Eksakta : Berkala Ilmiah Bidang MIPA
VOLUME 27 NO 04 2026, pp 479-494
ISSN : Print 1411-3724 — Online 2549-7464
DOI : https://doi.org/10.24036/eksakta/vol27-iss04/667

Eksakta
Berkala Ilmiah Bidang MIPA
http://www.eksakta.ppj.unp.ac.id/index.php/eksakta

Article

Comparative Performance of Gradient Boosting Algorithms for Household Food Resilience Classification during The COVID-19

Article Info

Article history :
Received March 16, 2026
Revised April 01, 2026
Accepted April 04, 2026
Published August 30, 2026
In Press

Keywords :
COVID-19,
food resilience,
gradient boosting,
machine learning,
SHAP

Nabila Syukri[1], Sachnaz Desta Oktarina[1*], Septian Rahardiantoro[1]

[1]Study Program of Statistics and Data Science, School of Data Science, Faculty of Mathematics, and Informatics, IPB University, Bogor, Indonesia

Abstract. The need for data-driven analysis to understand socio-economic vulnerability, particularly in relation to food security, has intensified due to the pandemic. The growing volume of survey data demands analytical methods that can capture multidimensional relationships. Over time, food security evolves into the concept of food resilience, reflecting a household's capacity to withstand or recover from adverse conditions. This study uses machine learning techniques to categorize household food resilience, based on data from the World Bank's High Frequency Phone Survey (HFPS), covered 2,868 households in Indonesia. A comparative evaluation of gradient boosting algorithms (XGBoost, LightGBM and CatBoost) was conducted. Model performance was evaluated using accuracy, sensitivity, specificity, and AUC across ten repeated train-test splits, with statistical significance assessed using Friedman and Wilcoxon tests. The results show that CatBoost performed best and most consistently, achieving mean accuracy of 0.7592 and mean AUC of 0.8331, which is significantly higher than that of competing models. SHAP analysis further indicates that baseline vulnerability, financial concerns, income capacity, food price shocks and unmet healthcare needs are important features for identifying resilient households. These findings demonstrate that gradient boosting, particularly CatBoost, provides strong predictive power and interpretability to support data-driven decision-making in classifying food resilience.

Corresponding Author :
Sachnaz Desta Oktarina
Study Program of Statistics and Data Science, School of Data Science, Faculty of Mathematics, and Informatics, IPB University, Bogor, Indonesia
Email : sachnazdes@apps.ipb.ac.id

1. Introduction

Food security is a major global concern and is explicitly emphasized in the Sustainable Development Goals (SDG 2: Zero Hunger) [1-2]. However, food security is highly vulnerable to external factors, like climate change, political conflicts, economic disruptions, health crises, and pandemics [3-4]. One of the recent global shocks is COVID-19, represented the most disruptive and impacting food systems, with Indonesia experiencing significant consequences [5-6]. National statistic shows that Prevalence of Undernourishment (PoU) had been decreased trend from 10.73% in 2015 to 7.63% in 2019, but increased sharply to 8.34% in 2020, highlighting the negative impact of the pandemic on food security [7-8].

In crisis situation, monitoring socio-economic vulnerability requires a shift from static measures of food security to the concept of food resilience. Resilience is defined as the capacity to anticipate, absorb, recover from, and adapt to disruptions [9]. While traditional food security focuses on a snapshot in time, resilience emphasizes the capacity to recover and maintain stability despite ongoing shocks [10]. In Indonesia, although food security has been widely studied, the concept of food resilience has received comparatively less attention, despite its increasing relevance during the pandemic period [11].

Analyzing food resilience is an inherently complex task, as it involves multidimensional and dynamic factors [12]. Conventional statistical approaches often rely on linear assumption that may inadequately capture nonlinear relationships and interactions among variables[13-14]. In recent years, machine learning approaches have been increasingly applied in socio-economic research due to their flexibility in modelling complex patterns and handling high-dimensional data [15-16]. Among these methods, gradient boosting algorithms have demonstrated strong predictive performance by effectively capturing nonlinear relationships and feature interactions [17]. Algorithms such as eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGBM), and Categorical Boosting (CatBoost) have demonstrated superior predictive performance across various classification problems, making them suitable for identifying resilient and non-resilient households.

Previous studies in Indonesia have applied machine learning techniques primarily to classify food security status using cross-sectional dataset such as SUSENAS. For instance, Dharmawan et al [18] conducted research on classifying food insecurity using Random Forest, SVM, ANN and XGBoost, reporting Random Forest as the best model with of accuracy of 79%. Similarly, Khikmah et al [19] evaluated several ensemble methods, including Random Forest, Gradient Boosting, Extra Trees, and Rotation Forest, and found random Forest as the best performance with balanced accuracy of 65.79%. Fransiska et al [20] extended this line of research by comparing Random Forest and Generalized Random Forest (GRF) to predict household food insecurity, demonstrating that GRF achieved better performance in with balanced accuracy of 55%.

Although these studies demonstrate the potential of machine learning in food security analysis, they predominantly focus on cross-sectional data and emphasize broad comparisons across ensemble learning families, particularly between bagging and boosting approaches. As a result, they tend to capture static aspects of food insecurity and may not adequately represent the dynamic nature of household resilience under prolonged shocks. In contrast, this study focuses on an internal comparison among gradient boosting algorithms—XGBoost, LightGBM, and CatBoost—to provide a more detailed understanding of how different boosting strategies perform in modelling complex socioeconomic data. This narrower focus enables a more precise evaluation of algorithmic behavior within the same learning paradigm, which remains limited in the existing literature.

Beyond predictive performance, interpretability is crucial in this context because decision makers require transparent information regarding the classification results to design effective intervention programs [21-22]. Explainable Artificial Intelligence (XAI) methods address this need by clarifying how models generate predictions [23]. One widely used is SHAP (SHapley Additive Explanations), which quantifies the contribution of each predictor to a model's output based on cooperative game

theory [24]. Unlike conventional feature importance, SHAP indicates whether each predictor has a positive or negative association with model output [25]. This approach enhancing the practical usability of complex gradient boosting models in policy context.

This study is among the first to explicitly compare gradient boosting algorithms within a unified framework using high-frequency panel data for food resilience classification, addressing the limited evidence on boosting-based approaches in longitudinal socio-economic analysis. The key research problem addressed in this study is to determine which gradient boosting algorithm provides the most accurate and reliable classification of household food resilience, while maintaining interpretability for policy-relevant insights. Thus, this study contributes to the literature by providing a comprehensive framework that combines predictive modelling, feature engineering, and explainable machine learning to better capture the dynamic nature of household food resilience.

2. Research Method

2.1. Materials

This study adopts a quantitative approach using a machine learning workflow to classify household food resilience. The overall research framework consists of data preprocessing, model development, and explainable analysis utilizing SHAP to the best-performing model. Gradient boosting was selected in developing the models due to their ability to handle complex data efficiently, and capture non-linear relationships and interaction among predictors, which are commonly present in socio-economic data [26]. The workflow of the study is illustrated in Figure 1.

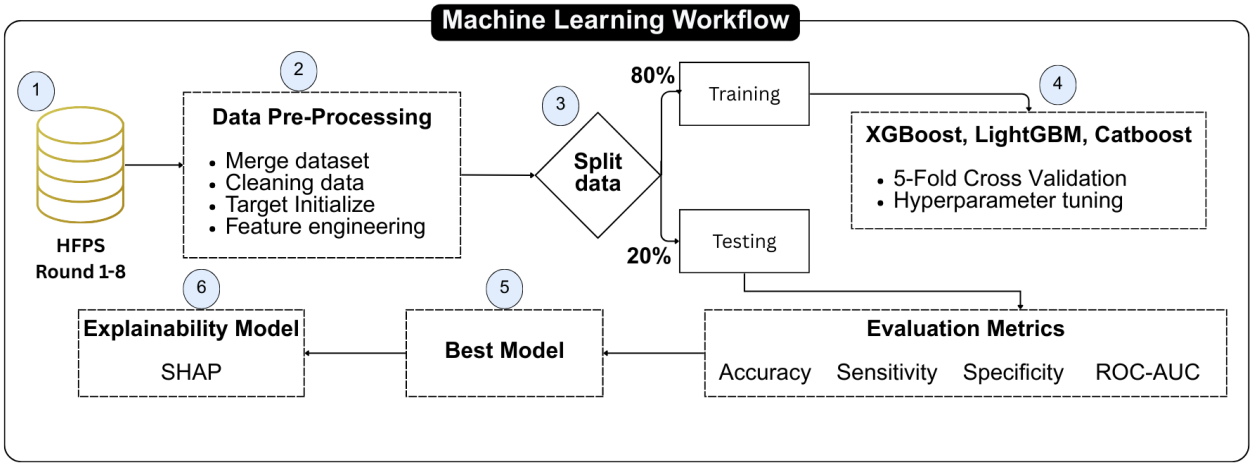


Figure 1. Research Workflow

2.2. Data Source

The dataset was obtained from the High Frequency Phone Survey (HFPS) conducted by World Bank in Indonesia during the COVID-19 pandemic [27]. The survey consists of eight rounds and includes eleven thematic modules covering household demographic, employment and income, food security, access to health service, education, concern and subjective welfare, knowledge about COVID-19, and social safety nets. The baseline survey interviewed 4,371 households which were subsequently re-contacted in the following rounds to monitor socioeconomic conditions during the pandemic.

Several survey questions were repeated across rounds, allowing the analysis of household dynamics over time. These repeated observations were later transformed into summary features during the feature engineering stage. In addition to structured questionnaire responses, the survey also includes open-ended questions describing household perception. These responses were recorded in textual format and were processed as text variables in the feature engineering stage.

2.3. Data Acquisition and Data Cleaning

The initial step of data preprocessing involved merging datasets from the eight survey rounds using a unique household identifier. Since some households did not participate in all rounds, observations with incomplete records were excluded from the analysis to ensure data consistency. Restricting the sample to a balanced panel inevitably raises concerns about potential selection bias due to household attrition [28]. To address this, an attrition check using baseline household characteristics with logistic regression model show limited predictive performance (accuracy ≈ 53.89, AUC ≈ 55.62%), indicating that dropout is not systematically driven by observable factors [29]. This provides reassurance that the balanced panel used in this study is unlikely to introduce substantial bias, thereby supporting the validity of the predictive framework.

Following sample acquisition, data validation was performed to detect anomalies and inconsistent entries. Certain placeholder values, such as 999.99, were used in the survey to indicate "do not know" responses. These were treated as missing and imputed using the median of non-missing observations within each feature. Median imputation was chosen because it is robust to outliers and preserves the sample size [30]. The final dataset thus comprised 2,868 households, complete and consistent for predictive modeling.

2.4. Response Variable

In supervised classification tasks, defining the response variable is essential. The response variable in this study is household food resilience, which is categorized into two classes: resilient (1) and non-resilient (0). The classification was derived from the trajectory of household food security status measured using the Food Insecurity Experience Scale (FIES) across eight survey rounds. Households are considered food secure when they do not experience food shortages and do not reduce their food consumption. An illustration of the construction of the food resilience status is provided in Figure 2.

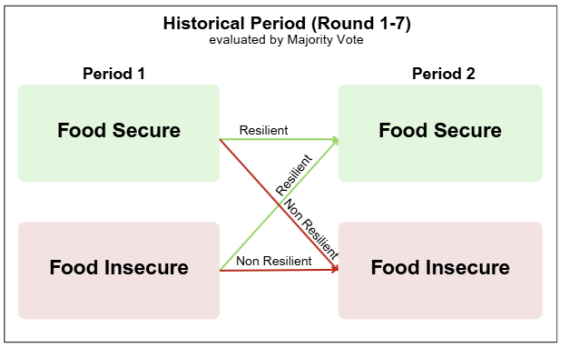


Figure 2. Resilience Concept

Following the resilience framework proposed by Villacís et al. [10], households are classified as resilient if they are able to maintain food security or recover from food insecurity to a food-secure condition. While the original framework was designed for two periods, this study extends the concept to eight survey rounds, thereby generating seven consecutive transitions that capture household dynamics throughout the pandemic.

To operationalize resilience in this multi-round setting, a majority voting rule was applied across the transitions. Households were classified as resilient (coded 1) if food security was the dominant state across transitions and they remained food secure in the final observation (Round 8). This criterion reflects the household's ability to sustain or regain food security after pandemic. Majority voting across transitions is consistent with ensemble decision theory, ensuring that classifications reflect dominant rather than temporary states. Conversely, households were classified as non-resilient (coded 0) if food insecurity dominated across transitions or if they were food insecure in the final observation, regardless of earlier conditions. The detailed classification criteria are summarized in Table 1.

Table 1. The Description of Response Variable

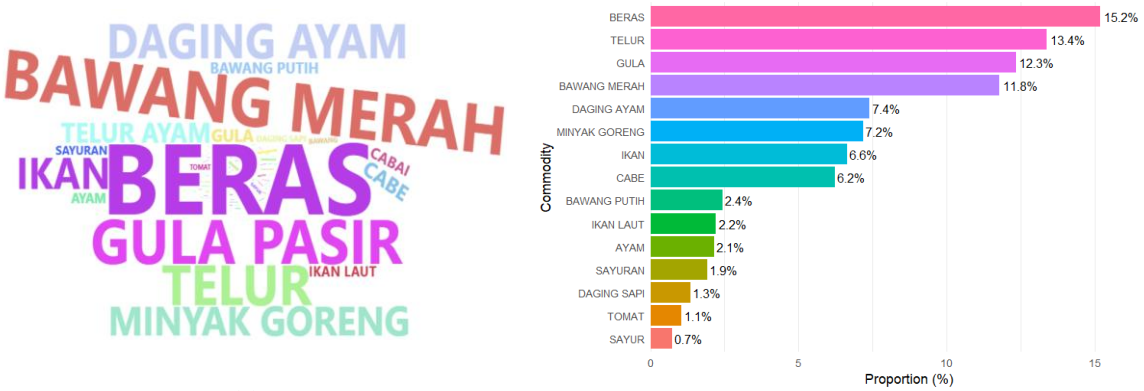
Label	Description
1 = Resilient	Households that maintained food security as the dominant state across transitions and remained food secure in the final observation (Round 8)
0 = Non resilient	Households that were food insecure in the final period (Round 8), or households with food insecurity dominating transitions regardless of their final condition

2.5. Feature Engineering

Feature engineering was conducted to transform the raw survey data into informative predictor variables suitable for machine learning algorithms [31]. Since the HFPS dataset consists of repeated observations across multiple survey rounds, the raw data inherently contains temporal dynamics that cannot be directly utilized by standard classification algorithms. To address this, feature engineering was designed to represent household dynamics over time by summarizing repeated observations into fixed-length predictors. This transformation enables the model to capture not only the level of household conditions but also their stability, variability, and exposure to shocks throughout the pandemic period. The selection of predictor variables was guided by both domain knowledge and data-driven considerations. The process focused on two primary areas: the conversion of unstructured text into numerical indicators and the aggregation of temporal data into stationary summary features.

2.5.1 Preprocessing Text Variable

One of the survey questions asked household to mention three commodities that they perceived as experiencing price increases during the pandemic. This question was designed as an open-ended response, allowing respondents to freely report commodities. Consequently, the responses were recorded in textual form and contained variations in spelling and wording. To make these responses consistent to machine learning model, the textual variables were transformed into numerical representations [32]. The preprocessing steps include case folding and standardization (for example, "cabe," "cabai merah," and "lombok" were standardized into "cabai"). Figure 3a presents the word cloud before cleaning, while Figure 3b displays the fifteen most frequently reported commodities experiencing price increases.



(a) (b)

Figure 3. Preprocessing Text

The bar plot in Figure 3b illustrates the distribution of various commodities that households reported as experiencing price increases. To ensure model parsimony while retaining explanatory relevance, only the most influential commodities were selected as predictors. The four most frequently reported

items—rice (beras), eggs (telur), sugar (gula), and shallots (bawang merah)—were included due to their central role in household consumption patterns. In addition, chili (cabai) was incorporated as a predictor despite not ranking among the top five commodities. This inclusion is theoretically justified by its historically high price volatility during the COVID-19 period, making it a sensitive indicator of food-related economic pressure [33].

2.5.2 Summary Features

The aggregation of repeated variables was performed using summary statistics such as mean, mode, count, and coefficient of variation. These statistics were selected to capture complementary aspects of household dynamics [34]. Mean values represent the average condition of households, reflecting their general socioeconomic status during the observation period, coefficient of variation (CV) captures instability and fluctuations. Mode represents the dominant state experienced by households, aligning with the concept of resilience as a sustained condition rather than a temporary state. Meanwhile, count capture the frequency of adverse events or disruptions, reflecting the intensity of shocks experienced during the pandemic.

This transformation aims to simplify the complex information, reduce redundancy and multicollinearity, also improve the ability of machine learning model to detect meaningful patterns associated with household food resilience. The detailed summary feature construction is presented in Table 2.

Table 2. Summary Features Construction from Repeated Survey Variables

Feature engineering	Statistic	Rationale
Reduced Work hours	Mode	Captures the dominant state of reduced working hours relative to pre-pandemic conditions
Work hour variation	CV	Measures variability in household's working hour across survey round
Work frequency	Count	Total instances of work participation or income-generating activities across round
Job change frequency	Count	Frequency of job transitions, reflecting labor market instability
Average monthly income	Mean	Represents average household income during the pandemic
Income variation	CV	Quantifies income fluctuation, indicating financial instability
Income drop	Mode	Binary indicator of whether income most frequently declined relative to pre-pandemic
Quarantine experienced	Count	Number of rounds where household head was unable to work due to quarantine restrictions
Job disruption frequency	Count	Frequency of irregular work, unemployment, or job changes due to COVID-19
Unmet need for medication	Count	Number of rounds where the household reported barriers to accessing essential medicines
COVID-19 concern	Mean	Average Likert score of household anxiety regarding the pandemic
Financial concern	Mean	Average Likert score of financial stress experienced by the household
Debt ownership	Mode	Binary indicator showing whether borrowing was the dominant coping mechanism across rounds
Frequency of social safety nets	Count	Number of social assistance programs received across rounds

2.5.3 List of Predictor Variables

Following the preprocessing and feature engineering procedures described in the previous section, the variables generated from the summary feature construction (Table 2), together with other relevant household characteristic obtained from the HFPS dataset were compiled into the final predictor set. After combining, a total of 42 predictor variables were applied for model development. These variables

were used as inputs for the classification models to predict household food resilience. To facilitate the interpretation, the predictors were grouped into five categories as presented in Table 3 with the complete list of predictor variables.

Table 3. Predictor Variables

Predictor Group	Code	Variables
Demographic Carctheristics	$X_1 - X_8$	Residential location; House ownership status; Household size; Gender of Household head; Age of household head; Education of household head; Dependency ratio (%); Death status of household head
Employement and Income Loss	$X_9 - X_{25}$	Employment type; multiple worker; business ownership; online business; open business during COVID-19; ability to pay electricity; digital financial transactions; reduced work hours; work hour variation; work frequency; job change frequency; average monthly income; income variation; income drop; quarantine experience; job disruption frequency
Health and Food Shocks	$X_{26} - X_{35}$	Ability to buy medicine; COVID-19 symptoms; unmet needs medication; rice price shock; sugar price shock; shallot price shock; egg price shock; chili price shock; cooking oil scarcity; pre-pandemic vulnerability status.
Subjective Welfare	$X_{36} - X_{38}$	Covid-19 concern, financial concern, debt ownership
Social safety nets	$X_{39} - X_{42}$	Receiving remittance, BPJS health insurance, BPJS Employment insurance, frequency of social safety nets

2.6 eXtreme Gradient Boosting

Gradient boosting is an ensemble technique that build predictive models by sequentially combining multiple decision trees [35]. Each new tree is trained to correct the prediction errors made by previous trees, thereby improving overall model accuracy. eXtreme Gradient Boosting (XGBoost) is an advanced implementation of the gradient boosting framework. In this process, each successive tree is designed to minimize the residual errors of the previous ones by optimizing a specific loss function. A distinguishing feature of XGBoost is the inclusion of explicit regularization terms in its objective function, which penalizes model complexity to effectively prevent overfitting and improve generalization performance on unseen data [36].

2.7 LightGBM

LightGBM is a gradient boosting framework designed for high efficiency and scalability. Unlike traditional boosting models that grow trees level-wise, LightGBM employs a leaf-wise growth strategy, which splits the leaf with the highest loss reduction, resulting in deeper, asymmetrical trees [37]. To optimize computational speed, it utilizes two key techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS reduces the data size by retaining observations with larger gradients (those with higher errors) while sampling those with smaller gradients, ensuring the model focuses on the most informative data points. Meanwhile, EFB reduces feature dimensionality by bundling mutually exclusive sparse features together without losing critical information.

2.8 CatBoost

CatBoost (Categorical Boosting) is a specialized gradient boosting algorithm that excels in handling datasets with numerous categorical features [38]. It utilizes a unique technique called ordered boosting to overcome the gradient estimation bias (target leakage) that often occurs in traditional boosting methods [39]. This approach enhances the robustness of the training process and allows the model to process categorical variables directly through efficient encoding schemes, making it highly effective for complex socio-economic datasets.

2.9 Model Evaluation

Model performance was evaluated using several classification metrics derived from the confusion matrix, including accuracy, sensitivity, and specificity [40]. In addition, the Area Under Curve (AUC) was used as the primary evaluation metric because it summarizes model performance across all possible classification thresholds [41]. The AUC values, which indicate how well the model identifies, are presented in Table 4.

Table 4. AUC Values Category

AUC Value	Category
0.9 – 1.00	Excellent Classification
0.80 – 0.89	Good Classification
0.70 – 0.79	Fair Classification
0.60 – 0.69	Poor Classification
0.50 – 0.59	Failure

Furthermore, to statistically assess the comparative performance among gradient boosting models, the Friedman test was employed. As a non-parametric test, it is particularly suitable for comparing multiple algorithms under repeated iterations [42].When the Friedman Test indicated significant differences, the Wilcoxon signed-rank test was conducted as a post-hoc analysis to identify which pairs of models differed significantly [43].

2.10 Shapley Additive Explanations (SHAP)

SHAP is a model-agnostic interpretability framework grounded in cooperative game theory. It quantifies the contribution of each feature to a specific classification prediction by calculating its Shapley value, which represents the average marginal contribution of a feature across all possible combinations of predictors [44]. By attributing the difference between the model's actual prediction and the average prediction to individual features, SHAP provides both local explanations for individual households and global insights into the overall drivers of model prediction [45]. The SHAP value for a specific feature j is calculated using the weighted average of all possible marginal contribution across all feature coallitions. It is expressed as follows:

$$\Phi_j = \sum_{S \subseteq M_{(j)}} \frac{|S|!\,(M - |S| - 1)!}{M!} \big(v(S \cup \{j\}) - v(S)\big), j = 1, \dots, M$$

Where M is the total numeber of predictor, S present a subset (coalition) of feature that exclude feature j, $v(S)$ denotes the model's prediction, and the term $[v(S \cup \{j\}) - v(S)]$ represents the marginal contribution of feature j to the prediction. Since this study employs tree-based gradient boosting models, TreeSHAP was implemented to compute Shapley values efficiently by leveraging the hierarchical structure of decision trees.

3. Results and Discussion

3.1 Descriptive Analysis

The initial stage of the analysis is the exploration of the distribution of household food resilience to obtain an overview of patterns within the HFPS sample. Presenting this distribution is essential for identifying the proportions of resilient and non-resilient households, thereby providing necessary context for subsequent analysis of the determinants that may influencing resilience. The proportion of the resilience status is shown in Figure 4.

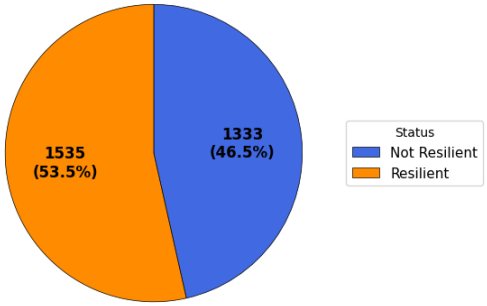


Figure 4. Proportion of Household Food Resilience Status

As shown in Figure 4, the distribution of food resilience among the 2,868 households indicates that 53.5% (1,535 household) were classified as resilient, while 46.5% (1,333 household) were classified as non-resilient. Because the class proportions are relatively balanced, specific handling for imbalanced data was not required for this classification case.

3.2 Classification Models

The first step for building the model is splitting the dataset into training 80% and testing 20%. Building model is process in the training set. After that, initialize hyperparameter among the algorithms. Each algorithm has its own set of hyperparameters, which can influence the model's learning process and the quality of the production it produces [46].Model training and hyperparameter optimization were performed exclusively on the training set, while the testing set was reserved for final performance evaluation [47]. To ensure fair comparison, hyperparameter search spaces were designed to be comparable across models. The explored parameter ranges are presented in Table 5.

Table 5. Hyperprameter Range

Hyperparameter	XGBoost	LightGBM	CatBoost
Number of trees (n_estimator/iterations)	200-600	200-600	200-600
Maximum tree depth (max_depth)	3-8	-	3-8
Learning rate (eta)	0.03-0.15	0.03-0.15	0.03-0.15
Proportion of sample each tree (subsample)	0.7-1.0	0.7-1.0	-
Proportion of features each tree (colsample_bytree)	0.7-1.0	0.7-1.0	-
Number of leaves (num_leaves)	-	20-100	-
Minimum sample per leaf (min_data_in_leaf)	-	20-50	-

The tuning process was performed using Optuna, a Bayesian optimization framework with a Tree-structured Parzen Estimator (TPE) approach. This optimization process employed 5-fold cross-validation scheme, with the Area Under the Curve (AUC) as the objective metric to identify the optimal parameter combinations. Based on tuning results, the XGBoost model achieved its best configuration with 490 trees, a maximum depth of 3, a subsample ratio of 0.93, and proportion feature of 0.78. LightGBM attained its optimal performance with 492 trees, 69 leaves and minimum of 34 samples per leaf. Meanwhile, CatBoost reached its best setting with 586 iterations (trees), a tree depth of 5, and a learning rate of 0.023.

This configuration emphasizes that optimal performance of a gradient boosting algorithm depends heavily on the distinct learning strategic. In XGBoost, the moderate number of trees with low depth, the use of subsamples and feature proportions, suggest a stochastic gradient boosting strategy

to maintain generalization and reduce overfitting [35]. LightGBM, with its large number of leaves, emphasizes the flexibility of a more complex tree structure, consistent with its characteristics as the fastest algorithm [48]. Meanwhile, CatBoost, with its low learning rate and moderate depth, exhibits a conservative strategy that emphasizes prediction stability. In conclusion, hyperparameter selection is not merely technical, but rather a strategy to balance accuracy, speed, and generalization ability, where each algorithm has unique characteristics that are reflected in its optimal configuration [49].

After the best configurations were obtained, the models were initially evaluated using a confusion matrix to visualize their predictive performance which provides a detailed summary of the classification results by comparing predicted labels with the actual class labels. The confusion matrix obtained from the testing dataset (20%) is presented in Figure 5.

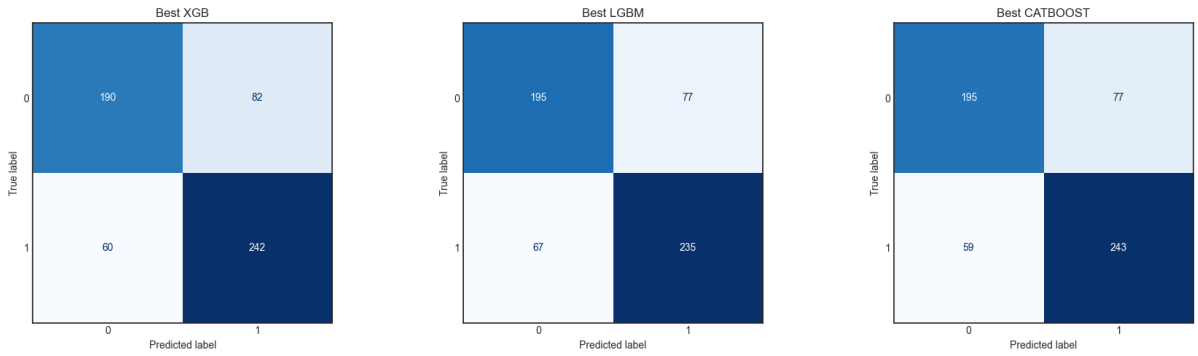


Figure 5. Confusion Matrix of XGBoost, LightGBM, and CatBoost

Figure 5 presents the confusion matrix of XGBoost, LightGBM and CatBoost evaluated on the fixed testing set using their optimal hyperparameters. The comparative evaluation highlights that all three algorithms reveal broadly similar predictive patterns, where the diagonal cells representing correct classifiations appear darker than those indicating misclassifications. This consistency may be attributed to the shared boosting framework, which iteratively build ensemble of weak learners to minimize classification errors. Despite this identical foundation, the models exhibit nuanced differences in their error distribution.

In terms of true positive (TP), both XGBoost (242) and CatBoost (243) demonstrate stronger ability to correctly identify positive class (resilient households), compared to LightGBM (235). This suggest that XGBoost and CatBoost are more effective in capturing resilience signals within the data. The false negative (FN) reveal a critical distinction, LightGBM record the highest FN (67), meaning it misses more resilient households by misclassifying them as non-resilient. CatBoost, with the lowest FN (59), minimizes this risk, thereby offering a more reliable detection of resilience. Regarding true negatives (TN), LightGBM and CatBoost (both 195) slightly outperform XGBoost (190). This indicates that LightGBM and CatBoost are marginally better at correctly identifying non-resilient households. In terms of false positive (FP), XGBoost produces the highest number (82), reflecting a tendency to overestimate resilience by misclassifying non-resilient households as resilient. Meanwhile, LightGBM and CatBoost both record fewer FP (77), showing a more conservative stance in resilience classification.

To ensure the stability of the result, the evaluation was extended by repeating the train-test split procedure across ten different random seeds. This repetition allowed for the assessment of model stability under varying data partitions, thereby reducing the risk of performance bias due to a single split. The performance of the model was evaluated using Accuracy, Sensitivity, Specificity, and AUC. In addition, statistical significance was assessed using the Friedman test. The average performance across the ten repetitions, along with the Friedman significance test results, is summarized in Table 6, providing a more comprehensive view of the comparative strengths and weaknesses of each algorithm

Table 6. Model Performance and Friedman Significance Test (Mean ± SD)

Metric	XGBoost (mean ± sd)	LightGBM (mean ± sd)	CatBoost (mean ± sd)	Friedman (p-value)
Accuracy	0.7531 ± 0.0023	0.7491 ± 0.0002	$\mathbf{0.7592 \pm 0.0033}$	1.83×10^{-4}
Sensitivity	$\mathbf{0.8007 \pm 0.0034}$	0.7781 ± 0.0052	0.7993 ± 0.0047	4.44×10^{-4}
Specificity	0.7004 ± 0.0036	$\mathbf{0.7169 \pm 0.0002}$	0.7147 ± 0.0053	5.60×10^{-4}
AUC	0.8304 ± 0.0007	0.8142 ± 0.0002	$\mathbf{0.8331 \pm 0.0007}$	4.54×10^{-5}

As shown in Table 6, CatBoost consistently achieves the highest average performance in terms of Accuracy (0.7592 ± 0.0033) and AUC (0.8331 ± 0.0007), indicating that approximately 75.9% of the observations were correctly classified and that the model demonstrates superior discriminative ability. XGBoost remains highly competitive in Sensitivity (0.8007 ± 0.0034), suggesting its strength in identifying resilient households. In contrast, LightGBM achieves the highest Specificity (0.7169 ± 0.0002), indicating a stronger ability to correctly classify non-resilient households. The Friedman test confirmed that these differences are statistically significant across all metrics ($p < 0.001$), necessitating further pairwise investigation. To identify specific differences between model pairs, a post-hoc analysis was performed using the Wilcoxon signed-rank test with Bonferroni correction, as presented in Table 7.

Table 7. Pairwise Comparison Results

Pairwise	Accuracy	Sensitivity	Specificity	AUC
XGBoost vs LightGBM	0.025^{*}	0.017^{*}	0.016^{*}	0.006^{*}
LightGBM vs CatBoost	0.017^{*}	0.017^{*}	0.666	0.006^{*}
XGBoost vs CatBoost	0.037^{*}	1.000	0.024^{*}	0.006^{*}

*indicates statistical significance at the 5% level

The results reveal that CatBoost significantly outperforms LightGBM across most evaluation metrics ($p < 0.05$), except for Specificity ($p = 0.666$), where the difference is not statistically significant. This suggests that although LightGBM achieves slightly higher specificity than CatBoost, the difference is not strong enough to be considered statistically meaningful. Furthermore, when compared to XGBoost, CatBoost demonstrates significantly better performance in Accuracy, Specificity and AUC ($p < 0.05$), while no significant difference is observed in Sensitivity ($p > 0.05$). This indicates that the higher sensitivity observed in XGBoost is not statistically significant. These results highlight that CatBoost provides as the most consistent predictive performance across multiple evaluation criteria.

CatBoost's superior performance is attributed by its algorithmic design, which is particularly well-suited for heterogeneous and noisy socio-economic survey data. CatBoost incorporates an ordered boosting scheme that reduce prediction sift and mitigates overfitting, leading to better generalization [50]. In addition, its ability to process categorical variables efficiently allows the model to better utilize information [51]. The relatively low variance observed across repeated experiments further indicates that CatBoost is more robust to sampling variability, an important property when working with survey data characterized by heterogeneity and measurement noise [35].

Recent interdisciplinary studies have reported similar finding, where gradient boosting consistently outperforms other machine learning algorithms in classification tasks [51]. In socioeconomic or food security context, the obtained AUC value over 0.80 can be categorized as good classification performance [52]. Our results fall within this benchmark, reinforcing the stability of gradient boosting for resilience classification. Nevertheless, prior studies in food security often identify Random Forest as the best-performing model [18-20], suggesting that model superiority is context-

dependent, influenced by data characteristics, feature representation, and problem formulation. Within this high-frequency dataset, CatBoost's design appears to provide a distinct advantage.

In summary, the results indicate that while all gradient boosting algorithms provide reliable performance, CatBoost offers the most robust and balanced predictive capability. This makes it particularly suitable for applications where both accuracy and stability are critical. In the context of household food resilience, this reliability is essential for policy targeting and vulnerability assessment, where misclassification could lead to inefficient resource allocation.

3.3 Feature Importance Analysis

Following the identification of CatBoost as the best model, the next step focuses on interpreting the model's predictions to uncover the underlying drivers of household food resilience. This analysis is conducted using SHAP (SHapley Additive Explanations), which provides a theoretically grounded framework to decompose model predictions into individual feature contributions. The explainability are presented in Figure 6, which displays the top fifteen predictors with highest important.

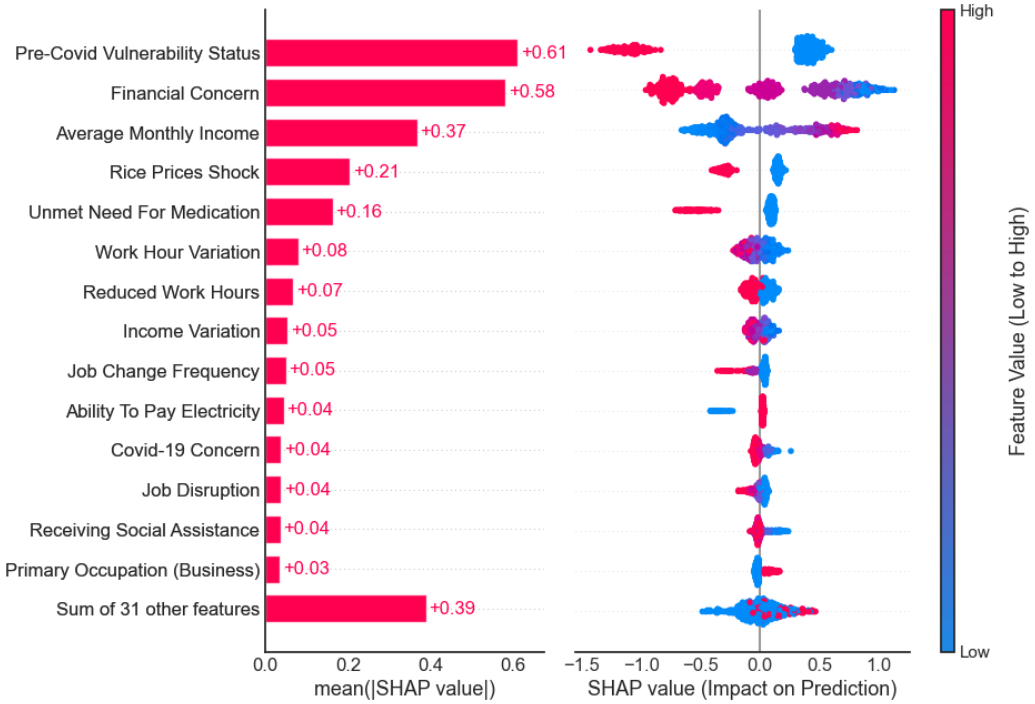


Figure 6. SHAP Feature Importance of The CatBoost Model

Figure 6 presents a dual perspective on feature importance. The left panel (barplot) illustrates the global importance of features, calculated as mean absolute SHAP value across all observations, thereby identifying the variables with the strongest overall influence on the model. The right panel (beeswarm plot) provides a more detailed visualization of the direction of feature contributions. Each point represents an individual observation, with red points indicating high feature values, and blue representing low feature values.

The barplot results highlight three broad dimensions-structural factors, economic capacity, and exposure to shock-that emerge as important factors. The beeswarm plot further illustrates how these predictors affect classification outcomes. Households with higher levels of baseline vulnerability tend to be associated with lower predicted probabilities of resilience. Similarly, higher levels of financial concern are generally linked with lower resilience predictions. In contrast, higher household income levels tend to increase the predicted probability of resilience. Indicators related to economic shocks

and access to resources, such as rice price shocks and unmet medical needs, also show noticeable contributions to the model output.

Other variable within the top-ranked predictors are derived from the feature engineering process, particularly those capturing variability, frequency, and persistence of conditions across survey rounds. Variables such as work hour variation, income instability, and frequency of job change consistently appear among the top contributors. This finding provides empirical evidence of the importance of feature engineering in this study. By transforming repeated survey observations into summary statistics, the modelling framework is able to capture temporal dynamics that would otherwise be lost in cross-sectional representations. The dominance of these engineered features in the SHAP ranking suggests that incorporating repeated measures enhances the model's ability to detect meaningful patterns associated with resilience.

Furthermore, the SHAP results demonstrate that the model captures complex, nonlinear relationships between predictors and the response variable. For instance, the impact of average income and financial concern is not strictly linear, as their effects vary across different ranges of values. This reinforces the suitability of gradient boosting models for analyzing socio-economic data, where interactions and nonlinearities are inherent.

Overall, the SHAP analysis not only improves the interpretability of the classification model but also provides substantive insights into the mechanisms underlying household food resilience. The results highlight that resilience is shaped by a combination of structural factors (baseline vulnerability), economic capacity (income and financial stress), and exposure to shocks (price and health-related disruptions). Importantly, the findings also underscore that capturing temporal dynamics through feature engineering is essential for accurately representing resilience in longitudinal survey data.

4. Conclusion

This study demonstrates that gradient boosting algorithms are effective for classifying household food resilience using high-frequency survey data. Among the evaluated models, CatBoost consistently achieves the best and most stable performance across multiple evaluation metrics, indicating its suitability for capturing complex patterns in socio-economic data. The integration of SHAP further extends the contribution of this study by providing interpretable insights into the drivers of household food resilience. The analysis reveals that baseline vulnerability, financial stress, income capacity, and exposure to economic and health-related shocks are the most influential factors. These results reinforce the conceptual understanding of resilience as a function of both structural conditions and adaptive capacity, while also demonstrating how explainable machine learning can bridge predictive accuracy with substantive interpretation.

Future research may extend this framework by incorporating alternative explainability methods, such as LIME or permutation-based approaches, as well as exploring hybrid modelling strategies that integrate temporal modelling techniques to further enhance the representation of dynamic household behavior.

References

[1] KC, U., Campbell-Ross, H., Godde, C., Friedman, R., Lim-Camacho, L., & Crimp, S. (2024). A systematic review of the evolution of food system resilience assessment. *Global Food Security*, 40, 100744.

[2] Yenerall, J., & Jensen, K. (2022). Food security, financial resources, and mental health: Evidence during the COVID-19 pandemic. *Nutrients*, 14(1), 161.

[3] Wambede, N. M., Sharon, A., JoyFred, A., Mulabbi, A., & Robert, T. (2025). Adaptations to climate variability in Northern Uganda: Implications for food security. *Journal of Tropical Crop Science*, 12(2), 261–273.

[4] Rahardiantoro, S., & Sakamoto, W. (2021). Clustering regions based on socio-economic factors which affected the number of COVID-19 cases in Java Island. *Journal of Physics: Conference Series*, 1863(1), 012014.
[5] Satria, A., Anggraini, E., Widyastutik, Nuryartono, N., Amaliah, S., Helmi, A., et al. (2022). Membangun resiliensi sistem pangan Indonesia. *Policy Brief Pertanian, Kelautan Dan Biosains Tropika*, 4(4), 338–345.
[6] Oktarina, S. D., Nurkhoiry, R., Amalia, R., Pradikto, I., & Rahutomo, S. (2022). Dampak ketidakpastian COVID-19, iklim, dan kompleksitas lainnya pada industri kelapa sawit. *Warta PPKS*, 27(2), 70–77.
[7] BPS-Statistics Indonesia. (2025). *Prevalence of undernourishment*. Jakarta: BPS.
[8] Akbar, A., Darma, R., Fahmid, I. M., & Irawan, A. (2023). Determinants of household food security during the COVID-19 pandemic in Indonesia. *Sustainability*, 15(5), 4131.
[9] Hassan, R., Di Martino, M., & Daher, B. (2025). Quantifying sustainability and resilience in food systems: A systematic analysis for evaluating the convergence of current methodologies and metrics. *Frontiers in Sustainable Food Systems*, 9, 1479691.
[10] Villacis, A. H., Badruddoza, S., & Mishra, A. K. (2024). A machine learning-based exploration of resilience and food security. *Applied Economic Perspectives and Policy*, 46(4), 1479–1505.
[11] Ronalia, P., Hartono, D., & Misdawita, M. (2023). The impact of resilience on household food insecurity in Indonesia. *Jurnal Ekonomi Pembangunan*, 21(1), 25–38.
[12] Machefer, M., Thomas, A. C., Meroni, M., Veiga Lopez Pena, J. M., Ronco, M., Corbane, C., et al. (2025). Potential and limitations of machine learning modeling for forecasting acute food insecurity. *Global Food Security*, 45, 100859.
[13] Aguilera, T., & Jatmiko, Y. A. (2023). Pemodelan tingkat kerawanan pangan rumah tangga di Indonesia tahun 2021 dengan pendekatan regresi logistik ordinal. *Indonesian Journal of Applied Statistics*, 5(2), 90.
[14] Rahardiantoro, S., Juhanda, A. R. N., Kurnia, A., Aswi, A., Sartono, B., Handayani, D., et al. (2024). Spatio-temporal modeling to identify factors associated with stunting in Indonesia using a modified generalized Lasso. *Spatial and Spatio-temporal Epidemiology*, 51, 100694.
[15] Hartono, A., Dewi, L. A., Yuniarti, E., Putri, S. T. H., & Harahap, T. S. (2024). Machine learning classification for detecting heart disease with K-NN algorithm, decision tree and random forest. *Eksakta: Berkala Ilmiah Bidang MIPA*, 24(4), 513–522.
[16] Kyriazos, T., & Poga, M. (2024). Application of machine learning models in social sciences: Managing nonlinear relationships. *Encyclopedia*, 4(4), 1790–1805.
[17] Kern, C., Klausch, T., & Kreuter, F. (2019). *Tree-based machine learning methods for survey research*. https://doi.org/10.31235/osf.io/7w3sc .
[18] Dharmawan, H., Sartono, B., Kurnia, A., Hadi, A. F., & Ramadhani, E. (2022). A study of machine learning algorithms to measure the feature importance in class-imbalance data of food insecurity cases in Indonesia. *Communications in Mathematical Biology and Neuroscience*, 2022, 1-15.
[19] Khikmah, K. N., Sartono, B., Susetyo, B., & Dito, G. A. (2024). Performance comparative study of machine learning classification algorithms for food insecurity experience by households in West Java. *Jurnal Online Informatika*, 9(1), 128–137.
[20] Fransiska, H., Soleh, A. M., Notodiputro, K. A., & Erfiani. (2025). Evaluation of machine learning models based on household food insecurity data in Indonesia. *BIO Web of Conferences*, 171, 02011.
[21] Abd El-Ghani, S. S., Mansour, T. G. I., & Esleem, S. A. (2025). The most important economic and social indicators of the challenges facing food security for the most important crops in Egypt. *Environmental and Sustainability Indicators*, 27, 100808.

[22] Yuliani, E., Sartono, B., Wijayanto, H., Hadi, A. F., & Ramadhani, E. (2022). Study of features importance level identification of machine learning classification model in sub-populations for food insecurity. *AIP Conference Proceedings*, 2668(1), 050012.

[23] Borys, K., Schmitt, Y. A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C. M., et al. (2023). Explainable AI in medical imaging: An overview for clinical practitioners - Saliency-based XAI approaches. *European Journal of Radiology*, 162, 110787.

[24] Oktora, S. I., Matualage, D., Notodiputro, K. A., & Sartono, B. (2025). Data-driven insights into underdeveloped regencies: SHAP-based explainable artificial intelligence approach. *International Journal of Artificial Intelligence Research*, 9(1), 1-12

[25] Saranya, A., & Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7, 100230.

[26] Rane, N. L., Mallick, S. K., & Rane, J. (2025). *Artificial intelligence and machine learning for enhancing resilience: Concepts, applications, and future directions*. Deep Science Publishing.

[27] World Bank. (2023). *Indonesia: High-frequency monitoring of COVID-19 impacts rounds 1-8, 2020-2023 (HIFY 2020-2023)*. World Bank Open Data. https://doi.org/10.48529/VC74-5V86

[28] Letta, M., Montalbano, P., Morales Opazo, C., & Petruccelli, F. (2025). Measuring and testing vulnerability to food insecurity for prediction and targeting. *Economics & Human Biology*, 59, 101552.

[29] Jacobsen, E., Ran, X., Liu, A., Chang, C. C. H., & Ganguli, M. (2021). Predictors of attrition in a longitudinal population-based study of aging. *International Psychogeriatrics*, 33(8), 767–778.

[30] Joel, L.O., Doorsamy, W. & Paul, B.S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *Int J Data Sci Anal* , 20, 6357–6373.

[31] Katya, E. (2023). Exploring feature engineering strategies for improving predictive models in data science. *Research Journal of Computer Systems and Engineering*, 4(2), 201–215.

[32] Safrizal, A. N., & Agustian, S. (2024). Classification of COVID-19 vaccine sentiment using K-Nearest Neighbor and Fasttext on Twitter. *Eksakta: Berkala Ilmiah Bidang MIPA*, 25(3), 362–371.

[33] Rahman, I. A., Siregar, P. S. G., & Suharsih, S. (2024). Analysis of the potential and volatility of large chili and cayenne chili after COVID-19 in Yogyakarta City. *Journal of International Conference Proceedings*, 6(6), 497–508.

[34] Cascarano, A., Mur-Petit, J., Hernández-González, J., Camacho, M., de Toro Eadie, N., Gkontra, P., et al. (2023). Machine and deep learning for longitudinal biomedical data: A review of methods and applications. *Artificial Intelligence Review*, 56(2), 1711–1771.

[35] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967.

[36] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[37] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.

[38] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 6638–6648.

[39] Syamkalla M., Khomsah S., Nur Y., (2024). Implementasi Algoritma Catboost Dan Shapley Additive Explanations (SHAP) Dalam Memprediksi Popularitas Game Indie Pada Platform Steam. *Jurnal Teknologi Informasi dan Ilmu Komputer,* 11(4), 777-786.

[40] Putri, N., Parmikanti, K., & Gusriani, N. (2024). Robust linear discriminant analysis with modified one-step M-estimator Qn scale for classifying financial distress in banks: Case study. *Eksakta: Berkala Ilmiah Bidang MIPA*, 25(2), 219–230.
[41] Ratnaningayu, N. D., Tedjo, A., & Panigoro, S. S. (2024). The implementation of machine learning algorithms for breast cancer biomarker validation in metabolomics studies. *Eksakta: Berkala Ilmiah Bidang MIPA*, 25(4), 468–483.
[42] Yunita, A., Pratama, M. I., Almuzakki, M. Z., Ramadhan, H., Akhir, E. A. P., Mansur, A. B. F., et al. (2025). Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models. *MethodsX*, 15, 103462.
[43] Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1), 6086.
[44] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56–67.
[45] Chowdhury, T. R., & Dey, S. (2024). A comparative survey of SHAP and LIME: Explaining machine learning models for transparent AI. *International Journal of Innovative Research in Technology*, 11(1), 827–835.
[46] Gomes Mantovani, R., Horváth, T., Rossi, A. L. D., Cerri, R., Barbon Junior, S., Vanschoren, J., et al. (2024). Better trees: An empirical study on hyperparameter tuning of classification decision tree induction algorithms. *Data Mining and Knowledge Discovery*, 38(3), 1364–1416.
[47] Nafis, M. A. N., Susilaningrum, D., Ulama, B., & Kusrini, D. E. (2024). Assessing the impact of household socioeconomic factors on clean and healthy living behaviors with binary logistic regression: A study in Probolinggo Regency. *Eksakta: Berkala Ilmiah Bidang MIPA*, 25(4), 436–447.
[48] Tang, M., Peng, Z., & Wu, H. (2021). Fault detection for pitch system of wind turbine-driven doubly fed based on IHHO-LightGBM. *Applied Sciences*, 11(17), 8030.
[49] Ilemobayo, J. A., Durodola, O., Alade, O., Awotunde, O. J., Olanrewaju, A. T., Falana, O., et al. (2024). Hyperparameter tuning in machine learning: A comprehensive review. *Journal of Engineering Research and Reports*, 26(6), 388–395.
[50] Raparthi, M., Dhabliya, D., Kumari, T., Upadhyaya, R., & Sharma, A. (2024). Implementation and performance comparison of gradient boosting algorithms for tabular data classification. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 461–479).
[51] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, 7(1), 94.
[52] de Hond, A. A. H., Steyerberg, E. W., & van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, 4(12), e853–e855.