© Sungai Siput Community College, Ministry of Higher Education, Malaysia

JTH

https://jthkkss.com/
e-ISSN 2805-4431
DOI: https://doi.org/10.53797/jthkks.v6i1.3a.2025

Enhancing Nutritional Predictive Models: Addressing Class Imbalance with Machine Learning

Liliana Swastina[1], Bahbibi Rahmatullah[1*], Bambang Lareno[2], Asmara Alias[1], Achmad Hidayanto[3], M Khairudin[4]

[1]Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, MALAYSIA

[2]Calvin Institute of Technology, Jakarta, INDONESIA

[3]Faculty of Computer Science, University of Indonesia, West Java, Depok, INDONESIA

[4]Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, INDONESIA.

*Corresponding author email: bahbibi@meta.upsi.edu.my

Received 03 May 2025; Accepted 13 June 2025; Available online 30 June 2025

Abstract: Malnutrition remains a critical public health concern, particularly in low-resource settings where early detection is essential yet often constrained by limited infrastructure. While machine learning (ML) has emerged as a promising tool for nutritional risk prediction, many existing models fail to address class imbalance, resulting in biased outcomes and poor minority class detection. This study introduces an optimized ML framework that integrates imbalance-handling techniques—specifically SMOTE and Bagging—into the classification of stunting, wasting, and underweight among children in Banjarmasin, Indonesia. A curated dataset from 26 community health centers was used to train and evaluate five algorithms (Neural Network, Random Forest, Decision Tree, Logistic Regression, and XGBoost) across three treatment phases. Performance was assessed using 10-fold cross-validation and multi-method statistical validation, including ANOVA, Kruskal-Wallis, Dunn's, and Friedman tests. XGBoost consistently outperformed other models, achieving the highest accuracy (90.7%) and F1 scores across all indicators. The integration of oversampling and ensemble methods yielded substantial improvements in minority class detection, with F1 score gains ranging from 1.15% to 419.42%. Spatial validation revealed regional disparities, underscoring the need for adaptive modeling strategies. These findings contribute to the development of scalable, equitable, and context-aware nutritional surveillance systems, offering actionable insights for targeted interventions and public health policy.

Keywords: Nutritional Status, Predictive model, Class imbalance, Machine learning

1. Introduction

In Indonesia, according to a study by the Directorate of Community Nutrition Development of the Health Ministry, changes in the weight of children under five years old over time can be an early indication of changes in their nutritional status (Lareno et al., 2020). If a child's weight does not increase within six months, they are 12.6 times more likely to experience malnutrition than children whose weight continues to increase. The prevalence of stunting in Indonesia has been a significant issue, particularly in certain regions. For instance, South Kalimantan Province, with Banjarmasin as its capital, had a stunting prevalence of 34.2% in 2017, which was an increase of 1.1% from the previous year (Dinkes Kalsel, 2023). This placed the province among the ten highest stunting rates. However, in recent years, there has been a notable improvement. The national prevalence of stunting in Indonesia decreased from 29.6% in 2017 to 21.6% in 2022, and further to approximately 14.6% in 2024. In Banjarmasin, the prevalence of stunting fell to 17% in 2023 (Dinkes Banjarmasin, 2024).

Corresponding author: bahbibi@meta.upsi.edu.my

The Indonesian government has established the Posyandu (*Pos Pelayanan Terpadu,* or neighborhood integrated service post) under Puskesmas (*Pusat Kesehatan Masyarakat,* or community health facility). Posyandu provides services to monitor maternal, neonatal, and child health (MNCH) (Setjen Kementerian Kesehatan RI, 2021). Banjarmasin has 26 Puskesmas, each overseeing 8-12 Posyandu (Dinkes Banjarmasin, 2024). Efforts to reduce stunting require expanding access to health services and facilities, but the government's budget is limited. Thus, the budget needs to be optimized by prioritizing areas required more urgently. One approach is to utilize health data from Posyandu with machine learning (ML) to predict nutritional status and generate a comprehensive map of public health conditions in Banjarmasin. This data-driven approach can help target interventions more effectively and make the most of limited resources (Fazraningtyas et al., 2024; Sinambela et al., 2024; Swastina et al., 2024).

Many studies on ML algorithms have been conducted to assess children's nutrition. Table 1 highlights four commonly used algorithms in nutritional research: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), and Neural Network (NN). Among these, RF performed best in 8 out of 15 comparison papers. However, in another study, LR outperformed other algorithms, including RF (Ferdowsy et al., 2021). Similarly, DT was found to outperform NB in (Yuliansyah et al., 2020). It is essential to note that these four algorithms are not the only ones with high accuracy. XGB demonstrated superior performance compared to RF(Bitew, et al., 2022; Pang, et al., 2021) In two papers that included XGB alongside the other four algorithms, XGB outperformed them.

Table 1: Summary of comparative study research

Authors	Algorithm	Result
(Alqahtani et al., 2021)	**RF**, Multilayer Perception (MLP)	The RF method outperforms MLP in predicting obesity at an early stage with a high accuracy of 96.70%.
(Bansod et al., 2020)	**ID3**, Naïve Bayes (NB)	The accuracy for the ID3 is 85%, and Naïve Bayes is 57%.
(Bitew et al., 2022)	LR, NN, RF, k-NN, Extreme Gradient Boosting (**XGB**)	The XGB offers better predictive accuracy in stunting, wasting, and underweight, with 67.7%, 88.0%, and 75.7%, respectively.
(Fenta et al., 2021)	LR, NN, **RF**, Least Absolute Shrinkage and Selection Operator (L-1 regularization LR), L-2 regularization (Ridge), Elastic net	The RF algorithm was selected as the best ML model. In order of importance; urban-rural settlement, literacy rate of parents, and place of residence were the major determinants of disparities in nutritional status for under-five children in Ethiopian administrative zones.
(Ferdowsy et al., 2021)	k-NN, **LR**, SVM, NB, CART, RF, MLP, AdaBoost, Gradient Boosting (GB)	The LR achieves the highest accuracy of 97.09% for the classification of obesity risk.
(Hammond et al., 2019)	**LASSO**, RF, and GB Regression	The best-performing LASSO models predicted obesity with an AUC of 81.7% for girls and 76.1% for boys.
(Hemo & Rayhan, 2021)	**RF**, DT	The RF has a better performance with an accuracy of 70.1% and 72.4% for predicting stunting and underweight, respectively.
(Khan et al., 2022)	LR-STEP, LR-LASSO, DT, RF, **GB**, SVM, NN, Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA)	The GB has performed the best in terms of the smallest misclassification error (ME) for predicting stunted growth.
(Momand et al., 2020)	**RF**, NB, LR, PART Rule	PART and RF were suitable algorithms for predicting the malnutrition status of preschool-age children in Afghanistan.
(Pang et al., 2021)	DT, LR, NN, **XGB**, SVM with RBF kernel, Gaussian NB (GNB), Bernoulli NB (BNB)	The XGB yielded 0.81% AUC and achieved statistically significantly better performance on standard classifier metrics for the prediction of early childhood obesity.
(Rahman et al., 2021)	**LR**, RF, SVM	The LR identified five risk factors for stunting and underweight, and four for wasting. The RF achieved high accuracy in classifying stunted (88.3%), wasted (87.7%), and underweight (85.7%) children.
(Ridwan & Sari, 2021)	**C4.5,** NB	The C4.5 is 0.93% better in accuracy than NB for the classification of toddler nutrition status based on the anthropometric index.
(M. M. Shahriar et al., 2019)	**NN**, DT, SVM, RF, NB	The NN shows the best result with accuracy close to 86.0%, 70.0%, and 67.30% respectively, with wasting, underweight, and stunting.
(Talukder & Ahammed, 2020)	NN, LDA, SVM, LR, **RF**	The RF accurately predicted malnutrition in Bangladeshi children with 68.51% accuracy, 94.66% sensitivity, and 69.76% specificity.
(Yuliansyah et al., 2020)	SVM, K-NN, RF, **DT**, NB	DT is superior for weight data for age and height for an age, while K-NN is superior for weight data for height.

Additionally, most of these earlier frameworks did not address the issue of class imbalance caused by rare occurrences of malnutrition (stunting, wasting, and underweight) in the community. This research investigates the impact of addressing class imbalance on the performance of predictive models used to assess children's nutritional status. Therefore, it is crucial to investigate NN, RF, DT, LR, and XGB with more thorough tests, specifically addressing the issue of class imbalance. The aim of this study is to enhance ML nutritional predictive models by evaluating model performance, comparing algorithms, and addressing class imbalance to achieve accurate nutritional status predictions of Banjarmasin’s children.

2. Methodology

This research employs the experimental research method. The method includes four critical phases, which consist of (1) Data Gathering and Data Pre-processing, (2) Proposed Approach, (3) Test Model and Experiment, and (4) Result and Evaluation Metrics. Each phase is crucial to achieving the research objectives.

2.1 Data Gathering and Data Pre-processing

The population in this study consisted of mothers and their children in the city of Banjarmasin, South Kalimantan Province, Indonesia. Data was collected from 26 Puskesmas in Banjarmasin. The acquired data was used to obtain relevant attributes that fit the input algorithm format. Relevant features for children's nutritional status, linked to the BDHS 2014 (National Institute of Population Research and Training (NIPORT), 2016) are shown in Table 2.

Table 2: Selected features for children's nutritional status from different studies

Features	(Talukder & Ahammed, 2020)	(Hemo & Rayhan, 2021)	(Rahman et al., 2021)	Proposed
Child				
● Age	✓	✓	✓	✓
● Sex		✓	✓	✓
● Size at birth		✓		
● Birth Order		✓	✓	✓
● Twin			✓	
Maternal				
● Age		✓	✓	
● Education	✓	✓	✓	✓
● Current working status			✓	✓
● Media exposure		✓		
● BMI	✓	✓		✓
● Age at first birth				✓
● Preceding birth interval	✓			✓
● Currently breastfeeding		✓		
Paternal				
● Education			✓	✓
Household				
● Toilet facility			✓	
● Drinking water			✓	
● Wealth index	✓	✓	✓	✓
Community				
● Place of residence	✓	✓	✓	✓
● Division	✓	✓	✓	✓

2.2 Proposed Approach

The proposed framework is evaluated against previous prediction frameworks (Ferdowsy et al., 2021), (Rahman et al., 2021), and (M. Shahriar et al., 2019) with Figure 1(a) illustrating both the initial and optimized versions - without and with class imbalance handling, respectively. Nutritional status predictions use NN, RF, DT, XGB, and LR, all trained on identical datasets and validated through 10-fold cross-validation, dividing the training set into equal portions and repeating the learning process 10 times to mitigate overfitting (Witten et al., 2016). After optimization, all models are tested with specific data treatments for the best performance.

Treatment I: Reducing Class Labels - Class labels were simplified into two categories for comparability, transforming the WHZ score to binomial for use with LR.

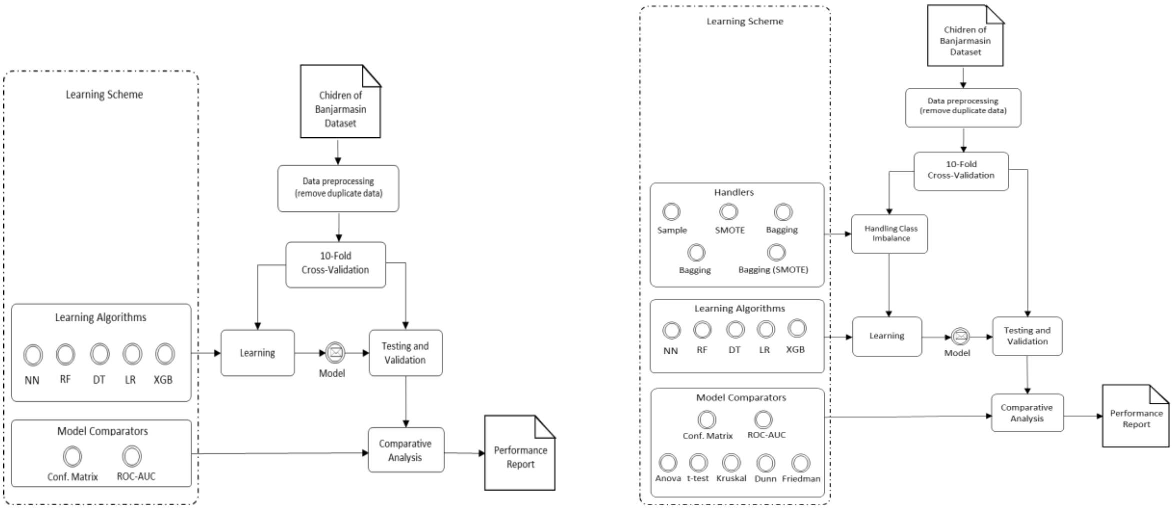
Treatment II: Data Restructuring - Features like maternal employment and education were generalized to prevent overfitting, ensuring model clarity and robustness.

Treatment III: Addressing Class Imbalance - Class imbalance, previously unaddressed, is now managed with techniques like sampling and weighting, as shown in the optimized framework Figure 1(b).

In datasets with binomial or binary label classes, one class may be much more common than another, posing a challenge for predictive modeling because most machine learning algorithms assume an equal number of examples for each class. As a result, models trained with imbalanced data tend to perform poorly, especially for minority classes. There are two main methods to overcome class imbalance: sampling and weighting. However, not all ML algorithms accept weights or sampling. Sampling itself is divided into under-sampling and over-sampling. The proposed framework will combine the following techniques to address class imbalances:

- Sample Technique (Under-sampling): This strategy removes examples from the majority class in a training dataset to balance the class distribution. When used in conjunction with an oversampling strategy for the minority class, this combination often outperforms under-sampling alone (Fernández et al., 2018).

- SMOTE (Synthetic Minority Oversampling Technique): This method generates synthetic samples for the minority class by interpolating among existing minority class instances(Fernández et al., 2018). SMOTE has been applied in several previous studies in the context of predicting or identifying nutritional status (Fernández et al., 2018). The reliability of handling class imbalances using SMOTE will be tested and compared with other techniques.
- Bagging (Bootstrap Aggregating): Introduced to enhance classification by combining classifications of randomly produced training sets (Breiman, 2001). The Bagging classifier divides a training set into numerous new sets through random sampling and creates models using these new training sets. This technique lowers variance and helps prevent overfitting, improving the performance of unstable techniques such as NN, classification, and regression trees. It also effectively handles class imbalance.



(a) (b)

Fig. 1: The proposed framework. (a) Initial framework (b) Optimized framework

2.3 Test Model and Experiment

This stage involves implementing the proposed approach and testing it through experimentation. The model's performance is compared to related frameworks to predict children's nutritional status. Key metrics, such as accuracy, are used to evaluate the effectiveness of the proposed framework.

2.4 Results and Evaluation Metrics

The experiments use AUC to assess classifier performance, supporting cross-study comparability (Rahmatullah & Noble, 2014). Precision, Recall, and F1 score offer further insights into model effectiveness. Sensitivity and Specificity are also evaluated to ensure accurate detection of nutritional issues (Talukder & Ahammed, 2020).

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positive + True\ Negative} \tag{1}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$Specificity = \frac{True\ Negative}{False\ Positive + True\ Negative} \tag{3}$$

$$F1\ Score = \frac{2\ x\ True\ Positif}{2\ x\ True\ Positif\ +\ False\ Positive\ +\ False\ Negative} \tag{4}$$

To assess the statistical significance of model performance across multiple spatial regions, a combination of parametric and non-parametric tests was employed. One-way ANOVA was applied to determine whether mean differences existed among groups, followed by paired t-tests for specific pairwise comparisons when assumptions of normality and homogeneity of variance were met. In cases where these assumptions were violated, the Kruskal–Wallis test was utilized as a robust non-parametric alternative to detect differences in median ranks. Subsequent pairwise comparisons were conducted using Dunn's test, which controls for familywise error rates and is appropriate following a significant Kruskal–Wallis result. To evaluate consistency in model rankings across repeated measures or folds, the Friedman test was implemented, offering a non-parametric approach analogous to repeated-measures ANOVA. This multi-method strategy ensured that both central tendency and ordinal relationships were rigorously examined under varying distributional conditions (Dinno, 2015; Rozaini & Khalid, 2024).

3. Results and Discussion

This section overviews the study's data and compares the performance of models in predicting nutritional status. Key metrics, including accuracy, AUC, and F1 score, are highlighted to assess each model's effectiveness. The results provide insight into the framework's strengths and areas for improvement.

3.1 Data

Data were collected from 26 Puskesmas in Banjarmasin, primarily from 2021, due to limited maternal data availability from 2018 to 2021. Initially, 75,252 rows of child data and 12,546 rows of maternal e-cohort data were collected from the following Puskesmas: Terminal, Cempaka Putih, 9 Nopember, Sungai Mesa, Sungai Bilu, Pekauman, Kelayan Timur, Kelayan Dalam, Gadang Hanyar, Karang Mekar, Pekapuran Raya, Cempaka, Teluk Dalam, Basirih Baru, Banjarmasin Indah, Pelambuan, Sungai Jingah, S. Parman, Alalak Selatan, Alalak Tengah, Kuin Raya, Teluk Tiram, Pemurus Baru, Pemurus Dalam, and Beruntung Jaya. After preprocessing, the dataset was refined to 5,664 rows, encompassing data from all Puskesmas.

3.2 Model Performance

Optimal settings for each algorithm were determined through testing, revealing the following configurations: NN) Hidden layers with sizes 100-50-50, learning rate of 0.1, using 10-fold cross-validation with shuffled sampling; RF: 100 trees, depth of 30, gain ratio criterion, using 10-fold cross-validation with shuffled sampling; DT: Depth of 30, gain ratio criterion, using 10-fold cross-validation with stratified sampling; XGB: Tree Booster, Approximate Method, depth 8, sub-sample ratio 80:20, using 10-fold cross-validation with stratified sampling; LR: L-BFGS solver (without regularization), using 10-fold cross-validation with shuffled sampling. Table 3 and Figure 2 summarize the best results from each treatment, showing relative changes (Δ) compared to previous experiments as percentages. For example, NN accuracy improved by 0.22% with data restructuring (Treatment II) compared to binomial labeling (Treatment I) and increased by 1.23% with class imbalance handling (Treatment III) compared to data restructuring.

Table 3: Summary of results using different models, treatments, and handlers

Model	Treatment	Handler	Acc.	Δ%	AUC	Δ%	F1 Score	Δ%
NN	I	None	0.892	-	0.907	-	0.695	-
	II	None	0.894	0.22%	0.900	-0.77%	0.696	0.14%
	III	Bagging (7:3) + SMOTE 10 neighbors	0.907	1.23%	0.896	-0.44%	0.707	1.15%
RF	I	None	0.823		0.876		0.099	
	II	None	0.824	0.12%	0.859	-1.94%	0.103	4.04%
	III	SMOTE 5 neighbors	0.805	-2.31%	0.848	-1.28%	0.535	419.42%
DT	I	None	0.818	-	0.686	-	0.105	-
	II	None	0.815	-0.37%	0.655	-4.52%	0.128	21.90%
	III	Bagging (7:3) + Sample 300	0.854	4.79%	0.823	25.65%	0.314	145.31%
LR	I	None	0.849	-	0.774	-	0.576	-
	II	None	0.857	0.94%	0.773	-0.13%	0.587	1.91%
	III	Bagging (7:3)	0.910	6.18%	0.884	14.36%	0.662	12.78%
XGB	I	None	0.901	-	0.895	-	0.688	-
	II	None	0.906	0.55%	0.905	1.12%	0.708	0.55%
	III	SMOTE 10 neighbors	0.907	0.11%	0.907	0.22%	0.732	3.39%

A slight increase in accuracy was observed in the Neural Network model following the application of ensemble methods (up to 90.4%) with Bagging. However, its F1 score remains relatively unchanged, emphasizing the need for precision-recall balance. A substantial improvement in F1 score was recorded for the Random Forest model (419.24%), particularly under Treatment III, where SMOTE was applied. DT gains an F1 score enhancement (21.50%) with Bagging (Treatment II), but accuracy decreases by 3.07%, likely due to inherent variance. LR stands out with high accuracy (91.0%) and a significant F1 score increase (12.78%) using Treatment III and Bagging, ensuring balanced performance. XGB optimizes under Treatment III, achieving the highest accuracy (90.7%) and F1 Score (0.732). Class imbalance handling improves F1 scores by approximately 1.15% to 12.78% (NN, LR, XGB). RF and DT achieve F1 scores exceeding 100%. However, accuracy and AUC values vary between -2.31% and 25.65%.

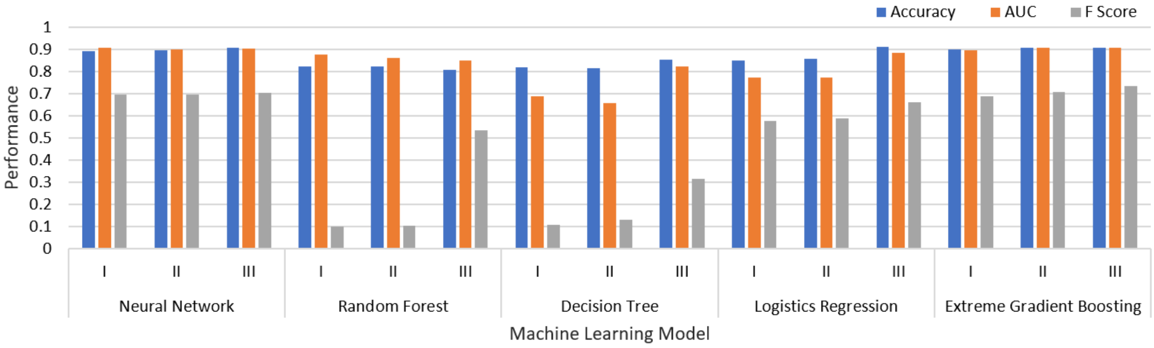


Fig. 2: Performance comparison of five machine learning models (NN, RF, DT, LR, XGB) across three treatments, evaluated using Accuracy, AUC, and F1 Score

It should be noted that accuracy and AUC metrics tend to be biased toward the majority class, thereby limiting their reliability in evaluating model performance under class imbalance conditions. The F1 score, which balances precision and recall, is crucial in mitigating this bias. While combining minor classes may offer a partial solution, handling class imbalance remains a complex challenge requiring algorithm-specific strategies. Algorithms must be carefully selected, tested, and compared across different datasets to achieve optimal performance.

In the search for feature impact on model performance, particularly concerning the F1 score (Saarela & Jauhiainen, 2021)—as illustrated in Figure 3, the features Posyandu affiliation, Maternal age, Puskesmas affiliation, Maternal blood type, and Wealth status consistently rank among the top five across all models examined.

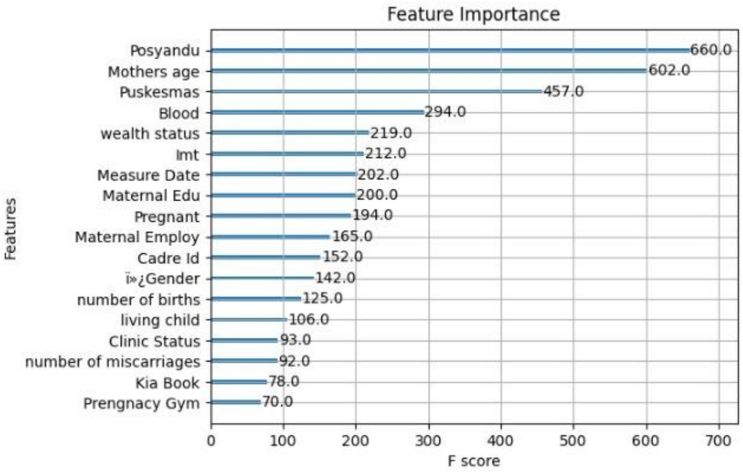


Fig. 3: Ranked feature importance based on F1 Score contribution, highlighting Posyandu affiliation, maternal age, and household wealth as top predictors.

The results of phased geographic validation indicate that the proposed model performs consistently across most regions, particularly in the North, West, and Central Banjarmasin areas. However, performance variations in South and East Banjarmasin underscore the need for adaptive thresholds and longitudinal model refinement. These findings reinforce the model's relevance as an early prediction tool.

3.3 Model Evaluation and Validation

Across all validation folds, the highest accuracy was consistently obtained by XGBoost (M = 0.909, SD = 0.005), followed closely by the Neural Network (M = 0.894, SD = 0.003). Moderate performance was recorded for Logistic Regression (M = 0.864, SD = 0.007), while the lowest accuracy scores were observed for Decision Tree and Random Forest (M = 0.820 and 0.826, respectively). A statistically significant difference in accuracy among the models was identified through a one-way ANOVA, $F(4, 45) = 112.37$, $p < .001$. Post hoc comparisons using Bonferroni-adjusted pairwise t-tests confirmed that XGBoost and Neural Network significantly outperformed Random Forest and Decision Tree ($p < .001$), with Logistic Regression occupying an intermediate position.

Regarding the F1 score, which reflects the balance between precision and recall, the highest mean was again achieved by XGBoost (M = 0.712, SD = 0.013), followed by the Neural Network (M = 0.705, SD = 0.007). Logistic Regression yielded a mean F1 score of 0.611 (SD = 0.017), whereas Decision Tree and Random Forest exhibited substantially lower scores, with Random Forest recording the lowest mean of 0.123 (SD = 0.024). A significant difference in F1 score across models was revealed by ANOVA, $F(4, 45) = 89.56$, $p < .001$. These findings were further validated using the Kruskal–Wallis test ($\chi^2 = 38.21$, $p < .001$), and Dunn's post hoc analysis confirmed that Random Forest's performance was significantly inferior to that of the other models.

To assess consistency across folds, the Friedman test was applied to both accuracy and F1 score metrics. Significant differences in model rankings were detected (accuracy: $\chi^2(4) = 38.6$, $p < .001$; F1 score: $\chi^2(4) = 36.9$, $p < .001$), reinforcing the robustness of the observed performance gaps.

In summary, XGBoost and Neural Network were identified as the most effective classifiers in this study. Their superior accuracy and F1 scores, combined with consistent performance across folds, suggest strong reliability and predictive capability compared to the other models evaluated.

3.4 Applying Framework

Tables 4 and 5 illustrate the varying effectiveness of handling methods for nutritional status indicators, such as Weight by Age Z score (WAZ) and Height by Age Z score (HAZ). For instance, while RF and XGB show relative consistency across WAZ and HAZ labels, their performance may not match that of WHZ. It is crucial to evaluate various handling methods systematically across all indicators for robust results.

Table 4: ML model performance against WAZ label

Model	Handler	Accuracy	AUC	F1 score
NN	Bagging (7:3)	0.8786	0.7854	0.1824
	Bagging (7:3) + sample 300†	0.7901	0.7799	0.3758
	SMOTE 5 neighbors	0.8887	0.7293	0.2308
	Bagging (7:3) + SMOTE 10 neighbors*	0.8607	0.7625	0.2749
RF	Bagging (7:3)	0.8883	0.7228	NaN
	Bagging (7:3) + sample 300	0.8042	0.6995	0.2752
	SMOTE 5 neighbors*†	0.8414	0.7295	0.2780
	Bagging (7:3) + SMOTE 5 neighbors	0.8791	0.6544	0.0234
DT	Bagging (7:3)	0.8873	0.6397	NaN
	Bagging (7:3) + sample 300*	0.8757	0.6030	0.0228
	SMOTE 5 neighbors†	0.2599	0.5159	0.1497
	Bagging (7:3) + SMOTE 5 neighbors	0.8728	0.6003	0.0505
LR	Bagging (7:3)*	0.8704	0.6845	0.2538
	Bagging (7:3) + sample 300†	0.7369	0.7329	0.3242
	SMOTE 5 neighbors	0.7788	0.6619	0.2490
	Bagging (7:3) + SMOTE 5 neighbors	0.7679	0.6835	0.2456
XGB	None	0.8772	0.7660	0.2112
	SMOTE 5 neighbors*†	0.8520	0.7500	0.2644

*Best for WHZ †Best for WAZ

Table 5: ML model performance against HAZ label

Model	Handler	Accuracy	AUC	F1 score
NN	Bagging (7:3)	0.8867	0.7198	0.2403
	Bagging (7:3) + sample 300†	0.7483	0.7196	0.3348
	SMOTE 5 neighbors	0.8469	0.6918	0.2491
	Bagging (7:3) + SMOTE 10 neighbors*	0.8727	0.7060	0.2757
RF	Bagging (7:3)	0.8872	0.7239	NaN
	Bagging (7:3) + sample 300	0.8112	0.6985	0.3004
	SMOTE 5 neighbors*†	0.8067	0.7113	0.3153
	Bagging (7:3) + SMOTE 5 neighbors	0.8751	0.6386	0.0719
DT	Bagging (7:3)	0.8858	0.6962	NaN
	Bagging (7:3) + sample 300*	0.8732	0.5801	0.0576
	SMOTE 5 neighbors†	0.1738	0.5246	0.1967
	Bagging (7:3) + SMOTE 5 neighbors	0.8756	0.5927	0.0919
LR	Bagging (7:3)*	0.8562	0.6225	0.1134
	Bagging (7:3) + sample 300†	0.6966	0.6605	0.2711
	SMOTE 5 neighbors	0.7344	0.5467	0.1781
	Bagging (7:3) + SMOTE 5 neighbors	0.7764	0.5788	0.1890
XGB	None	0.8761	0.7130	0.1742
	SMOTE 5 neighbors*†	0.8548	0.7110	0.2788

*Best for WHZ †Best for HAZ

Tables 4 and 5 compare the model's performance with that of related studies (See Table 6). While LR achieved 91.0% accuracy, it fell short of (Ferdowsy et al., 2021) 97.09%, for obesity prediction. RF showed strong accuracy (80.7%, 80.5%, and 84.1%), surpassing Hemo and Rahman (2021), and Talukder and Ahammed (2020), but not reaching Shahriar et al. (2019)'s higher benchmarks of 88.3%, 87.7%, and 85.7%.

The NN algorithm demonstrated strong performance with accuracy rates of 90.7% for wasting, 79.0% for being underweight, and 74.8% for stunting, surpassing results in Shahriar et al. (2019) but not reaching the levels reported by

Ferdowsy et al. (2021) and Rahman et al. (2021). Direct comparisons may be unfair without considering the F1 score, which measures precision-recall balance. XGB showed excellent performance with accuracy rates of 85.48%, 90.7%, and 85.20% for stunting, wasting, and underweight, surpassing benchmarks from Bitew et al. (2022).

In summary, while LR has not exceeded the benchmarks set by Ferdowsy et al. (2021)RF, NN, and XGB have shown significant advancements in child nutritional status detection. It is acknowledged that these frameworks have yet to exceed the high accuracy rates achieved by Rahman et al. (2021), yet the omission of the F1 score in their study calls for a cautious interpretation of these results. The necessity of a comprehensive metric evaluation is emphasized to ensure the development of accurate machine learning models that are also balanced and effective in practical healthcare applications (Zhang et al., 2023)

Table 6: Results comparison

(Author, Year)	Algorithm	Best	Stunted	Wasted	Underweight
The proposed study	RF		80.7%	80.5%	84.1%
	NN		74.8%	90.7%	79.6%
	DT		17.4%	85.4%	25.9%
	LR		69.7%	91.0%	73.7%
	XGB		85.5%	90.7%	85.2%
(Bitew et al., 2022)	LR, NN, RF, k-NN, XGB	**XGB**	67.7%	88.0%	75.7%
Ferdowsy et al., 2021	k-NN, LR, SVM, NB, CART, RF, MLP, AdaBoost, PGB	**LR**	Reported only the highest accuracy of 97.09% for obesity risk.		
Hemo & Rayhan, 2021	RF, DT	**RF**	70.1%	-	72.4%
Rahman et al., 2021	LR, RF, SVM,	**LR**	88.3%	87.7%	85.7%
			Reported only accuracy, without the critical F1 score, precision, recall, or specificity		
Shahriar et al., 2019	NN, DT, SVM, RF, NB	**NN**	67.3%	86.0%	70.0%
Talukder & Ahammed, 2020	NN, LDA, SVM, LR, RF	**RF**	68.51%	-	-

3.5 Implementation

It is necessary to identify risk factors for individuals affected by malnutrition and to understand their spatial distribution to pinpoint clusters of vulnerability within regions. Additionally, it is crucial to examine the relationship between risk factors associated with family characteristics and the spatial distribution of residences where family members suffer from malnutrition. This information serves as valuable input for government and public health policymakers. Consequently, utilizing the clustering results, an estimated public health condition map (dashboard) of Puskesmas can be generated based on their rank, which includes Good Health (Green), Moderate Health (Yellow), and Moderate Health with Concern (Red), as illustrated in Figure 4.

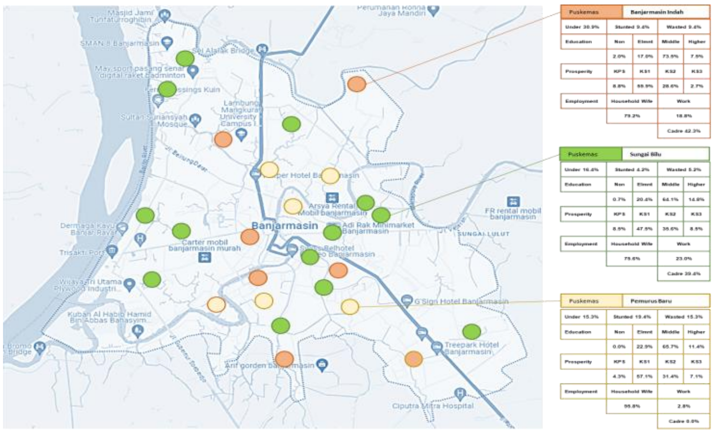


Fig. 4: Cluster-based health map of Banjarmasin's Puskesmas, categorized into three public health levels: Green (Good), Yellow (Moderate), and Red (Concern).

3.6 Limitation

While the proposed framework demonstrates promising results in predicting nutritional status, several limitations must be acknowledged. First, the dataset was derived exclusively from Puskesmas in Banjarmasin, which may limit the generalizability of the findings to other regions with different demographic or socioeconomic profiles. Spatial validation was conducted to assess whether a predictive model trained on the Banjarmasin region could be generalized across its five sub-regencies. Statistical tests revealed significant performance disparities. ANOVA indicated differences in accuracy ($p = .02$) and F1 score ($p < .01$), with Banjarmasin Utara and Tengah outperforming the parent region. Kruskal-Wallis and Dunn's post-hoc tests confirmed these findings, particularly highlighting Utara's superior accuracy and recall. Friedman's test further demonstrated inconsistent model rankings across regions ($\chi^2 = 18.2$, $p < .01$), suggesting that local data structures influenced algorithmic effectiveness. These results imply that the Banjarmasin-trained model lacks sufficient generalizability. Spatial validation revealed performance disparities across sub-regencies, suggesting that local data structures significantly influence model effectiveness.

Second, although class imbalance was addressed using SMOTE and Bagging techniques, the selection of hyperparameters (e.g., number of neighbors, sampling ratios) was based on empirical tuning rather than automated optimization, which may affect reproducibility. Additionally, the study did not incorporate temporal data or longitudinal tracking, which could enhance predictive accuracy and support early intervention strategies.

Lastly, while the F1 score was prioritized to mitigate bias from imbalanced data, other fairness metrics—such as precision-recall trade-offs across subgroups—were not explored. Future work should consider fairness-aware modeling to ensure equitable health predictions across diverse populations.

4. Conclusion

This study confirms that addressing class imbalance significantly enhances the predictive performance of machine learning models in nutritional classification tasks. XGBoost emerged as the most reliable algorithm, consistently achieving the highest accuracy and F1 scores across all indicators—wasting, underweight, and stunting. Neural Networks also demonstrated strong performance, particularly for WAZ and HAZ predictions when combined with Bagging techniques.

The integration of SMOTE and ensemble methods notably improved minority class detection, underscoring the importance of using the F1 score as a primary evaluation metric under imbalanced data conditions. Spatial validation revealed regional performance disparities, suggesting that adaptive modeling strategies—such as localized fine-tuning—may be necessary to ensure equitable prediction outcomes.

Overall, the proposed framework offers a scalable, data-driven approach to nutritional surveillance, supporting more targeted public health interventions in resource-limited settings. Future research should incorporate longitudinal data and fairness-aware modeling to enhance generalizability, precision, and policy relevance. This framework lays the foundation for integrating predictive analytics into national nutritional surveillance systems, enabling proactive and equitable health interventions.

Acknowledgement

The authors would like to express their sincere gratitude to Universiti Pendidikan Sultan Idris, MALAYSIA for providing the necessary resources and facilities to complete this study. The authors also appreciate the constructive comments from colleagues and reviewers, which have greatly improved the quality of this manuscript.

Appendix A:
This appendix provides supplementary tables and figures that support the findings presented in the main text. Table A1 shows the detailed performance metrics for each machine learning model tested, while Figure A1 illustrates the class distribution before and after the balancing techniques were applied. These additional materials are intended to give readers deeper insight into the experimental results.

References

Alqahtani, A., Albuainin, F., Alrayes, R., Al Muhanna, N., Alyahyan, E., & Aldahasi, E. (2021). Obesity Level Prediction Based on Data Mining Techniques. *IJCSNS International Journal of Computer Science and Network Security*, *21*(3), 103.

Bansod, J., Amonkar, M., Naik, A., Vaz, T., Sanke, M., & Aswale, S. (2020). Prediction of Child Development using Data Mining Approach. *International Journal of Computer Applications*, *177*(44), 13–17. https://doi.org/10.5120/ijca2020919955

Bitew, F. H., Sparks, C. S., & Nyarko, S. H. (2022). Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public Health Nutrition*, *25*(2), 269–280. https://doi.org/10.1017/S1368980021004262

Breiman, L. (2001). Random forests. Machine Learning. *Kluwer Academic Publishers. Manufactured in The Netherlands.*, *45(1)*, 5–32.

Dinkes Banjarmasin. (2024). *Profil Kesehatan Kota Banjarmasin Tahun 2023*. Dinas Kesehatan Banjarmasin.

Dinkes Kalsel. (2023). *Profil Kesehatan Provinsi Kalimantan Selatan 2022*.

Dinno, A. (2015). Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. *The Stata Journal*, *15*(1), 292–300.

Fazraningtyas, W. A., Rahmatullah, B., Salmarini, D. D., Ariffin, S. A., & Ismail, A. (2024). Recent advancements in postpartum depression prediction through machine learning approaches: a systematic review. *Bulletin of Electrical Engineering and Informatics*, *13*(4), 2729–2737. https://doi.org/10.11591/eei.v13i4.7185

Fenta, H. M., Zewotir, T., & Muluneh, E. K. (2021). A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian administrative zones. *BMC Medical Informatics and Decision Making*, *21*(1), 1–12. https://doi.org/10.1186/s12911-021-01652-1

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*(August), 100053. https://doi.org/10.1016/j.crbeha.2021.100053

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. In *Journal of Artificial Intelligence Research* (Vol. 61). https://doi.org/10.1613/jair.1.11192

Hammond, R., Athanasiadou, R., Curado, S., Aphinyanaphongs, Y., Abrams, C., Messito, M. J., Gross, R., Katzow, M., Jay, M., Razavian, N., & Elbel, B. (2019). Predicting childhood obesity using electronic health records and publicly available data. *PLoS ONE*, *14*(4). https://doi.org/10.1371/journal.pone.0215571

Hemo, S. A., & Rayhan, M. I. (2021). Classification tree and random forest model to predict under-five malnutrition in Bangladesh. *Biom Biostat Int J*, *10*(3), 116–123. https://doi.org/10.15406/bbij.2021.10.00337

Khan, J. R., Hossain, M. B., & Awan, N. (2022). Community-level environmental characteristics predictive of childhood stunting in Bangladesh - a study based on the repeated cross-sectional surveys. *International Journal of Environmental Health Research*, *32*(3), 473–486. https://doi.org/10.1080/09603123.2020.1777947

Lareno, B., Swastina, L., & Tan, F. (2020). The Mapping of Malnutrition and Stunting Through Web-Based Support System. *Asia Pacific Institute of Advanced Research (APJCECT)*, *6*(2), 30–39.

Momand, Z., Mongkolnam, P., & ... (2020). Data mining based prediction of malnutrition in Afghan children. *... on Knowledge and ...*. https://ieeexplore.ieee.org/abstract/document/9059388/

National Institute of Population Research and Training (NIPORT). (2016). Bangladesh Demographic and Health survey 2014. In *Dhaka, Bangladesh, and Rockville, Maryland, USA*.

Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, *150*(April). https://doi.org/10.1016/j.ijmedinf.2021.104454

Rahman, S. M. J., Ahmed, N. A. M. F., Abedin, M. M., Ahammed, B., Ali, M., Rahman, M. J., & Maniruzzaman, M. (2021). Investigate the risk factors of stunting, wasting, and underweight among under-five Bangladeshi children and its prediction based on machine learning approach. *PLoS ONE*, *16*(6 June 2021), 1–11. https://doi.org/10.1371/journal.pone.0253172

Rahmatullah, B., & Noble, J. A. (2014). Anatomical Object Detection in Fetal Ultrasound: Computer-Expert Agreements. *Communications in Computer and Information Science*, *404 CCIS*, 207–218. https://doi.org/10.1007/978-3-642-54121-6_18

Ridwan, A., & Sari, T. N. (2021). The comparison of accuracy between naïve bayes classifier and c4.5 algorithm in classifying toddler nutrition status based on anthropometry index. *Journal of Physics: Conference Series*, *1764*(1), 1–6. https://doi.org/10.1088/1742-6596/1764/1/012047

Rozaini, N. A., & Khalid, Z. M. (2024). *One-Way and Two-Way Analysis of Variance (ANOVA) using Parametric and Non-parametric Methods* (Vol. 22).

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*(2). https://doi.org/10.1007/s42452-021-04148-9

Setjen Kementerian Kesehatan RI. (2021). *Profil Kesehatan Indonesia Tahun 2020*.

Shahriar, M., Iqubal, M. S., Mitra, S., & Das, A. K. (2019). A deep learning approach to predict malnutrition status of 0-59 month's older children in Bangladesh. *Proceedings - 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2019*, 145–149.

https://doi.org/10.1109/ICIAICT.2019.8784823

Sinambela, D. P., Rahmatullah, B., Lah, N. H. C., & Selamat, A. W. (2024). Machine learning approaches for predicting postpartum hemorrhage: a comprehensive systematic literature review. In *Indonesian Journal of Electrical Engineering and Computer Science* (Vol. 34, Issue 3, pp. 2087–2095). Institute of Advanced Engineering and Science. https://doi.org/10.11591/ijeecs.v34.i3.pp2087-2095

Swastina, L., Rahmatullah, B., Saad, A., & Khan, H. (2024). A systematic review on research trends, datasets, algorithms, and frameworks of children's nutritional status prediction. *IAES International Journal of Artificial Intelligence*, *13*(2), 1866–1875. https://doi.org/10.11591/ijai.v13.i2.pp1868-1877

Talukder, A., & Ahammed, B. (2020). Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition*, *78*, 110861. https://doi.org/10.1016/J.NUT.2020.110861

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*.

Yuliansyah, H., Winiarti, S., Arfiani, I., & Sari, N. (2020). Comparison and Analysis of Classification Algorithm Performance for Nutritional Status Data. *International Journal of Computer Applications*, *176*(20), 14–20. https://doi.org/10.5120/ijca2020920157

Zhang, B., Rahmatullah, B., Wang, S. L., Zaidan, A. A., Zaidan, B. B., & Liu, P. (2023). A review of research on medical image confidentiality related technology, coherent taxonomy, motivations, open challenges and recommendations. *Multimedia Tools and Applications*, *82*(14). https://doi.org/10.1007/s11042-020-09629-4