j) = \psi a''(\eta_{ij})$ [13].

Given the canonical function $g(a')^{-1}$, which connects the mean with the systematic component, the GLMM formula can be expressed by Eq. (2).

$$\begin{gathered} \mu_j = E(y_j|b_j), j = 1,2,\dots,J, \\ g(\mu_j) = \eta_j, \\ \eta_j = X_j\beta + Z_j b_j, \\ b_j \sim N_Q(0,\psi), \end{gathered} \quad (2)$$

where j is the group index, and J is the number of groups. n_j is the number of observations in the j-th group and $\sum_{j=1}^{J} n_j = J$. η_j is a linear predictor vector of dimension n_j, where X_j is a fixed-effect predictor variable matrix of size $n_j \times P$, $P = p + 1$ and β is a coefficient vector of predictor variables of size P. Z_j is a $n_j \times Q$ matrix of random effects regression, b_j is a Q-dimensional vector of coefficients (including random intercepts), and ψ is a $Q \times Q$ matrix of the variance of random effects. Fixed effects were identified using parameters related to the entire population, whereas random effects were identified using group-specific parameters. Estimation methods include the maximum likelihood, restricted maximum likelihood, and penalized quasi-likelihood [13], [16].

The generalized mixed effect random forest (GMERF) presented by Pallagati *et al.* (2021) [13]. Fundamentally, the GMERF replaces the linear function of fixed effects in a traditional GLMM with a RF method. Given the canonical function $g(a')^{-1}$, which connects the mean with the systematic component, the GMERF model can be expressed by Eq. (3).

$$\begin{gathered} \mu_j = E(y_j|b_j), j = 1,2,\dots,J, \\ g(\mu_j) = \eta_j, \\ \eta_j = f(x_j) + Z_j b_j, \\ b_j \sim N_Q(0,\psi), \end{gathered} \quad (3)$$

Consequently, the GMERF extends the capabilities of GLMMs by enabling the modelling of nonlinear and interactive fixed effects through the integration of an ensemble tree structure, while still effectively addressing dependencies within grouped data via random effects.

The modification from Generalized Mixed-Effects Random Forest (GMERF) to Generalized Mixed-Effects Generalized Random Forest (GMEGRF) entails the integration of Generalized Random Forest (GRF) components within the established GMERF framework. GMERF, a hybrid of Generalized Linear Mixed Models (GLMM) and RF, is augmented by GRF's capacity to address heterogeneous treatment effects and broader statistical estimation tasks. This modification is realized by substituting the conventional 'fitting random forest' step in the standard GMERF with a modified GRF procedure. GMEGRF can be expressed by Equation (3), where function $f(x_j)$ in $\eta_j = f(x_j) + Z_j b_j$ in Equation (3) is computed using the GRF algorithm in GMEGRF method. The initial approach for estimating the parameters of a GMEGRF involved an iterative algorithm that alternated between fitting a GRF for fixed effects and a generalized linear mixed model for random effects. Through this integration, GMEGRF is anticipated to enhance predictive accuracy, effectively manage complex nonlinearities, and simultaneously account for hierarchical structures and random effects within the data, thereby yielding a more comprehensive and precise model. The GMEGRF and GRF algorithms as follows:

Algorithm 1: **Generalized Mixed Effect Generalized Random Forest**

Input:

y- vector with responses y_{ij}

cov- data frame with all covariates

gr- vector with the grouping variable for each observations

Algorithm 1: **Generalized Mixed Effect Generalized Random Forest**

$znam$- vector with names of covariates to be used as random affects
$xnam$- vector with names of covariates to be used as fixed affects
fam- distribution of y
b_0- optional matrix of initial values for each $\underline{b_i}$
$toll$- threshold to decide whether our estimation coveraged or not
$itmax$- maximum number of iterations

$Z \leftarrow (1; cov[znam])$ {to include also the random intercept}

Initialize b to a matrix of $zero$ (if is b_0 not given) {Each column b$[i,]$ of b will be the i-th random coefficients $\underline{b_i}$}
$all.b[0] = b$

fit a GLM model using y as response and cov as matrix of covariates
$eta \leftarrow$ estimated η_{ij} by the GLM model
$i \leftarrow 1$
while $it < itmax$ and not $conv$ do
  $targ \leftarrow eta - Z \times b$
  fit a GRF model using $targ$ as target and cov as predictor matrix
  $f\,x \leftarrow$ fitted values of the GRF model
  fit the GLMM $\eta_{ij} - f(x_{ij}) = \underline{Z_{ij}^T} \times \underline{b_i}$
  $all.b[it] \leftarrow b \leftarrow$ the estimated b from the model
  $M \leftarrow max(abs(b - all.b[it - 1]))$
  $(i,j) \leftarrow argmax(abs(b - all.b[it - 1]))$
  $tr \leftarrow M \,/\, all.b[it - 1][i,j]$
  if $tr < toll$ then
    $conv \leftarrow$ true
  else
    $conv \leftarrow$ false
  end if
  $it \leftarrow i + 1$
end while
if not $conv$ then
  give a warning
end if

Algorithm 2: Generalized Random Forest with Honesty and Subsampling

All tuning parameters are prespecified, including the number of trees C and the subsampling rate s used in Subsample.
Procedure GRF
set of examples S, test point x, weight vector $\alpha \leftarrow ZEROS(|\,S\,|)$
for $c = 1$ to total number of trees C do
set of examples $I \leftarrow SUBSAMPLE(S, s)$
sets of examples $J_1, J_2 \leftarrow SPLITSAMPLE(I)$
tree $T \leftarrow GRADIENTTREE(J_1, X)$
$N \leftarrow NEIGHBORS(x, T, J_2)$
Returns those elements of J_2 that fall into the same leaf as x in the tree T
for all example $e \in N$ do
  $\alpha[e] \mathrel{+}= 1/|\,N\,|$
Output:
$\hat{\theta}(x)$, the solution to (2) with weights α/C

Table 2 summarizes the modeling approaches considered in this study by presenting the type of model each method represents and its ability to support mixed-effects modeling. This comparison is essential to understand the structural differences between methods, especially in the context of hierarchical or cluster data.

Table 2. The Summary Methods

Method	Model Type	Support for Mixed Models
Random Forest (RF)	Ensemble Decision Tree	No
Generalized Random Forest (GRF)	Generalization of RF for local parameter estimation	No
GMERF	Random Forest + Mixed Effects	Yes
GMEGRF	GRF + Mixed Effects	Yes

3. RESULTS AND DISCUSSION

This section is divided into several subsections. The first subsection concerns the results of the descriptive analysis, the second subsection concerns the model results, and finally, the discussion.

3.1 Summary Statistics

The analysis in this study was conducted on 25,873 of the 25,890 observations (households) initially included in the data. This was a result of the data cleansing process, which involved the deletion of household records containing 'do not know' or 'no responses.' The collected data were analyzed using advanced statistical techniques to interpret the effectiveness of the methods used. The proportion of Y can be expressed as follows:

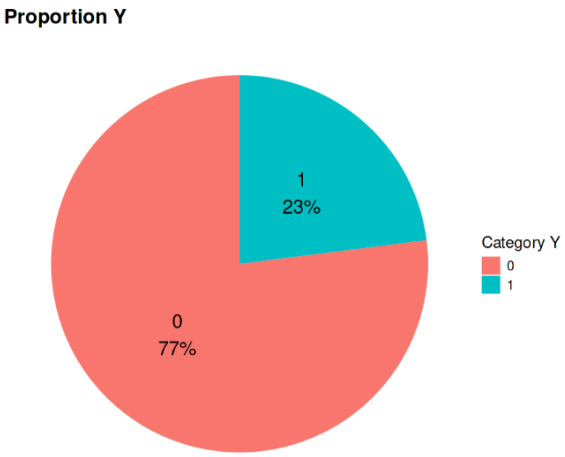

Figure 3. The Proportion of Y
(Source: R Application in kaggle)

Fig. 3 shows that 77% of households are classified as food secure (category 0), and 23% of households experience food insecurity (category 1). Although the majority of households (77%) are classified as food secure, 23% of households still experience food insecurity. This figure is significant considering the negative impacts of food insecurity on individuals' health, education, and productivity. This condition requires serious attention as it can hinder the achievement of sustainable development goals, particularly SDG 2, which targets zero hunger.

Prior to model development, the variance of the explanatory variables was assessed to identify potential near-zero variance predictors (NZV). The X4 (literacy status of head of household) and X18 (village-owned enterprise benefit recipient status) exhibited exceedingly low variances, indicating a lack of meaningful variability within these predictors, primarily because most observations shared a single, predominant value, indicating a lack of meaningful variability within these predictors. Consequently, to mitigate potential issues of model instability and ensure the inclusion of informative variables, X4 and X18 were excluded from subsequent analyses. This decision was based on the premise that variables with near-zero variance contribute minimally to the predictive capability of the model and can adversely affect its robustness.

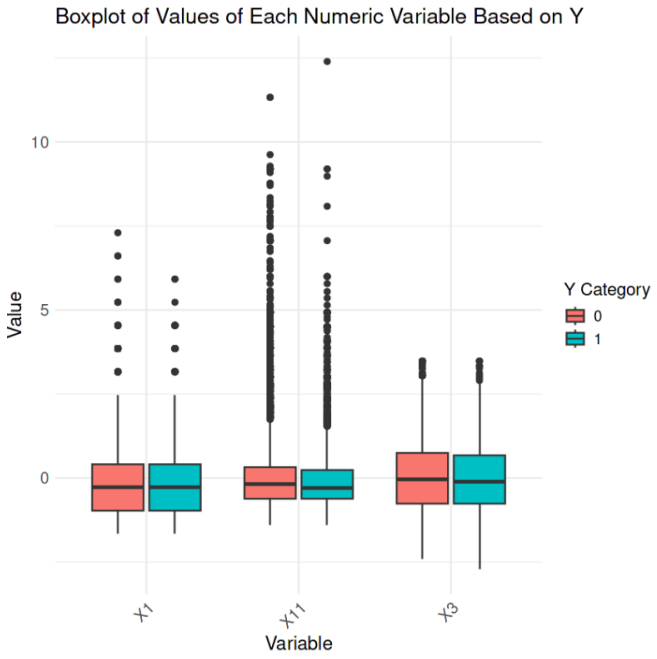

Figure 4. The Boxplot of Numeric Variable Predictor
(Source: R Application in kaggle)

Fig. 4 visualizes the distribution of values for the numerical variables number of household members (X_1), age of head of household (X_3), and house size (X_{11}), grouped by category Y (0 and 1). In general, the medians for all three variables tend not to differ significantly between categories Y=0 and Y=1, indicating that the average number of household members, age of head of household, and house size are not significantly different between the two groups. However, there were significant differences in the data spread, particularly in the house size variable (X_{11}). For variable X_{11}, a very wide data spread and numerous outliers were observed in category Y=1, indicating extreme variations in house size within that group. This suggests that although the average house size does not differ significantly, there are houses with significantly larger or smaller sizes in group Y=1 than in group Y=0.

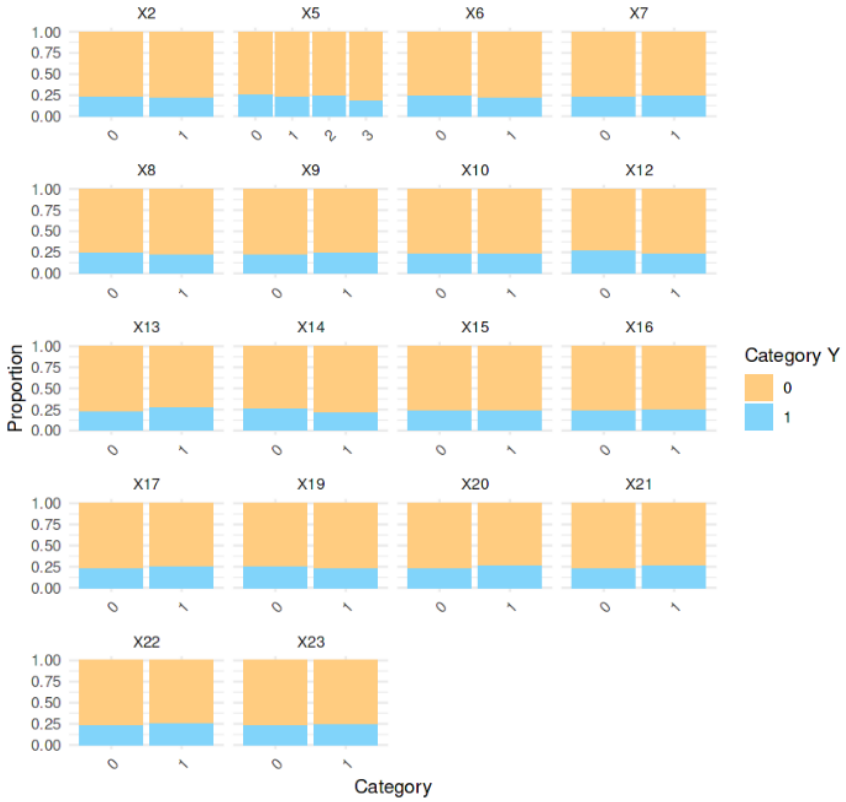

Figure 5. The Proportion of Category X by Category Y
(Source: R Application in kaggle)

Fig. 5 shows that the graph visualizes the proportions of categorical predictor variables (X_2, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}, X_{19}, X_{20}, X_{21}, X_{22}, and X_{23}) for each value of category Y (0 and 1). This graph allows us to compare the distribution of predictor variables between the two groups of category Y. In general, there were significant differences in proportions between categories Y=0 and Y=1 for most predictor variables. To understand this relationship in more detail, statistical modeling was performed.

3.2 Modeling Results

This study employed standard machine learning techniques, specifically Random Forest (RF) and Generalized Random Forest (GRF), and mixed-effects machine learning approaches, namely Generalized Mixed-Effects Random Forest (GMERF) and Generalized Mixed-Effects Random Forest (GMEGRF), to model food insecurity data. In the RF and GRF models, the variable 'R' was operationalized as a fixed-effects predictor. Conversely, the mixed-effects models, GMERF and GMEGRF, considered 'R' as a random effects predictor to account for the inherent hierarchical structure of the data, wherein households within the same 'R' exhibited greater homogeneity compared to those in different 'R' units. The presence of significant sub-district-level random effects was statistically validated through the Likelihood Ratio Test and the Intraclass Correlation Coefficient (ICC), yielding a significant result ($p < 0.05$, ICC = 36.06%). Subsequently, the predictive capabilities of all the models were evaluated using a suite of performance metrics, including accuracy, sensitivity, specificity, and balanced accuracy.

Table 3. Performance Measure

Algorithm		Average Performance Measure			
		Accuracy	**Sensitivity**	**Specificity**	**Balance Accuracy**
Machine Learning	RF	0.7022422	0.8377061	0.2485139	0.5431100
	GRF	0.7695784	0.9699956	0.0984168	0.5342062
Mixed Models	GMERF	0.7141011	0.7595313	0.5940777	0.6768045
	GMEGRF	0.7497661	0.8100859	0.5478559	0.6789709

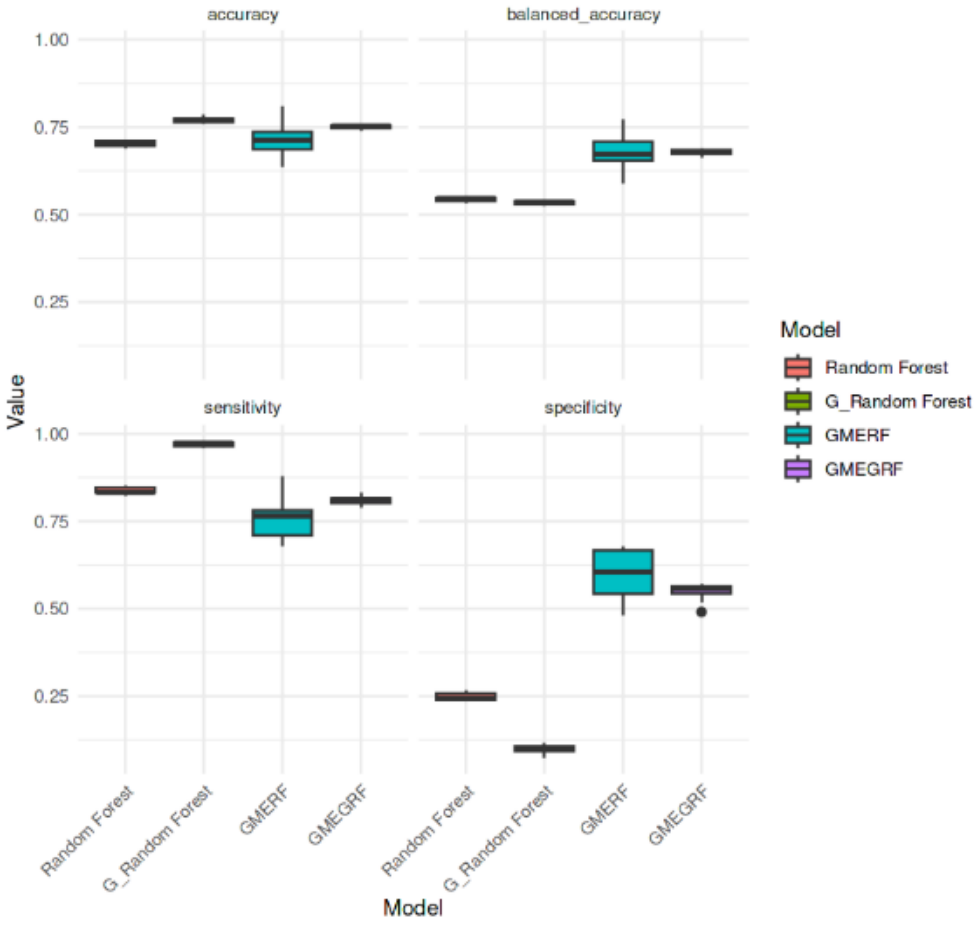

Figure 6. The Boxplots Comparing the Distribution of Evaluation Metrics Across Different Models *(Source: R Application in kaggle)*

Table 3 presents the average performance metrics of the four predictive models used in this study. To provide a more nuanced understanding of the performance distribution across these models, Fig. 6 shows the metrics (accuracy, sensitivity, specificity, and balanced accuracy) obtained using boxplots for each algorithm. The integration of information from both the table and figure facilitates a more comprehensive analysis of the strengths and limitations inherent in each modeling approach, particularly considering how the mixed-effects framework models random effects within the data.

Based on Table 3 and the visualizations in Fig. 6, the Generalized Random Forest (GRF) model exhibits an intriguing performance profile. The GRF demonstrated the highest average accuracy (0.7696) and excellent sensitivity (0.9700), indicating its superior capability in identifying positive cases. However, these models demonstrated a low specificity (0.0984), which was visually corroborated by the specificity boxplot, where the GRF values were concentrated at the lower end of the scale. This low specificity reveals a significant bias in GRF towards the majority class, wherein the model tends to misclassify negative instances as positive. Consequently, despite its high overall accuracy, the balanced accuracy of GRF (0.5342) was relatively low, reflecting an imbalance in its ability to predict both classes equally. Bias is particularly problematic given the imbalanced nature of our dataset, where the majority class (food secure) constitutes 77% of the observations, while the minority class (food insecure) is only 23%.

Conversely, mixed-effects approaches, represented by the Generalized Mixed-Effects Random Forest (GMERF) and, notably, the Generalized Mixed-Effects Random Forest (GMEGRF), offer a more balanced performance profile. This model is an extension of the machine learning model that incorporates a mixed-effects model structure into the tree-building process. Although their average accuracy and sensitivity were slightly lower than those of GRF, both models consistently exhibited substantially improved specificity, which was also confirmed by the specificity boxplots showing a higher distribution of values. Considering the enhanced equilibrium between sensitivity and specificity, coupled with a superior balanced accuracy, the Generalized Mixed-Effects Random Forest (GMEGRF) has emerged as the most optimal and recommended model for this classification task. The model includes fixed effects (the general effect of predictor variables) and random effects (variability between groups or clusters, such as sub-districts), allowing it to capture differences between groups. When the data has a clustered or hierarchical structure, or when there is heterogeneity between groups that influences the outcome (response), this model is more appropriate than RF or GRF and mitigates potential class imbalance. Table 3 reports an accuracy of 0.7498 and the highest balanced accuracy (0.6790) for GMEGRF among all models. This finding is further supported by the balanced accuracy boxplot, which illustrates a higher and more stable distribution of the values for GMEGRF. In addition to the evaluation results, the GMEGRF has a faster computation time compared to the GMERF.

3.3 Interpretation and Discussion

In this study, the optimal model employed was a modified mixed-effects model, the Generalized Mixed-Effects Generalized Random Forest (GMEGRF). We have attempted to extract a specific output from the GMEGRF model. This model effectively segregates the variance in outcomes attributed to directly measured factors (fixed effects) from the variance resulting from interdistrict group differences (random effects). The random effects variance, quantified at 1.812, indicated a statistically significant variation across districts in influencing the research outcomes. This suggests that each district possesses unique characteristics that contribute to the observed differences in results, which cannot be solely explained by directly measured factors.

The Generalized Mixed-Effects Generalized Random Forest (GMEGRF) model employed in this study yielded a fixed effects intercept estimate of -0.26286, with a highly significant p-value of 1.19e-05. This indicates that the average log odds of the event, excluding random effects across district groups, is -0.26286. The model achieved convergence after four iterations, signifying that the optimization algorithm successfully identified a stable and efficient solution. Furthermore, the model demonstrated consistent performance.

Importantly, the output of the model offers practical value for real-world decision-making. The random effects component highlights which districts deviate most significantly from the average trend, allowing policymakers to target specific areas that may require additional support or intervention. Districts with unusually high or low random effects may signal underlying contextual factors—such as infrastructure, accessibility, or local policies—that warrant closer attention.

On the other hand, the fixed effects provide insight into household-level variables that are consistently associated with the outcome of interest. These variables can guide the design of household-focused programs, such as nutritional aid, health interventions, or educational outreach, by identifying which characteristics most strongly influence food insecurity or other target outcomes.

In summary, the GMEGRF model not only offers strong predictive performance but also yields interpretable components that can inform both area-based and individual-level interventions. By integrating these outputs into planning and resource allocation, stakeholders can make more informed, data-driven decisions that reflect the heterogeneity across and within districts.

4. CONCLUSION

This study successfully developed a Generalized Mixed-Effects Generalized Random Forest (GMEGRF) model, which demonstrated superior predictive performance for food insecurity compared to Random Forest (RF), Generalized Random Forest (GRF), and Generalized Mixed-Effects Random Forest (GMERF) models. The efficacy of the GMEGRF model is attributed to its ability to effectively partition variance between fixed and random effects, manage the hierarchical structure inherent in food insecurity data, produce balanced predictive outcomes, and achieve efficient model convergence. These findings establish GMEGRF as a robust and accurate tool for food insecurity prediction, offering valuable insights for policy formulation and effective interventions while underscoring the critical importance of addressing data structure and class imbalance.

To enhance model performance, particularly in scenarios involving extreme class imbalance, it is recommended to consider employing techniques such as oversampling, undersampling, or the application of class weights. Extremely imbalanced datasets often overlook minority classes, which are often the primary focus of analysis. By addressing the dominance of the majority class, the model becomes more inclusive and sensitive to groups requiring special attention.

Author Contributions

Herlin Fransiska: Conceptualization, Methodology, Writing-Original Draft, Software, Validation. Agus Mohamad Soleh: Software, Resources, Draft Preparation. Khairil Anwar Notodiputro: Formal analysis, Validation, and Resources. Erfiani: Visualization and Resources. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received a specific grant from the Directorate General of Higher Education, Research, and Technology of the Ministry of Education, Culture, Research, and Technology for funding this research through the 2024 Doctoral Research Scheme in accordance with Research Contract Number: 027/E5/PG.02.00.PL/2024 dated June 11, 2024.

Acknowledgment

This research was made possible by the generous support of the School of Data Science, Mathematics, and Informatics, IPB University; the Faculty of Mathematics and Sciences, Bengkulu University; and Badan Pusat Statistik (Statistics Indonesia).

Declarations

The authors declare no competing interest.

Declaration of Generative AI and AI-assisted Technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement (grammar, spelling, and clarity). The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. The authors reviewed and approved all final text

REFERENCES

[1] K. P. Myers and J. L. Temple, "TRANSLATIONAL SCIENCE APPROACHES FOR FOOD INSECURITY RESEARCH," *Appetite*, vol. 200, p. 107513, Sep. 2024, doi: https://doi.org/10.1016/j.appet.2024.107513.

[2] G. Nica-Avram, J. Harvey, G. Smith, A. Smith, and J. Goulding, "IDENTIFYING FOOD INSECURITY IN FOOD SHARING NETWORKS VIA MACHINE LEARNING," *J Bus Res*, vol. 131, pp. 469–484, Jul. 2021, doi: https://doi.org/10.1016/j.jbusres.2020.09.028.

[3] A. H. Villacis, S. Badruddoza, A. K. Mishra, and J. Mayorga, "THE ROLE OF RECALL PERIODS WHEN PREDICTING FOOD INSECURITY: A MACHINE LEARNING APPLICATION IN NIGERIA," *Glob Food Sec*, vol. 36, p. 100671, Mar. 2023, doi: https://doi.org/10.1016/j.gfs.2023.100671.

[4] C. Gao, C. J. Fei, B. A. McCarl, and D. J. Leatham, "IDENTIFYING VULNERABLE HOUSEHOLDS USING MACHINE-LEARNING," *Sustainability (Switzerland)*, vol. 12, no. 15, Aug. 2020, doi: https://doi.org/10.3390/su12156002.

[5] S. Gholami *et al.*, "FOOD SECURITY ANALYSIS AND FORECASTING: A MACHINE LEARNING CASE STUDY IN SOUTHERN MALAWI," *Data Policy*, vol. 4, no. 3, Oct. 2022, doi: https://doi.org/10.1017/dap.2022.25.

[6] J. J. L. Westerveld *et al*, "FORECASTING TRANSITIONS IN THE STATE OF FOOD SECURITY WITH MACHINE LEARNING USING TRANSFERABLE FEATURES," *Science of The Total Environment*, vol. 786, p. 147366, Sep. 2021, doi: https://doi.org/10.1016/j.scitotenv.2021.147366.

[7] X. Shu and Y. Ye, "KNOWLEDGE DISCOVERY: METHODS FROM DATA MINING AND MACHINE LEARNING," *Soc Sci Res*, vol. 110, p. 102817, Feb. 2023, doi: https://doi.org/10.1016/j.ssresearch.2022.102817.

[8] A. Hajjem, F. Bellavance, and D. Larocque, "MIXED EFFECTS REGRESSION TREES FOR CLUSTERED DATA," *Stat Probab Lett*, vol. 81, no. 4, pp. 451–459, Apr. 2011, doi: https://doi.org/10.1016/j.spl.2010.12.003.

[9] A. Hajjem, F. Bellavance, and D. Larocque, "MIXED-EFFECTS RANDOM FOREST FOR CLUSTERED DATA," *J Stat Comput Simul*, vol. 84, no. 6, pp. 1313–1328, Jun. 2014, doi: https://doi.org/10.1080/00949655.2012.741599.

[10] A. Hajjem, D. Larocque, and F. Bellavance, "GENERALIZED MIXED EFFECTS REGRESSION TREES," *Stat Probab Lett*, vol. 126, pp. 114–118, Jul. 2017, doi: https://doi.org/10.1016/j.spl.2017.02.033.

[11] J. Hu and S. Szymczak, "A REVIEW ON LONGITUDINAL DATA ANALYSIS WITH RANDOM FOREST," *Brief Bioinform*, vol. 24, no. 2, pp. 1–11, Mar. 2023, doi: https://doi.org/10.1093/bib/bbad002.

[12] P. Krennmair and T. Schmid, "FLEXIBLE DOMAIN PREDICTION USING MIXED EFFECTS RANDOM FORESTS," *J R Stat Soc Ser C Appl Stat*, vol. 71, no. 5, pp. 1865–1894, Nov. 2022, doi: https://doi.org/10.1111/rssc.12600.

[13] M. Pellagatti, C. Masci, F. Ieva, and A. M. Paganoni, "GENERALIZED MIXED-EFFECTS RANDOM FOREST: A FLEXIBLE APPROACH TO PREDICT UNIVERSITY STUDENT DROPOUT," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 14, no. 3, pp. 241–257, Jun. 2021, doi: https://doi.org/10.1002/sam.11505.

[14] R. J. Sela and J. S. Simonoff, "RE-EM TREES: A DATA MINING APPROACH FOR LONGITUDINAL AND CLUSTERED DATA," *Mach Learn*, vol. 86, no. 2, pp. 169–207, Feb. 2012, doi: https://doi.org/10.1007/s10994-011-5258-3.

[15] J. L. Speiser *et al*, "BIMM TREE: A DECISION TREE METHOD FOR MODELING CLUSTERED AND LONGITUDINAL BINARY OUTCOMES," *Commun Stat Simul Comput*, vol. 49, no. 4, pp. 1004–1023, Apr. 2020, doi: https://doi.org/10.1080/03610918.2018.1490429.

[16] L. Fontana, C. Masci, F. Ieva, and A. M. Paganoni, "PERFORMING LEARNING ANALYTICS VIA GENERALISED MIXED-EFFECTS TREES," *Data (Basel)*, vol. 6, no. 7, p. 74, Jul. 2021, doi: https://doi.org/10.3390/data6070074.

[17] D. Kusumaningrum *et al*, "FOUR-PARAMETER BETA MIXED MODELS WITH SURVEY AND SENTINEL 2A SATELLITE DATA FOR PREDICTING PADDY PRODUCTIVITY," *Smart Agricultural Technology*, vol. 9, Dec. 2024, doi: https://doi.org/10.1016/j.atech.2024.100525.

[18] P. C. Chen, M. M. Yu, J. C. Shih, C. C. Chang, and S. H. Hsu, "A REASSESSMENT OF THE GLOBAL FOOD SECURITY INDEX BY USING A HIERARCHICAL DATA ENVELOPMENT ANALYSIS APPROACH," *Eur J Oper Res*, vol. 272, no. 2, pp. 687–698, Jan. 2019, doi: https://doi.org/10.1016/j.ejor.2018.06.045.

[19] L. Breiman, "RANDOM FORESTS," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: https://doi.org/10.1023/A:1010933404324.

[20] S. W. Raudenbush and A. S. Bryk, "HIERARCHICAL LINEAR MODELS: APPLICATIONS AND DATA ANALYSIS METHODS," *Applications and data analysis methods (Vol. 1)*, 2002.doi: https://doi.org/10.3758/s13428-017-0971-x

[21] M. Fokkema, N. Smits, A. Zeileis, T. Hothorn, and H. Kelderman, "DETECTING TREATMENT-SUBGROUP INTERACTIONS IN CLUSTERED DATA WITH GENERALIZED LINEAR MIXED-EFFECTS MODEL TREES," *Behav Res Methods*, vol. 50, no. 5, pp. 2016–2034, 2018, doi: 10.3758/s13428-017-0971-x.

[22] S. Athey, J. Tibshirani, and S. Wager, "GENERALIZED RANDOM FORESTS," *https://doi.org/10.1214/18-AOS1709*, vol. 47, no. 2, pp. 1148–1178, Apr. 2019, doi: https://doi.org/10.1214/18-AOS1709.

[23] E. Zhou and D. Lee, "GENERATIVE ARTIFICIAL INTELLIGENCE, HUMAN CREATIVITY, AND ART," *PNAS Nexus*, vol. 3, no. 3, Mar. 2024, doi: https://doi.org/10.1093/pnasnexus/pgae052.

[24] H. Fransiska, A. M. Soleh, K. A. Notodiputro, and Erfiani, "EVALUATION OF MACHINE LEARNING MODELS BASED ON HOUSEHOLD FOOD INSECURITY DATA IN INDONESIA," in *BIO Web of Conferences*, EDP Sciences, Apr. 2025. doi: https://doi.org/10.1051/bioconf/202517102011.

[25] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "BIG DATA PREPROCESSING: METHODS AND PROSPECTS," *Big Data Anal*, vol. 1, no. 1, Dec. 2016, doi: https://doi.org/10.1186/s41044-016-0014-0.

[26] I. K. Nti, O. Nyarko-Boateng, and J. Aning, "PERFORMANCE OF MACHINE LEARNING ALGORITHMS WITH DIFFERENT K VALUES IN K-FOLD CROSSVALIDATION," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: https://doi.org/10.5815/ijitcs.2021.06.05.

[27] G. Y. Lee, L. Alzamil, B. Doskenov, and A. Termehchy, "A SURVEY ON DATA CLEANING METHODS FOR IMPROVED MACHINE LEARNING MODEL PERFORMANCE," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.07127

[28] P. Agasthi *et al* "PREDICTION OF PERMANENT PACEMAKER IMPLANTATION AFTER TRANSCATHETER AORTIC VALVE REPLACEMENT: THE ROLE OF MACHINE LEARNING," *World J Cardiol*, vol. 15, no. 3, pp. 95–105, Mar. 2023, doi: https://doi.org/10.4330/wjc.v15.i3.95.

[29] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "MULTI-LABEL CLASSIFIER PERFORMANCE EVALUATION WITH CONFUSION MATRIX," ACADEMY AND INDUSTRY RESEARCH COLLABORATION CENTER (AIRCC), Jun. 2020, pp. 01–14. doi: https://doi.org/10.5121/csit.2020.100801.

[30] S. H. Hasanah *et al*, "GOJEK DATA ANALYSIS THROUGH TEXT MINING USING SUPPORT VECTOR MACHINE (SVM) AND K-NEAREST NEIGHBOR (KNN)," *BAREKENG: J. Math. & App*, vol. 19, no. 2, pp. 889–0902, 2025, doi: https://doi.org/10.30598/barekengvol19iss2pp889-902.

[31] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: MULTI-LABEL CONFUSION MATRIX," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3151048.

[32] I. Sriliana, S. Nugroho, W. Agwil, and E. D. Sihombing, "EVALUATION OF MULTIVARIATE ADAPTIVE REGRESSION SPLINES ON IMBALANCED DATASET FOR POVERTY CLASSIFICATION IN BENGKULU PROVINCE," *Barekeng*, vol. 19, no. 2, pp. 1143–1156, Jun. 2025, doi: https://doi.org/10.30598/barekengvol19iss2pp1143-1156.

[33] H. A. Salman, A. Kalakech, and A. Steiti, "RANDOM FOREST ALGORITHM OVERVIEW," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: https://doi.org/10.58496/BJML/2024/007.