International Journal of
Technology, Public Administration, and Education
Vol. 2 No. 1. June 2026, pp. 43-53

Stunting Prediction Without Derived Diagnostic Variables: A Comparison Of Random Forest, Lightgbm, Xgboost, And Logistic Regression

Muhammad Resha[1*], Apriana Toding[2]
[1] Universitas Teknologi AKBA Makassar, Makassar, South Sulawesi, Indonesia
[2] Department of Electrical Engineering, Universitas Kristen Indonesia Paulus, South Sulawesi, Indonesia

First AUTHOR : https://orcid.org/0000-0002-0335-3739/ mresha@akba.ac.id
Second AUTHOR : https://orcid.org/0000-0002-8668-1729 / apriana.toding@ukipaulus.ac.id

Corresponding author email: mresha@akba.ac.id

Article Info
Received: 07/03/2026
Revised: 24/04/2026
Accepted: 17/06/2026
Online: 30/06/2026

Abstract
Z-score-based stunting screening requires reference standards and calculation procedures that may constrain the efficiency of community health services. Evidence remains limited on whether machine-learning models can classify stunting using only basic anthropometric measurements without derived diagnostic variables.
This study developed and evaluated a Z-score-free, leakage-aware classification framework using age, sex, weight, and height. A total of 40,071 records of children aged 0–59 months from Jeneponto Regency, Indonesia, collected between 2021 and 2024 were analyzed. XGBoost, Random Forest, LightGBM, and Logistic Regression were compared using a stratified 80:20 hold-out split and stratified five-fold cross-validation.
Median imputation and SMOTE were applied exclusively to the training data. In the hold-out evaluation, Random Forest achieved 95.81% accuracy, a 94.86% F1-score, an MCC of 0.913, and a Cohen's kappa of 0.913, whereas LightGBM achieved the highest recall (95.83%) and ROC-AUC (99.28%).
Cross-validation produced a consistent pattern, with Random Forest obtaining the highest mean F1-score (94.84% ± 0.34). SHAP analysis of XGBoost identified height and age as the dominant contributors to prediction. These findings indicate that tree-based models may support preliminary screening without using Z-scores as predictors. External and temporal validation is required before broader implementation in community health posts..

Keywords: Anthropometry, data breach, machine learning, stunting, Z-Score

INTRODUCTION

Stunting reflects chronic linear growth failure and remains an important public health concern in low- and middle-income countries. Its consequences extend beyond physical growth and are associated with cognitive development, educational attainment, productivity, and long-term health(Anku & Duah, 2024)(Resha & Toding, 2026). Studies conducted in Bangladesh, Rwanda, Ghana, and India indicate that stunting patterns are shaped by child, maternal, household, socioeconomic, and

environmental factors, underscoring the continuing importance of early detection in nutritional interventions (Ntawuyirushintege et al., 2025) (Arya et al., 2025)

Stunting status is determined from height or length relative to age according to the WHO Child Growth Standards. Its application depends on accurate age and anthropometric measurements, access to reference tables or calculators, and the ability of health workers to interpret the results (Rahman et al., 2021) (Resha & Toding, 2026). Errors in anthropometric measurement and age estimation can alter both individual classification and prevalence estimates (Grange et al., 2024) . In community-based services such as integrated health posts (Posyandu), streamlined screening procedures are needed to accelerate case identification while retaining confirmation through official standards.

Machine learning has been used to classify stunting and malnutrition through Logistic Regression, Random Forest, XGBoost, LightGBM, and ensemble models. Tree-based algorithms often perform well on tabular data because they can capture nonlinear interactions; however, findings vary across studies according to population characteristics, feature sets, class distributions, and validation strategies (Islam et al., 2024) (Shen et al., 2023) . Some studies continue to emphasize accuracy while providing insufficient control over derived diagnostic variables and preprocessing-related data leakage.(Apicella et al., 2025)

A key research gap is the limited evidence that stunting can be classified using only age, sex, weight, and height without including Z-scores or nutritional-status categories as predictors. This study developed a Z-score-free, leakage-aware framework, compared three tree-based ensemble models with Logistic Regression using hold-out evaluation and stratified five-fold cross-validation, and applied SHAP to examine prediction patterns. The study aimed to assess the feasibility of a lightweight approach as the basis for a Posyandu screening aid rather than as a replacement for official nutritional-status assessment.(David, 2019)

RESEARCH METHOD

Research Design

This study used a comparative quantitative design based on controlled machine-learning experiments to develop and evaluate stunting-screening models without using Z-scores or derived nutritional-status indicators as predictors. Three tree-based ensemble algorithms Extreme Gradient Boosting, Random Forest, and Light Gradient Boosting Machine were compared with Logistic Regression as a linear baseline(Vickers & Holland, 2021). All models were trained and evaluated using the same feature set, preprocessing procedures, data partitions, and performance metrics to ensure consistent comparison.

The experimental workflow comprised annual-data consolidation, data-quality assessment, removal of features with potential leakage, stratified data splitting, missing-value imputation, class balancing in the training data, model training, internal validation, and model-interpretability analysis (Stiglic et al., 2020). Strict separation between training and test data was maintained to prevent test-set information from influencing imputation, synthetic sample generation, or model fitting

Data Sources and Unit of Analysis

The study data were obtained from four CSV files containing anthropometric records of children aged 0–59 months in Jeneponto Regency, Indonesia, from 2021 to 2024. Annual files were merged at the individual-record level. The initial dataset contained 40,072 rows. After one blank row was removed, 40,071 labeled records were retained for modeling. All records that remained after the initial quality assessment were included; no additional sampling was performed.

The class distribution comprised 23,867 non-stunted records (59.56%) and 16,204 stunted records (40.44%). Label distributions were not uniform across years: the 2021, 2022, and 2023 files contained only records labeled as stunted, whereas the 2024 file contained both classes. The stratified evaluation was therefore intended to estimate discrimination within the pooled dataset rather than temporal generalizability. Temporal validation was not conducted because the first three years did not contain both outcome classes.

Outcome Definition and Predictor Variables

The study outcome was height-for-age status as recorded in the source data and encoded into two classes: stunted and non-stunted. Stunting was designated as the positive class for all precision, recall, and F1-score calculations. Height-for-age status was retained only as the target label and was not included among the model inputs(Riley & Collins, 2023).

Predictors were restricted to four basic measurements routinely available at Posyandu: age in months, sex, weight in kilograms, and height in centimeters. All Z-score variables, height-for-age categories, weight-for-age status, weight-for-height status, and other derived nutritional indicators were excluded from the predictor matrix. This strategy was intended to ensure that the models learned patterns from basic anthropometric measurements without receiving diagnostic information that was mathematically or conceptually proximate to the target label.

Table 1. Definitions and roles of variables in modeling

Variable	Data Type	Unit/Category	Role in Analysis	Preprocessing
Height-for-age status	Binary categorical	Stunting/not stunting	Outcome or target label Encoded as binary target	Encoded as binary target
Age	Numeric	Months	Predictor	Median imputation in training pipeline
Sex	Binary categorical	Two categories as per data source	Predictor	Encoded as binary variable
Weight	Numeric	Kilograms	Predictor	Unit error correction and median imputation
Height	Numeric	Centimeters	Predictor	Used as numeric value
Z-Score and derived nutritional status indicators	Numeric/Categorical	Diverse	Not used as predictor	Removed before modeling

The procedures used to generate the original labels in the source system—including the reference standard and calculation process used to determine height-for-age status—were not documented in the study files. The recorded labels were therefore treated as available ground-truth labels from secondary data, and stunting status was not recalculated.

Data Quality Inspection and Preprocessing

Initial screening identified blank rows, missing values, unit-entry errors, and variables that could introduce data leakage. One completely blank row was removed. Two age values and one weight value were missing. Missing numerical values were handled through median imputation (Nduwayezu, Zhao, et al., 2025) . To prevent information leakage, the median was estimated exclusively from the training data and then applied to validation or test data through the same pipeline.

Several weight values displayed clear unit-entry errors. Infant weights recorded between 2,800 and 3,100 were converted to 2.8–3.1 kg, and one value of 91 was corrected to 9.1 kg. After correction, weight ranged from 2.0 to 28.5 kg. Corrections were limited to explicitly identifiable unit errors; the source study did not report winsorization, distribution-based outlier exclusion, or other numerical transformations.

Sex was encoded as a binary variable before modeling. All Z-score columns and derived nutritional-status indicators were excluded from the feature set. This structure-based exclusion was implemented to prevent target leakage, in which a model receives inputs that directly or indirectly contain information used to define the target label.

Exploratory Analysis

Exploratory analysis described class composition, yearly label distributions, the relationship between height and age by stunting status, and associations among numerical anthropometric variables. Linear relationships among age, weight, and height were evaluated using Pearson correlation coefficients. These analyses were used to characterize the data structure and potential interactions but were not used to select or remove features before modeling.

The age–height distribution was examined separately by outcome class to identify the degree of class separation and overlap. This examination also assessed whether the classification boundary appeared nonlinear, providing an empirical basis for comparing Logistic Regression with tree-based algorithms

Data Splitting and Class Imbalance Handling

The data were divided into training and test sets using a stratified 80:20 random split. Stratification preserved the proportions of stunted and non-stunted records in both subsets. A random_state of 42 was used to improve reproducibility.

The Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training data after the split. The test data were not oversampled and retained their original distribution. During cross-validation, median imputation and SMOTE were re-estimated exclusively within the training portion of each fold, while the validation fold was not used to derive imputation values or synthetic observations. This design prevented synthetic samples or distributional information from the validation and test data from entering model training.

Operationally, each training pipeline followed the sequence of numerical missing-value imputation, class balancing using SMOTE, and model fitting. All distribution-dependent steps were fitted using training data only

Model Development and Configuration

Four classification algorithms were evaluated: XGBoost, Random Forest, LightGBM, and Logistic Regression. The three tree-based algorithms were selected to evaluate their ability to capture nonlinear interactions among age, height, weight, and sex. Logistic Regression served as a linear baseline to determine whether additional model complexity provided measurable performance gains.

Hyperparameters were specified manually and held constant across experiments. Grid search, random search, Bayesian optimization, and nested cross-validation were not conducted (Nduwayezu, Mansourian, et al., 2025). Accordingly, the reported results represent fixed-configuration performance rather than the globally optimal performance of each algorithm.

Table 2. Configuration of classification algorithms

Algorithm	Main Configuration
XGBoost	n_estimators=300; learning_rate=0.05; max_depth=5; subsample=0.8; colsample_bytree=0.8; objective="binary:logistic"; eval_metric="logloss"; tree_method="hist"; random_state=42
Random Forest	n_estimators=300; criterion="gini"; max_depth=None; min_samples_split=2; min_samples_leaf=1; bootstrap=True; random_state=42
LightGBM	n_estimators=300; learning_rate=0.05; num_leaves=31; max_depth=-1; subsample=0.8; colsample_bytree=0.8; objective="binary"; random_state=42
Logistic Regression	solver="liblinear"; penalty="l2"; C=1.0; random_state=42

Fixed configurations enabled an initial comparison under controlled computational requirements. However, these results cannot establish that one algorithm family is universally superior because each model's hyperparameter space was not systematically optimized.

Validation Strategy

Model performance was evaluated using two internal-validation schemes. The first used a 20% hold-out test set that was not involved in imputation, oversampling, or model training. The second used stratified five-fold cross-validation on the modeling dataset. In each fold, four partitions were used for training and one for validation, with class proportions preserved through stratification.

All procedures that could learn distributional characteristics particularly median imputation and SMOTE were repeated within the training portion of each fold. Cross-validation performance was summarized as the mean and standard deviation across five folds. Agreement between hold-out and cross-validation results was used to evaluate the relative stability of each algorithm.

Because the 2021–2023 data contained only stunted records, temporal validation using earlier years for training and later years for testing was not performed. The validation results therefore primarily represent internal performance within the pooled Jeneponto dataset.

Performance Metrics

Classification performance was evaluated using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (ROC-AUC), Matthews correlation coefficient (MCC), and Cohen's kappa. Precision, recall, and F1-score were calculated with stunting as the positive class. Recall was emphasized because false-negative classifications may delay assessment and follow-up. MCC and Cohen's kappa complemented accuracy by accounting for agreement across both outcome classes.

Table 3. Evaluation metrics and analytical functions

Metric	Function in evaluation
Accuracy	Measures the proportion of all correct predictions
Precision	Measures the proportion of stunting predictions that are truly labeled as stunting
Recall	Measures the proportion of stunting cases successfully identified by the model
F1-score	Summarizes the balance between precision and recall
ROC-AUC	Measures the model's ability to distinguish between two classes at various decision thresholds
MCC	Measures the quality of classification by considering all components of the confusion matrix
Cohen's Kappa	Measures the agreement between predictions and labels after accounting for agreement occurring by chance

Hold-out performance was reported for each model using all metrics. During cross-validation, the mean and standard deviation of each metric were used to assess performance stability across folds.

Model Interpretability

Model interpretability was assessed using model-based feature importance and SHapley Additive exPlanations (SHAP). SHAP was applied to XGBoost because this algorithm was the original focus of the modeling framework and supports tree-based prediction explanations.

Gain-based feature importance quantified the relative contribution of each feature to improved split quality during tree construction. A SHAP summary plot was then used to assess the magnitude and direction of contributions from age, sex, weight, and height. Global importance rankings were based on the mean absolute SHAP value for each feature.

A SHAP dependence plot examined the relationship between height and its contribution to prediction, with age treated as the interaction variable. This analysis assessed whether the influence of

height varied across child age. Because SHAP was applied only to XGBoost, the interpretation describes the XGBoost decision process and does not directly explain Random Forest or LightGBM predictions.

Model Interpretability

Interpretability is analyzed using model-based feature importance and SHapley Additive exPlanations. SHAP analysis is applied to XGBoost because this algorithm is the initial focus of the research framework and supports tree-structured prediction explanations. Gain-based feature importance is used to measure the relative contribution of each feature in improving the quality of separation during tree formation. Subsequently, the SHAP summary plot is used to assess the magnitude and direction of the contributions of age, gender, weight, and height to the model's output. The global importance ranking is determined based on the average absolute SHAP values for each feature.

SHAP dependence plot is used to examine the relationship between height and its contribution to the prediction with age as an interaction variable. This analysis aims to determine whether the influence of height changes according to the child's age. Since SHAP is only applied to XGBoost, the interpretability results describe the prediction mechanism of XGBoost and do not directly explain the decision-making processes of Random Forest or LightGBM.

Computing Environment and Reproducibility

All analyses were performed in Python 3.11 using Pandas and NumPy for data management, Scikit-learn for data splitting, modeling, and evaluation, Imbalanced-learn for SMOTE implementation, and the XGBoost, LightGBM, and SHAP libraries. Experiments were conducted using CPU-based computation without GPU acceleration, allowing the workflow to be reproduced on general-purpose computing devices.

RESULTS AND DISCUSSION

Characteristics and Data Distribution

The initial dataset consists of 40,072 anthropometric records of children aged 0–59 months. After one empty row was removed, 40,071 labeled records were used in the analysis. Of that number, 23,867 records or 59.56% were categorized as not stunted, while 16,204 records or 40.44% were categorized as stunted. Summary of the flow and data distribution is presented in Table 4.

Table 4. Characteristics of the research data set

Characteristics	Number, n	Percentage
Initial recording	40,072	–
Empty lines removed	1	–
Final recording analyzed	40,071	100.00%
Not stunted	23,867	59.56%
Stunted	16,204	40.44%

Label distributions differed across collection periods. All records from 2021, 2022, and 2023 were labeled as stunted, whereas the 2024 data contained both stunted and non-stunted records. Both classes were therefore present in the pooled dataset but were not evenly represented within each year.

The age–height distribution showed that children labeled as stunted generally occupied lower height ranges than non-stunted children of comparable age. Nevertheless, overlap between the two classes remained at several age–height combinations.

Pearson correlation analysis showed strong positive associations among the numerical anthropometric variables. Height was correlated with weight ($r = 0.94$) and age ($r = 0.93$), while age was correlated with weight ($r = 0.88$). The correlation matrix may be presented as supplementary material so that the main results remain focused on predictive performance

Pearson correlation analysis shows a strong positive relationship among numerical anthropometric variables. Height is correlated with weight at $r=0.94$ and with age at $r=0.93$, while the

correlation between age and weight is r=0.88. The correlation matrix can be included as supplementary material so that the main results section remains focused on prediction performance.

Model Performance on Hold-Out Test Set

The hold-out performance of the four algorithms is presented in Table 5. Random Forest achieved the highest accuracy (95.81%), F1-score (94.86%), MCC (0.913), and Cohen's kappa (0.913). LightGBM achieved the highest recall (95.83%) and ROC-AUC (99.28%).

XGBoost achieved 94.98% accuracy, 95.46% recall, a 93.90% F1-score, and a 99.12% ROC-AUC. Logistic Regression produced the lowest values across the major metrics, including 84.22% accuracy, an 81.32% F1-score, and a 91.53% ROC-AUC.

Table 5. Comparison of model performance on hold-out data

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)	MCC	Cohen's Kappa
Random Forest	95.81	94.00	95.74	94.86	99.09	0.913	0.913
LightGBM	95.72	93.72	95.83	94.77	99.28	0.912	0.911
XGBoost	94.98	92.39	95.46	93.90	99.12	0.897	0.896
Logistic Regression	84.22	77.99	84.94	81.32	91.53	0.679	0.677

Overall, the three tree-based models outperformed Logistic Regression. Differences between Random Forest and LightGBM were small: Random Forest performed best on threshold-dependent balanced metrics, whereas LightGBM achieved the highest recall and ROC-AUC.

Consistency of Performance in Cross-Validation

Stratified five-fold cross-validation produced a pattern consistent with the hold-out evaluation. Random Forest achieved the highest mean accuracy (95.81% ± 0.28) and mean F1-score (94.84% ± 0.34), as well as the highest mean MCC and Cohen's kappa (both 0.913 ± 0.006).

LightGBM achieved the highest mean recall (95.51% ± 0.40) and ROC-AUC (99.22% ± 0.07). XGBoost obtained a mean F1-score of 93.86% ± 0.35, whereas Logistic Regression obtained 81.29% ± 0.68.

Table 6. Model performance based on stratified five-fold cross-validation

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)	MCC	Cohen's Kappa
Random Forest	95.81 ± 0.28	94.51 ± 0.38	95.17 ± 0.34	94.84 ± 0.34	99.07 ± 0.12	0.913 ± 0.006	0.913 ± 0.006
LightGBM	95.54 ± 0.23	93.58 ± 0.29	95.51 ± 0.40	94.54 ± 0.28	99.22 ± 0.07	0.908 ± 0.005	0.908 ± 0.005
XGBoost	94.96 ± 0.28	92.46 ± 0.34	95.30 ± 0.54	93.86 ± 0.35	99.07 ± 0.06	0.896 ± 0.006	0.896 ± 0.006
Logistic Regression	84.18 ± 0.58	77.93 ± 0.68	84.94 ± 0.75	81.29 ± 0.68	91.57 ± 0.50	0.678 ± 0.012	0.676 ± 0.012

The low standard deviations of the three tree-based models indicated relatively stable performance across the five folds. Model rankings were also consistent with the hold-out results: Random Forest achieved the highest F1-score, LightGBM achieved the highest ROC-AUC and recall, and Logistic Regression remained the weakest model.

Variable Contribution to XGBoost Predictions XGBoost

XGBoost interpretability analysis identified height and age as the two most influential features. Gain-based feature importance assigned 42.34% importance to height, followed by age (34.84%), weight (12.05%), and sex (10.77%).

The mean absolute SHAP values produced a similar pattern. Height accounted for 50.54% of total absolute SHAP contribution and age for 37.04%, while sex and weight accounted for 8.60% and 3.83%, respectively.

Table 7. Feature importance and SHAP contribution in the XGBoost model

Variable	XGBoost gain importance (%)	Average contribution | SHAP | (%)
Height	42.34	50.54
Age	34.84	37.04
Weight	12.05	3.83
Gender	10.77	8.60

The SHAP summary plot indicated that lower height values generally contributed toward the stunted-class prediction, whereas higher height values shifted predictions toward the non-stunted class. Age also made a substantial contribution, although its direction and magnitude varied according to each child's age–height combination.

The SHAP dependence plot showed a nonlinear relationship between height and its contribution to prediction. SHAP values for height decreased as height increased, while age modified the magnitude of this contribution. Records with similar height values could therefore receive different predictive contributions when they belonged to different age groups.

DISCUSSION

Comparative Performance and Model Stability

This study demonstrated that high internal classification performance could be achieved using four basic measurements without including Z-scores or derived nutritional-status indicators. Random Forest provided the best balance across accuracy, F1-score, MCC, and kappa, whereas LightGBM achieved the highest recall and ROC-AUC. XGBoost remained competitive, while Logistic Regression underperformed the three tree-based models. The Random Forest result is consistent with malnutrition and stunting studies conducted in Bangladesh, India, and Rwanda, although other studies have identified XGBoost or alternative algorithms as the strongest model (Arya et al., 2025; Nduwayezu et al., 2025; Talukder & Ahammed, 2020). Model rankings should therefore be interpreted as specific to this dataset, configuration, and validation strategy.

F1-score, MCC, and kappa complemented accuracy by assessing balanced classification across both classes. Agreement between hold-out and cross-validation results, together with low standard deviations, supported relative internal stability. However, stable mean performance across folds does not guarantee stable individual predictions, adequate calibration, or transportability to new populations. Apparently stable models may change when redeveloped in different samples; external validation is therefore required to assess generalization across locations, periods, and measurement procedures (Collins et al., 2024; Riley & Collins, 2023).

LightGBM's slightly higher recall is relevant for screening because it may reduce missed stunting cases. Nevertheless, model selection should not rely on recall or ROC-AUC alone; evaluation of health prediction models should include discrimination, calibration, and decision utility at clinically meaningful thresholds (Collins et al., 2024; Efthimiou et al., 2024). XGBoost remained useful for interpretation through SHAP, but its explanations do not automatically represent Random Forest or LightGBM. Because interpretability depends on the model and reference data, the model selected for implementation should be explained directly (Lundberg & Lee, 2017; Stiglic et al., 2020).

The lower performance of Logistic Regression supports the presence of nonlinear age–height patterns that tree-based models may capture more readily. However, the baseline did not include interactions, splines, or nonlinear transformations, and the results therefore do not demonstrate that logistic regression is inherently unsuitable. Consistent model rankings indicate only initial internal stability; both the hold-out and cross-validation samples originated from the same pooled data source and did not test transportability across years, facilities, measurement personnel, or regencies.

Interpretation of Predictions from Anthropometric Measurements

The dominance of height and age in gain importance and SHAP was consistent with the structure of the stunting label, which is defined by height or length relative to age and sex. The SHAP dependence plot indicated that the contribution of height varied by age, suggesting that XGBoost learned a nonlinear boundary. Similar patterns have been reported in stunting studies from Rwanda, Papua New Guinea, Ghana, and Bangladesh, although those studies also included maternal, socioeconomic, sanitation, and feeding-related factors (Anku & Duah, 2024; Islam et al., 2024; Ndagijimana et al., 2023; Shen et al., 2023). SHAP explains how the model uses features; it does not establish causal relationships (Lundberg & Lee, 2017; Stiglic et al., 2020).

Excluding Z-scores and nutritional-status categories reduced the most direct form of target leakage; nevertheless, age, sex, and height remain the basic components used to construct the label. The model is therefore more appropriately described as a Z-score-free surrogate classifier than as an etiological model independent of the WHO definition. The smaller contributions of weight and sex do not justify removing them, because gain importance and SHAP quantify different aspects and globally weak features may remain important within specific subgroups. Ablation studies are needed to assess changes in discrimination, calibration, and subgroup error after feature removal.

Methodological Contribution and Novelty

The principal methodological contribution is a leakage-aware pipeline that separates training and test data and confines imputation and SMOTE to the training portion of each fold. This practice reduces optimistic performance estimates caused by validation information entering preprocessing; data leakage can artificially inflate performance and alter feature interpretation (Rosenblatt et al., 2024). Excluding Z-scores, height-for-age categories, weight-for-age status, weight-for-height status, and other derived diagnostic indicators provided a stricter test of whether the models could classify labels from routine measurements. The novelty is therefore operational and methodological rather than complete independence from the deterministic basis of stunting diagnosis.

Practical Implications for Posyandu-Based Screening

Four simple inputs create an opportunity to develop a lightweight screening module for Posyandu applications, but operational benefit cannot be inferred from algorithmic performance alone. Age, sex, and height are also required by standard Z-score calculators; implementation studies should therefore compare completion time, input errors, connectivity requirements, health-worker acceptance, and agreement with standard procedures. Digital technologies may support field services, but their effectiveness remains dependent on input quality, devices, child positioning, user training, and follow-up workflows (Mukherjee et al., 2026).

The model should be positioned as a triage aid rather than a replacement for official nutritional-status assessment. Predictions may prioritize children for repeat measurement or professional assessment, but probability estimates require calibration and decision thresholds should reflect the consequences of false negatives and false positives. Decision-curve analysis is needed to determine whether the model provides net benefit relative to referring all children or referring none (Vickers & Holland, 2021).

Strengths

This study had several methodological strengths. First, the large dataset provided sufficient observations for an internal comparison of four algorithms. Second, excluding derived diagnostic indicators reduced an explicit form of target leakage. Third, imputation and SMOTE were restricted to the training data, including within each cross-validation fold. Fourth, performance was reported not only through accuracy but also through precision, recall, F1-score, ROC-AUC, MCC, and Cohen's kappa. Fifth, SHAP provided an initial view of the patterns used by XGBoost to generate predictions. Together, these features provide greater transparency than experiments reporting only a single data split and one performance metric.

Limitations

The principal limitation was the temporal structure of the data. All records from 2021–2023 were labeled as stunted, whereas both classes first appeared together in 2024. This pattern may reflect

changes in recording scope, export procedures, or case selection. Random stratification mixed periods with different label mechanisms, meaning that the model may have learned recording characteristics rather than growth patterns alone; temporal validation was also not possible.

The unit of analysis was not confirmed to represent unique children. If a child appeared across several visits or years, record-level random splitting could place observations from the same child in both training and test sets, causing group leakage and inflated performance (Rosenblatt et al., 2024). Future datasets require anonymous identifiers and grouped splitting. Label validity could not be verified because reference standards, instrument calibration, measurement protocols, and assessor consistency were undocumented; errors in age or height may change the classification that the model subsequently learns (Grange et al., 2024).

Restricting the model to four predictors improved simplicity but limited its role to anthropometric-status classification and did not explain causes of stunting, which are also influenced by maternal factors, diet, infection, sanitation, socioeconomic conditions, and the environment (Anku & Duah, 2024; Arya et al., 2025; Rahman et al., 2021). Fixed hyperparameters, the absence of hold-out confidence intervals, calibration, fairness analysis, decision-curve analysis, and complete between-model statistical tests also limited claims of superiority and implementation readiness (Efthimiou et al., 2024).

Future research should use consistently collected annual datasets containing both classes, anonymous child identifiers, temporal and external validation, nested hyperparameter tuning, calibration assessment, ablation studies, subgroup analysis, and interpretation of the final candidate model. Prospective implementation studies should compare the model with an official Z-score calculator in terms of completion time, input errors, health-worker acceptance, costs, referral patterns, and case detection

CONCLUSION

This study showed that stunting status could be classified with high internal performance using age, sex, weight, and height without including Z-scores or derived nutritional-status indicators as predictors. Under the evaluated configurations, Random Forest provided the best balance across accuracy, F1-score, MCC, and Cohen's kappa, whereas LightGBM achieved the highest recall and ROC-AUC. Consistent hold-out and stratified five-fold cross-validation results indicated relative stability among the three tree-based models compared with Logistic Regression. SHAP analysis of XGBoost further identified height and age as the dominant contributors, with a nonlinear relationship that reflected the importance of interpreting height relative to age.

The main contribution of this study is a leakage-aware modeling framework that restricts inputs to basic anthropometric measurements and applies imputation and SMOTE exclusively within training. The framework may be developed as a lightweight screening aid for prioritizing follow-up assessment at Posyandu, but it cannot replace official nutritional-status evaluation. Generalizability remains constrained by the single-regency dataset, nonuniform temporal label distribution, uncertainty in label generation, and the absence of temporal and external validation. Future studies should evaluate the model prospectively across different locations and periods, assess calibration and decision thresholds, and directly compare its operational utility with standard Z-score calculation procedures.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to AKBA University of Technology Makassar and Universitas Kristen Indonesia Paulus for the support, cooperation, and facilities provided during the implementation of this research. Appreciation is also extended to all parties who contributed to the smooth running of this activity.

AUTHOR CONTRIBUTIONS

Conceptualization, M.R., and A.T.; Methodology, M.R.,; Software, M.R.; Formal Analysis, M.R., and A.T.; Investigation, A.T.; Data Curation, M.R.; Writing-Original Draft Preparation, M.R.,; Writing-Review & Editing, M.R., and A.T.; Project Administration, M.R.,; Funding Acquisition, A.T., All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Anku, E. K., & Duah, H. O. (2024). Predicting and identifying factors associated with undernutrition among children under five years in Ghana using machine learning algorithms. *PLOS ONE*, *19*(2), e0296625. https://doi.org/10.1371/journal.pone.0296625

Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review*, *58*(11), 339. https://doi.org/10.1007/s10462-025-11326-3

Arya, P. K., Sur, K., Kundu, T., Dhote, S., & Singh, S. K. (2025). Unveiling predictive factors for household-level stunting in India: A machine learning approach using NFHS-5 and satellite-driven data. *Nutrition*, *132*. https://doi.org/10.1016/j.nut.2024.112674

David, Z. (2019). Information leakage in financial machine learning research. *Algorithmic Finance*, *8*(1–2), 1–4. https://doi.org/10.3233/AF-190900

Grange, J. M., Mock, N. B., & Collins, S. M. (2024). Influence of non-directional errors in anthropometric measurements and age estimation on anthropometric prevalence indicators. *PLOS ONE*, *19*(9), e0304131. https://doi.org/10.1371/journal.pone.0304131

Islam, Md. M., Kibria, N. Md. S. J., Kumar, S., Roy, D. C., & Karim, Md. R. (2024). Prediction of undernutrition and identification of its influencing predictors among under-five children in Bangladesh using explainable machine learning algorithms. *PLOS ONE*, *19*(12), e0315393. https://doi.org/10.1371/journal.pone.0315393

Nduwayezu, G., Mansourian, A., Bizimana, J. P., & Pilesjö, P. (2025). Hybridizing spatial machine learning to explore the fine-scale heterogeneity between stunting prevalence and its associated risk determinants in Rwanda. *Geo-Spatial Information Science*, *28*(6), 2947–2967. https://doi.org/10.1080/10095020.2025.2459133

Nduwayezu, G., Zhao, P., Pilesjö, P., Bizimana, J. P., & Mansourian, A. (2025). Multilevel small-area childhood stunting risk estimation: Insights from spatial ensemble learning, agro-ecological and environmentally remotely sensed indicators. *Environmental and Sustainability Indicators*, *27*, 100822. https://doi.org/10.1016/j.indic.2025.100822

Ntawuyirushintege, S., Ahmed, A., Bucyibaruta, G., Siddig, E. E., Remera, E., Tediosi, F., & Wyss, K. (2025). Spatiotemporal Trends in Stunting Prevalence Among Children Aged Two Years Old in Rwanda (2020–2024): A Retrospective Analysis. *Nutrients*, *17*(17), 2808. https://doi.org/10.3390/nu17172808

Rahman, S. M. J., Ahmed, N. A. M. F., Abedin, Md. M., Ahammed, B., Ali, M., Rahman, Md. J., & Maniruzzaman, Md. (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), e0253172. https://doi.org/10.1371/journal.pone.0253172

Resha, M., & Toding, A. (2026). Predicting Independent Z-Score Stunting Through Fundamental Anthropometric Measurements Utilizing Extreme Gradient Boosting (XGBoost). JOURNAL ZETROEM, 8(2), 148–154. https://doi.org/10.36526/ztr.v8i2.8736

Riley, R. D., & Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, *65*(8). https://doi.org/10.1002/bimj.202200302

Shen, H., Zhao, H., & Jiang, Y. (2023). Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea. *Children*, *10*(10), 1638. https://doi.org/10.3390/children10101638

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, *10*(5). https://doi.org/10.1002/widm.1379

Vickers, A. J., & Holland, F. (2021). Decision curve analysis to evaluate the clinical benefit of prediction models. *The Spine Journal*, *21*(10), 1643–1648. https://doi.org/10.1016/j.spinee.2021.02.024