JOURNAL OF COMPUTER SCIENCE AND INFORMATICS ENGINEERING
VOL. 05, NO. 3, JULY 2026: 481-492
Doi: https://doi.org/10.55537/cosie.v5i3.1825

e-ISSN 2827-8356
p-ISSN 2827-9999

Threshold Optimized Random Forest and XGBoost with SHAP for Low Birth Length Risk Prediction

Rian Andi Saputra[1], Albert Yakobus Chandra[2]

[1,2]*Information Systems Study Program, Faculty of Information Technology, Mercu Buana University Yogyakarta, Yogyakarta, Indonesia*

[1]221210007@student.mercubuana-yogya.ac.id , [2]albert.ch@mercubuana-yogya.ac.id

Received: June 20, 2026 | Revised: July 21, 2026 | Accepted: July 31, 2026

Abstract

Stunting is a chronic nutritional disorder that impairs children's cognitive development and long-term productivity, making early risk identification essential. However, previous studies have mainly focused on stunting classification or conventional machine learning models, with limited attention to integrating probability threshold optimization and explainable artificial intelligence (XAI) for low birth length risk prediction. This study proposes a predictive framework using Random Forest (RF) and XGBoost based on 249,626 valid records from the 2024 Indonesian Nutritional Status Survey (SSGI), where low birth length was defined as a birth length below 48 cm. The models were optimized through hyperparameter tuning, probability threshold optimization, Stratified 5-Fold Cross Validation, and SHAP analysis. Evaluation on 10,000 testing records showed that Random Forest accuracy increased from 76.76% to 78.58%, while XGBoost achieved the highest accuracy of 84.07% at a threshold of 0.79, although recall decreased to 27%. SHAP identified birth weight as the most influential predictor. These findings demonstrate that threshold optimization improves predictive accuracy, while SHAP enhances model interpretability to support early screening and clinical decision-making for low birth length risk.

Keywords: Birth Length Risk; Random Forest; XGBoost; SHAP; Threshold Optimization

1. INTRODUCTION

Stunting is a chronic growth failure condition caused by prolonged nutritional deficiencies, defined as a Height-for-Age Z-Score (HAZ) below 2 standard deviations according to World Health Organization (WHO) growth standards. Children affected by stunting face higher risks of cognitive impairment, reduced learning capacity, lower productivity, and increased susceptibility to health problems throughout their lives. Despite continuous national intervention programs, stunting continues to present a significant epidemiological hurdle domestically. Based on the 2024 SSGI report, Indonesia's national stunting prevalence reached 19.8%, exceeding the government's 14% reduction target. This persistent gap underscores the urgency for more robust, data-driven approaches to early detection, given that machine learning methods have shown considerable promise in identifying at-risk children and facilitating timely nutritional interventions [1].

Low Birth Length (LBL), defined as a birth length below 48 cm, is one of the earliest indicators associated with stunting in later childhood. Infants born below this threshold are more likely to experience growth failure due to intrauterine factors, particularly maternal nutritional deficiencies affecting fetal linear growth [2]. In Indonesia, the Ministry of Health Regulation No. 2 of 2020 classifies birth length below 48 cm as a risk criterion requiring further monitoring and intervention. Therefore, this study adopts LBL risk as the prediction target to support earlier prevention efforts before stunting becomes clinically apparent.

e-ISSN 2827-8356
p-ISSN 2827-9999

Utilizing computational intelligence algorithms within epidemiological and nutritional diagnostics has gained traction in recent years for growth deficit forecasting, with historical literature documenting classification scores from 61% to 71.4% ,[3], [4], while Random Forest and XGBoost achieved accuracy values between 68% and 71% with AUC scores of 0.71–0.74 on demographic and health survey data . Childhood stunting is influenced by multiple factors including maternal health, household socioeconomic conditions, sanitation, food security, and environmental determinants [5].

Recent studies further highlight the effectiveness of ensemble and boosting-based algorithms: XGBoost combined with SHAP has demonstrated strong predictive performance with interpretable feature contributions [6]. Ensemble methods have outperformed conventional approaches in under-five stunting prediction [7], [8], [9], and machine learning has been successfully applied to identify spatial, environmental, and socioeconomic determinants of child growth outcomes [2]. Combining SMOTEENN, ensemble frameworks, and additive explanations-based interpretation has also shown promising results in household-level stunting prediction , [10], while Bayesian Optimization has further improved XGBoost performance and comparative evaluations have confirmed competitive results between Random Forest and XGBoost for stunting detection [11].

Model interpretability has become equally important alongside predictive performance in healthcare applications. Explainable Artificial Intelligence (XAI) enables healthcare professionals and policymakers to understand the factors influencing model decisions and supports evidence-based interventions. Among XAI techniques, the SHAP mechanism has emerged as a cornerstone methodology for quantifying feature contributions through cooperative game theory concepts [12], [13], providing valuable insights beyond conventional performance metrics . A recent review of explainable AI in biomedical sciences identified SHAP as the most frequently used post-hoc explanation technique, emphasizing its role in building trust and transparency in high-stakes healthcare applications [14], [15].

Notwithstanding these advances, notable gaps persist in the literature. Existing work predominantly targets stunting status classification rather than the earlier indicator of low birth length risk, and rarely incorporates nationally representative Indonesian datasets [16]. Probability threshold optimization is rarely investigated despite its substantial influence on classification outcomes. Furthermore, studies integrating hyperparameter tuning, model stability validation, and SHAP-based interpretation within a single predictive framework remain limited [6], [10], [12]. This study therefore proposes an ensemble learning framework based on Random Forest and XGBoost for predicting low birth length risk using the SSGI 2024 dataset. The novelty of this study lies in four key contributions. First, it shifts the prediction focus from stunting classification to the earlier physiological indicator of low birth length risk, enabling proactive intervention before stunting becomes clinically apparent. Second, it integrates systematic grid-search hyperparameter tuning, dual-objective threshold optimization (F1-score for screening and accuracy for diagnosis), Stratified 5-Fold Cross-Validation, and SHAP-based explainability within a unified ensemble framework applied to the large-scale 2024 SSGI national dataset. Third, it provides a comprehensive comparison of optimized Random Forest and XGBoost models using multiple evaluation metrics including MCC, which is particularly robust for imbalanced classification. Fourth, it translates model outputs into clinically interpretable feature contributions, bridging the gap between predictive analytics and public health practice in the Indonesian context.

Practically, this framework can serve as an early screening tool in public health settings, supporting timely preventive interventions and informing resource allocation decisions.

e-ISSN 2827-8356
p-ISSN 2827-9999

2. METHODS

2.1 Research Type and Design

This study is a machine learning-based experiment employing a quantitative approach. The research stages include data collection, preprocessing, baseline model development, hyperparameter optimization, model evaluation, and model interpretability using SHAP.

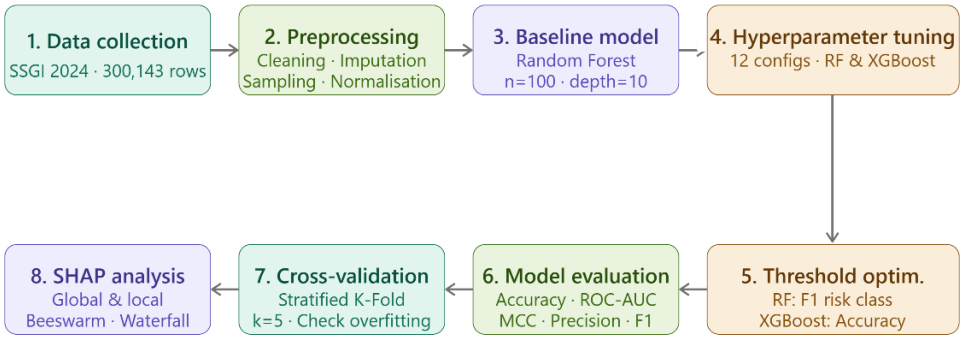

Figure 1. Research workflow illustrating the sequential stages of data preprocessing, model development, hyperparameter optimization, threshold calibration, cross-validation, and SHAP-based interpretation.

2.2 Data Collection

The data were extracted from the 2024 Nutritional Status Survey (SSGI), an official archive curated by the Indonesian Ministry of Health. The compiled registry contains 300,143 records and 14 features representing maternal, infant, environmental, and disease-history factors. The target variable was defined as Low Birth Length Risk (LBL < 48 cm) according to Indonesian Ministry of Health Regulation No. 2 of 2020.

2.3 Preprocessing and Exploratory Data Analysis (EDA)

Before model development, the SSGI 2024 dataset containing 300,143 records and 14 features underwent several preprocessing stages. Invalid codes (8, 88, 9, 99, 999, and 9999) were converted into missing values, followed by biological range filtering for birth length, birth weight, and maternal age. Records with missing target values were removed, resulting in 249,626 valid records. Missing numerical values were then imputed using the median of each feature. Because the cleaned dataset exceeded the available computational memory during repeated hyperparameter tuning, cross-validation, and SHAP analysis, stratified sampling was applied to obtain 50,000 representative records while preserving the original class distribution. This approach reduced computational cost without altering the class imbalance characteristics of the dataset, allowing the developed models to remain representative of the original population. The cleaned observations were then partitioned into an 80% development subset and a 20% validation matrix, and numerical features were standardized using StandardScaler. Finally, Exploratory Data Analysis (EDA) was conducted to examine data distributions, class imbalance, and potential risk factor patterns.The detailed preprocessing stages are presented in Table 1.

Table 1.Preprocessing Data Sampling

Stage	Initial Value	Value After Preprocessing	Description
Cleaning	Birth Weight = 9999	NaN	Invalid code converted to a missing value
Filtering	Birth Length = 70 cm	Removed	Outside the biological range of 30–60 cm
Imputation	Birth Weight = NaN	3150 grams	Replaced with the feature median
Sampling	Record No. 125000	Retained	Selected through Stratified Sampling

e-ISSN 2827-8356
p-ISSN 2827-9999

Normalization	Birth Weight = 3150	0.12	Standardized using StandardScaler

2.4 Baseline Model Development

The initial model was developed using the Random Forest algorithm with default configurations as the baseline performance benchmark: n_estimators = 100, max_depth = 10, class_weight = "balanced", and random_state = 42.

2.5 Hyperparameter Tuning

Optimization was conducted by evaluating 12 hyperparameter configurations for each algorithm using ROC-AUC as the primary selection criterion on a 90:10 train-validation split. For Random Forest, the explored parameters included n_estimators (150–300), max_depth (10–20), max_features (sqrt, log2), min_samples_leaf (1–2), and class_weight (balanced, balanced_subsample). For XGBoost [17], the explored parameters included learning_rate (0.05–0.10), max_depth (4–6), n_estimators (300–500), subsample (0.8–0.9), colsample_bytree (0.8–0.9), reg_lambda (0.5–1.5), min_child_weight (1–3), and scale_pos_weight to address class imbalance. The best-performing configuration was subsequently retrained using the full training dataset.

Hyperparameter tuning was performed using a systematic grid search on a 90:10 stratified split of the training data, with the ROC-AUC score serving as the primary selection criterion to ensure robust discriminative performance. For Random Forest, the search space comprised 12 distinct combinations of n_estimators (150, 200, 300), max_depth (10, 15, 20), max_features ('sqrt' or 'log2'), min_samples_leaf (1 or 2), and class_weight ('balanced' or 'balanced_subsample'). The optimal configuration, designated RF-01, employed n_estimators = 200, max_depth = 15, max_features = 'sqrt', min_samples_leaf = 1, and class_weight = 'balanced'. For XGBoost, the grid search similarly evaluated 12 configurations across learning_rate (0.05, 0.10), max_depth (4, 6), n_estimators (300, 400, 500), subsample (0.8, 0.9), colsample_bytree (0.8, 0.9), reg_lambda (0.5, 1.5), and min_child_weight (1, 3), with scale_pos_weight fixed at 4.46 to directly counter the inherent class imbalance. The best-performing XGBoost model, XGB-04, utilized learning_rate = 0.05, max_depth = 4, n_estimators = 400, subsample = 0.9, colsample_bytree = 0.9, reg_lambda = 0.5, and min_child_weight = 1. Following this selection, both winning configurations were retrained on the entirety of the full training partition to finalize the models for evaluation.

2.6 Threshold Optimization

Probability threshold optimization was performed to adjust model sensitivity according to the intended clinical objective. For the Tuned Random Forest model, the threshold was optimized using the F1-score of the Risk class because this metric provides a balanced trade-off between precision and recall for the minority class, making it more suitable for identifying children at risk of low birth length during early screening. In contrast, the Tuned XGBoost model was optimized using overall accuracy because the objective was to obtain the highest overall classification performance after hyperparameter optimization and class imbalance adjustment. Thresholds between 0.25–0.75 were evaluated for Random Forest, while thresholds between 0.15–0.85 were evaluated for XGBoost. Using different optimization criteria allowed each algorithm to be evaluated according to its primary objective, namely balanced minority-class detection for Random Forest and maximum overall predictive performance for XGBoost. The resulting thresholds represented screening-oriented (high recall) and diagnosis-oriented (high precision) scenarios.

The choice of optimization criteria reflected the primary intended use case of each model within the broader screening pipeline. For the Random Forest model, F1-score of the Risk class was selected because this model was designed as a screening tool where balanced detection of at-risk infants is critical—minimizing both false negatives (missed cases) and false positives (unnecessary follow-ups). In contrast, the XGBoost model was optimized for accuracy after class

e-ISSN 2827-8356
p-ISSN 2827-9999

imbalance had been addressed via scale_pos_weight, as this model was intended for confirmatory risk stratification where overall classification reliability is paramount. This dual-objective approach allows stakeholders to select between models based on their specific operational requirements: high-sensitivity screening (Random Forest at optimal F1 threshold) versus high-specificity diagnosis (XGBoost at optimal accuracy threshold).

2.7 Model Evaluation

Each model was evaluated using an unseen testing set consisting of 10,000 records. Evaluation metrics included Accuracy, Precision, Recall, F1-Score, ROC-AUC, and Matthews Correlation Coefficient (MCC). MCC was prioritized as a supplementary metric given its robustness under class imbalance, since it accounts for all confusion matrix elements proportionally.

2.8 Stability Validation Using Cross-Validation

To assess model stability and reduce potential bias from a single train-test split, Stratified K-Fold Cross-Validation (k = 5) was employed on the baseline Random Forest model [18]. Each fold rotation allowed the model to train and validate on non-overlapping subsets, with accuracy mean and standard deviation serving as stability indicators, while proximity between cross-validation and test accuracy was used as an overfitting diagnostic.

2.9 Model Interpretability Using SHAP

Model interpretability was performed using SHAP (SHapley Additive Explanations), a widely used explainable AI technique in healthcare applications [12]. The analysis employed PermutationExplainer on 100 representative testing samples due to memory limitations. Interpretation was conducted at the global level using feature importance and beeswarm plots, and at the local level using waterfall plots [19]. This analysis aimed to identify feature contributions and improve model transparency.

2.10 Implementation Details

All experiments were implemented in Python 3.10.12. Key libraries used included: scikit-learn 1.3.0 for Random Forest, preprocessing, cross-validation, and evaluation metrics; XGBoost 2.0.1 for gradient boosting; SHAP 0.44.0 for model interpretability; pandas 2.0.3 and numpy 1.24.3 for data manipulation; and matplotlib 3.7.2 with seaborn 0.12.2 for visualization. Experiments were conducted on a system equipped with an Intel Core i7-12700H processor, 16 GB RAM, and an NVIDIA GeForce RTX 3060 GPU. A fixed random seed of 42 was applied across all stochastic processes, including data splitting (train-test split, cross-validation fold generation), model initialization (random_state parameter), and sampling procedures. The complete code and supplementary materials are available from the corresponding author upon reasonable request.

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Exploratory Data Analysis (EDA)

The exploratory data analysis revealed class imbalance, with 81.7% Normal and 18.3% Risk observations. Low birth weight, maternal anemia, maternal chronic energy deficiency (CED), and unprotected drinking water sources were identified as major factors associated with Low Birth Length Risk. The distribution of categorical features is shown in Figure 2.

e-ISSN 2827-8356
p-ISSN 2827-9999

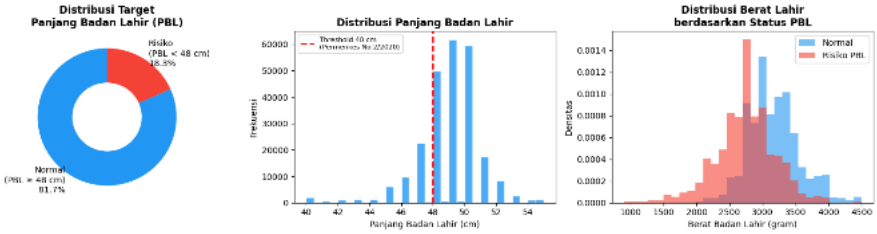

Figure 2. Exploratory data analysis showing the distribution of key categorical features associated with Low Birth Length Risk, including maternal anemia, chronic energy deficiency (CED), drinking water source, and area classification.

3.1.2 Baseline Random Forest

The initial Random Forest framework operated under its standard configuration: n_estimators = 100, max_depth = 10, class_weight = "balanced", and random_state = 42. The imbalance adjustment relied on the "balanced" class weight parameter.

Table 2.Classification Report of the Baseline Random Forest Model

Class	Precision	Recall	F1-Score	Support
Normal (0)	0.90	0.80	0.85	8,166
Low Birth Length Risk (1)	0.41	0.62	0.49	1,834
Overall Accuracy	—	—	0.77	10,000
Macro Average	0.66	0.71	0.67	10,000
Weighted Average	0.81	0.77	0.78	10,000
Matthews Correlation Coefficient (MCC)	—	—	0.37	—

3.1.3 Tuned Random Forest (Hyperparameter Optimization)

The exploration of 12 hyperparameter configurations using a 90:10 train-validation split identified RF-01 as the best-performing configuration with n_estimators = 200, max_depth = 15, max_features = "sqrt", min_samples_leaf = 1, and class_weight = "balanced". The optimal probability threshold was determined by maximizing the F1-score of the Risk class within a threshold range of 0.25–0.75 using increments of 0.02, resulting in an optimal threshold value of 0.50.

Table 3.Classification Report of the Tuned Random Forest Model

Class	Precision	Recall	F1-Score	Support
Normal (0)	0.90	0.83	0.86	8,166
Low Birth Length Risk (1)	0.44	0.58	0.50	1,834
Overall Accuracy	—	—	0.79	10,000
Macro Average	0.67	0.70	0.68	10,000
Weighted Average	0.81	0.79	0.80	10,000
Matthews Correlation Coefficient (MCC)	—	—	0.38	—

e-ISSN 2827-8356
p-ISSN 2827-9999

The improvement in accuracy was primarily driven by a reduction in false positives (1,633 to 1,354), though false negatives increased marginally from 693 to 770.

3.1.4 Tuned XGBoost

The exploration of 12 hyperparameter configurations for XGBoost resulted in the best-performing configuration, namely XGB-04, with the following parameters: learning_rate = 0.05, max_depth = 4, n_estimators = 400, subsample = 0.9, colsample_bytree = 0.9, reg_lambda = 0.5, min_child_weight = 1, and scale_pos_weight adjusted according to the class ratio. The optimal threshold was optimized based on overall accuracy within a range of 0.15–0.85 (step size = 0.01), resulting in a threshold value of 0.79.

Table 4.Classification Report of the Tuned XGBoost Model

Class	Precision	Recall	F1-Score	Support
Normal (0)	0.86	0.97	0.91	8,166
Low Birth Length Risk (1)	0.66	0.27	0.39	1,834
Overall Accuracy	—	—	0.84	10,000
Macro Average	0.76	0.62	0.65	10,000
Weighted Average	0.82	0.84	0.81	10,000
Matthews Correlation Coefficient (MCC)	—	—	0.39	—

3.1.5 Analysis of Two XGBoost Threshold Scenarios

To ensure a fair comparison, we also evaluated both models using a unified optimization criterion. When both models were optimized for the F1-score of the Risk class, XGBoost achieved an F1-Risk of 0.5095, while Random Forest achieved an F1-Risk of 0.50 (the threshold for Random Forest remained at 0.55, as originally optimized for F1-Risk). Similarly, when both were optimized for overall accuracy, XGBoost achieved 84.07% (using its original threshold of 0.79), and Random Forest reached 83.74% at an optimal threshold of 0.75. These supplementary analyses confirm that XGBoost's superior performance is robust regardless of the optimization objective.

Table 5.Analysis of Two XGBoost Threshold Scenarios

Metric	Threshold 0.50	Threshold 0.79
Accuracy	75.4%	84.1%
Precision (Low Birth Length Risk)	0.38	0.66
Recall (Low Birth Length Risk)	0.73	0.27
F1-Score (Low Birth Length Risk)	0.50	0.39
Matthews Correlation Coefficient (MCC)	0.36	0.39

3.1.6 Comparison of All Models

Table 6 summarizes the comparative performance across all three models. Tuned XGBoost consistently surpassed both Random Forest variants, recording 84.07% accuracy, 77.04% ROC-AUC, and an MCC of 0.388.

e-ISSN 2827-8356
p-ISSN 2827-9999

Table 6.Comparison of All Models Including MCC

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC
Baseline Random Forest	76.76%	81.30%	76.76%	78.40%	76.34%	0.37
Tuned Random Forest	78.58%	81.30%	78.58%	79.66%	76.42%	0.38
Tuned XGBoost	84.07%	81.95%	84.07%	81.26%	77.04%	0.39

3.1.7 ROC Curve

Figure 3 presents the ROC curves of the evaluated models. Tuned XGBoost achieved the highest ROC-AUC (0.7704), followed by Tuned Random Forest (0.7642) and Baseline Random Forest (0.7634), indicating that Tuned XGBoost provided the best overall discriminative performance for predicting Low Birth Length Risk.

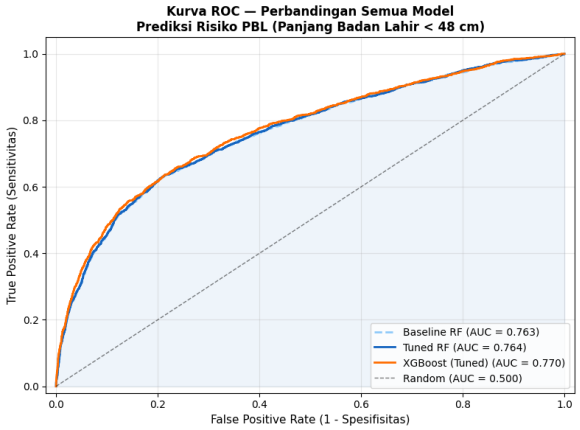

Figure 3. Receiver Operating Characteristic (ROC) curves comparing the discriminative performance of Baseline Random Forest, Tuned Random Forest, and Tuned XGBoost models for Low Birth Length Risk prediction. The area under the curve (AUC) values are reported in the legend.

3.1.8 Model Stability Using Cross-Validation

A 5-fold stratified cross-validation protocol yielded an average accuracy score of 76.34% (±0.73%), indicating stable model performance and no significant overfitting.

3.1.9 Model Interpretability Using SHAP

Table 7 and Figure 4 show that birth weight was the highest-ranking predictor (mean absolute SHAP score of 0.39), succeeded by maternal parity, child age, maternal age, and drinking water source.

Table 7.SHAP Feature Ranking

No.	Feature Code	Feature Name	Mean |SHAP Value|
1	birth_weight	Birth Weight (grams)	0.3900
2	maternal_parity	Maternal Parity (Number of Pregnancies)	0.0200
3	age_months	Child Age (Months)	0.0100
4	maternal_age	Maternal Age During Pregnancy (Years)	0.0100
5	water_source	Drinking Water Source	0.0100

e-ISSN 2827-8356
p-ISSN 2827-9999

6	sex	Child Sex	0.0090
7	area	Area Classification	0.0090
8	diarrhea	History of Diarrhea	0.0080
9	maternal_anemia	Maternal Anemia During Pregnancy (Hb < 11 g/dL)	0.0060
10	maternal_cek	Maternal Chronic Energy Deficiency (MUAC < 23.5 cm)	0.0030
11	pregnancy_interval	Pregnancy Interval	0.0010
12	ari	History of Acute Respiratory Infection (ARI)	0.0010
13	tb_contact	History of Tuberculosis Contact	0.0003
14	pneumonia	History of Pneumonia	0.0000

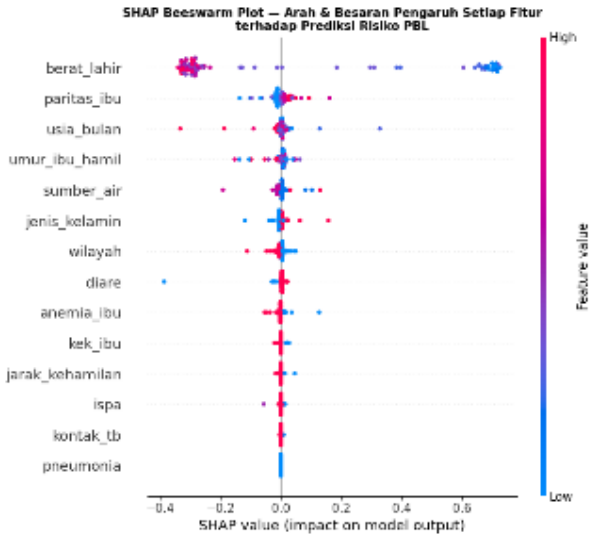

Figure 4.SHAP Summary Plot (Beeswarm) – Global Feature Importance

3.2 Discussion

3.2.1 Performance Improvement Through Optimization

The results demonstrate that hyperparameter tuning and threshold optimization improved predictive performance. Tuned XGBoost achieved the highest accuracy (84.07%), outperforming previous studies that reported accuracies of 61% and 71.4% [3]. The improvement may be attributed to the use of a large national dataset, class-imbalance handling, and systematic parameter optimization.

3.2.2 Clinical Trade-off: Accuracy vs. Sensitivity

Tuned XGBoost demonstrated a peak accuracy of 84% but a restricted risk-class recall of 27%, representing a direct consequence of adjusting the 0.79 threshold for optimal global accuracy. Lowering this operational cutoff to 0.50 effectively recovered the model sensitivity to 73%, though it downscaled the aggregate accuracy to 75.4%. Therefore, threshold selection should be aligned with the intended clinical objective.

e-ISSN 2827-8356
p-ISSN 2827-9999

3.2.3 Risk Factors Based on SHAP Analysis

Neonatal birth weight emerged as the primary predictor (mean |SHAP| = 0.39), consistent with previous studies identifying deficient birth mass as a leading catalyst for growth restriction and growth failure [5], [2]. Maternal factors (parity, anemia, and CED) and environmental factors (water source and residential area) also contributed substantially, supporting previous findings that stunting is influenced by maternal, child, and environmental conditions [8]. Furthermore, SHAP analysis improved model transparency by identifying clinically relevant predictors, consistent with findings reported in healthcare applications [20], [21].

3.2.4 Limitations

Certain constraints apply to this empirical work. Primarily, utilizing the Low Birth Length (LBL) parameter serves as a vulnerability indicator instead of a definitive clinical diagnosis (HAZ < -2 SD), as the SSGI 2024 dataset does not provide actual child height measurements required to calculate Z-scores. Second, stratified sampling of 50,000 records was performed due to RAM limitations; therefore, the use of the full dataset may further improve model robustness. Third, SHAP analysis was conducted on only 100 representative samples. Although [21] stated that a subset can provide adequate global representation, a larger sample size is recommended for future studies.

Fifth, although Stratified 5-Fold Cross-Validation was employed to assess internal stability, external validation using independent datasets (e.g., SSGI data from different years or regional health surveys from other provinces) was not performed due to data availability constraints. Future studies should pursue external validation to confirm generalizability across different populations and temporal periods.

3.2.5 Practical Implications for Public Health Screening

The proposed predictive framework has practical implications for public health screening by enabling earlier identification of infants at risk of Low Birth Length before growth failure becomes clinically apparent. Because the model relies on routinely collected maternal, neonatal, and environmental variables available in national health surveys and primary healthcare records, it can be integrated into existing maternal and child health services without requiring additional laboratory examinations.

For healthcare workers at community health centers (Puskesmas), the model may serve as a decision-support tool to prioritize infants requiring closer nutritional monitoring, growth assessment, and early intervention. At the policy level, local and national health authorities may utilize the prediction results to identify high-risk populations, allocate nutritional resources more efficiently, and support evidence-based planning for stunting prevention programs. Although further external validation is required before real-world deployment, the proposed framework demonstrates the potential of explainable machine learning to support data-driven public health screening and early preventive strategies.

4. CONCLUSION

This study successfully developed a low birth length risk prediction framework using Random Forest and XGBoost based on the 2024 Indonesian Nutritional Status Survey (SSGI) dataset. Hyperparameter optimization and probability threshold optimization improved the predictive performance of the developed models. Random Forest accuracy increased from 76.76% to 78.58%, while the optimized XGBoost model achieved the best overall performance with an accuracy of 84.07%, a ROC-AUC of 77.04%, and an MCC of 0.39.

SHAP analysis identified birth weight as the most influential predictor, followed by maternal parity, toddler age, and maternal age, thereby improving model interpretability. Overall, the findings demonstrate that integrating threshold optimization with explainable machine learning can improve predictive performance while providing transparent decision support for early identification of low birth length risk in public health screening and clinical settings.

e-ISSN 2827-8356
p-ISSN 2827-9999

5. SUGGESTIONS

Future research should prioritize the use of the full SSGI dataset, without stratified sampling, to further enhance model robustness. Additionally, expanding the SHAP analysis beyond the current sample of 100 representative instances would provide deeper insights into feature contributions and model behavior.

External validation remains a critical next step; evaluating the framework on SSGI data from different years (e.g., 2022, 2023) or on comparable datasets from neighboring countries would confirm its applicability across diverse demographic and epidemiological settings. Ultimately, the deployment of this framework into clinical decision support systems is recommended to facilitate real-world early screening for low birth length risk.

6. Acknowledgments

The authors would like to express their sincere gratitude to Universitas Mercu Buana Yogyakarta for providing academic support and a conducive research environment throughout this study. We are deeply grateful to our supervisors for their invaluable guidance, constructive feedback, and continuous encouragement during every stage of this research, from conceptualization to manuscript preparation. We also extend our appreciation to the Indonesian Ministry of Health for granting access to the 2024 Nutritional Status Survey (SSGI) dataset, which served as the primary data source for this analysis. Finally, we thank our colleagues and peers for their insightful discussions and technical assistance during the research process. Their support and contributions have been essential to the successful completion of this work.

7. Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Y. Xiong, X. Hu, J. Cao, L. Shang, and B. Niu, "A predictive model for stunting among children under the age of three," *Front. Pediatr.*, vol. 12, Sep. 2024, doi: 10.3389/fped.2024.1441714.

[2] G. Nduwayezu *et al.*, "Spatial Machine Learning for Exploring the Variability in Low Height-For-Age From Socioeconomic, Agroecological, and Climate Features in the Northern Province of Rwanda," *Geohealth*, vol. 8, no. 9, Sep. 2024, doi: 10.1029/2024GH001027.

[3] A. Ranjbar, F. Montazeri, M. V. Farashah, V. Mehrnoush, F. Darsareh, and N. Roozbeh, "Machine learning-based approach for predicting low birth weight," *BMC Pregnancy Childbirth*, vol. 23, no. 1, p. 803, Nov. 2023, doi: 10.1186/s12884-023-06128-w.

[4] A. Naik, G. G. Tejani, and S. J. Mousavirad, "SGO enhanced random forest and extreme gradient boosting framework for heart disease prediction," *Sci. Rep.*, vol. 15, no. 1, p. 18145, May 2025, doi: 10.1038/s41598-025-02525-7.

[5] C. Kalinda *et al.*, "Leveraging multisectoral approach to understand the determinants of childhood stunting in Rwanda: a systematic review and meta-analysis," *Syst. Rev.*, vol. 13, no. 1, p. 16, Jan. 2024, doi: 10.1186/s13643-023-02438-4.

[6] Z. Fan *et al.*, "XGBoost-SHAP-based interpretable diagnostic framework for knee osteoarthritis: a population-based retrospective cohort study," *Arthritis Res. Ther.*, vol. 26, no. 1, Dec. 2024, doi: 10.1186/s13075-024-03450-2.

[7] H. Shen, H. Zhao, and Y. Jiang, "Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea," *Children*, vol. 10, no. 10, p. 1638, Sep. 2023, doi: 10.3390/children10101638.

e-ISSN 2827-8356
p-ISSN 2827-9999

[8] T. K. Wudu, A. A. Endalew, and A. A. Dires, "Explainable hybrid machine learning model for predicting stunting and identifying key risk factors among Ethiopian children under five," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-46417-w.

[9] J. Munyemana, I. H. Kabano, B. Uzayisenga, A. R. Cyamweshi, E. Ndagijimana, and E. Kubana, "The role of national nutrition programs on stunting reduction in Rwanda using machine learning classifiers: a retrospective study," *BMC Nutr.*, vol. 10, no. 1, p. 98, Jul. 2024, doi: 10.1186/s40795-024-00903-4.

[10] N. El Furqany, M. Subianto, and A. Rusyana, "Hybrid Ensemble Learning with SMOTEENN and Soft Voting for Stunting Risk Prediction: A SHAP-Based Explainable Approach," *Journal of Applied Data Sciences*, vol. 6, no. 4, pp. 2989–3004, Dec. 2025, doi: 10.47738/jads.v6i4.829.

[11] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *J. Big Data*, vol. 10, no. 1, p. 58, May 2023, doi: 10.1186/s40537-023-00738-z.

[12] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," *Brain Inform.*, vol. 11, no. 1, p. 10, Dec. 2024, doi: 10.1186/s40708-024-00222-1.

[13] B. Paneru, B. Paneru, S. C. Sapkota, and R. Poudyal, "Enhancing healthcare with AI: Sustainable AI and IoT-Powered ecosystem for patient aid and interpretability analysis using SHAP," *Measurement: Sensors*, vol. 36, p. 101305, Dec. 2024, doi: 10.1016/j.measen.2024.101305.

[14] L. Malinverno *et al.*, "A historical perspective of biomedical explainable AI research," *Patterns*, vol. 4, no. 9, p. 100830, Sep. 2023, doi: 10.1016/j.patter.2023.100830.

[15] G. M. Setegn and B. E. Dejene, "Improving machine learning models through explainable AI for predicting the level of dietary diversity among Ethiopian preschool children," *Ital. J. Pediatr.*, vol. 51, no. 1, p. 91, Mar. 2025, doi: 10.1186/s13052-025-01892-1.

[16] M. Sinaga, F. Fujiati, and D. Halawa, "Designing a Stunting Prediction Model Using Machine Learning to Support SDGs Achievement in Indonesia," *sinkron*, vol. 9, no. 4, pp. 2070–2079, Oct. 2025, doi: 10.33395/sinkron.v9i4.15296.

[17] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[18] T. O. Omotehinwa, D. O. Oyewola, and E. G. Dada, "A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis," *Healthcare Analytics*, vol. 4, p. 100218, Dec. 2023, doi: 10.1016/j.health.2023.100218.

[19] Md. J. Raihan, Md. A.-M. Khan, S.-H. Kee, and A.-A. Nahid, "Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 6263, Apr. 2023, doi: 10.1038/s41598-023-33525-0.

[20] Md. Z. Islam, M. R. K. Chowdhury, M. Kader, B. Billah, Md. S. Islam, and M. Rashid, "Determinants of low birth weight and its effect on childhood health and nutritional outcomes in Bangladesh," *J. Health Popul. Nutr.*, vol. 43, no. 1, p. 64, May 2024, doi: 10.1186/s41043-024-00565-9.

[21] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.