Vol. 8, No. 2, April 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Ensemble Learning for Pediatric Stunting Detection: A Comparative Study of XGBoost, Random Forest, and LightGBM with Oversampling Techniques

Tri Sugihartono[1], Djoko Soetarno[2], Rahmat Sulaiman[3], Sarwindah[4], Marini[5], Fitriyani[6]

[1,3]Informatics Department, Information Technology Faculty, ISB Atma Luhur, Pangkalpinang, Indonesia

[2]Informatics Department, Binus University, Jakarta, Indonesia

[4]Business Digital Department, Economic Business Faculty, ISB Atma Luhur, Pangkalpinang, Indonesia

[5,6]Information System Department, Information Technology Faculty, ISB Atma Luhur, Pangkalpinang, Indonesia

Received:
September 12, 2025
Revised:
March 16, 2026
Accepted:
March 31 2026
Published:
April 12, 2026

Corresponding Author:
Author Name*:
Tri Sugihartono
Email*:
trisugihartono@atmaluhur.ac.id

DOI:
10.63158/journalisi.v8i2.1568

Abstract. Stunting, driven by chronic childhood malnutrition, remains a critical global public health concern. Early detection is persistently challenged by class imbalance in pediatric health datasets and the absence of systematic comparisons between oversampling strategies and ensemble classifiers. This study develops and evaluates an ensemble learning pipeline for stunting detection, benchmarking XGBoost, Random Forest, and LightGBM across five oversampling configurations — Original, SMOTE, ADASYN, Borderline-SMOTE, and SMOTE-ENN — using 10,000 pediatric health records from posyandu activities in Bangka Belitung Province, Indonesia. Seven anthropometric and demographic features were utilized, with stratified 80:20 train-test splitting and five-fold cross-validation. XGBoost with original imbalanced data achieved the highest Recall (0.9573) and a competitive F1-Score (0.9158), while LightGBM with SMOTE delivered the strongest balanced performance (F1-Score: 0.9160, ROC-AUC: 0.8431). SMOTE-ENN consistently underperformed across all classifiers. To our knowledge, this is the first study to simultaneously compare five oversampling strategies across three ensemble models within a unified framework, offering a foundation for high-sensitivity stunting surveillance in resource-constrained healthcare settings.

Keywords: Stunting Detection; Ensemble Learning; Imbalanced Classification; Oversampling; SMOTE

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Stunting — defined as low height-for-age below minus two standard deviations of the WHO Child Growth Standards median — is the most prevalent form of chronic childhood malnutrition globally. According to UNICEF, WHO, and the World Bank, approximately 148.1 million children under five were affected in 2022, representing 22.3% of the global under-five population [1]. Beyond its physical manifestations, stunting causes irreversible impairments in cognitive development, economic productivity, and lifelong health [2]. In Indonesia, the national stunting prevalence remained above 21% per the 2022 Indonesian Nutritional Status Survey (SSGI), obstructing sustainable development goals [3]. Early detection is therefore a national health priority, as the first 1,000 days of life represent the critical intervention window [4].

Machine learning has emerged as a promising tool for automated pediatric malnutrition screening, enabling faster and more scalable risk identification from routinely collected health records than traditional periodic anthropometric assessments [5], [6]. Studies applying single-algorithm approaches have established useful benchmarks: Novalina et al. [7] benchmarked multiple ML algorithms for stunting risk prediction in Indonesia; Ndagijimana et al. [8] applied ensemble methods for stunting prediction in Rwanda achieving strong discriminative performance; and Dewi et al. [9] demonstrated geographically weighted Random Forest for regional stunting analysis in East Java. These studies confirm the viability of ensemble models but do not systematically evaluate the interaction between classifier choice and oversampling strategy.

A second stream of research focuses specifically on class imbalance in malnutrition datasets. Pramana et al. [10] demonstrated early stunting detection using SMOTE integrated with ensemble learning, achieving competitive F1-Scores. Sugihartono et al. [11] systematically compared XGBoost, Random Forest, SVM, and k-NN optimized with SMOTE for stunting detection. Hamid and Subhiyakto [12] further evaluated Random Forest, SVM, and XGBoost with SMOTE, confirming SMOTE's utility. While these studies apply individual oversampling techniques, none provides a head-to-head comparison of multiple oversampling variants — SMOTE [13], Borderline-SMOTE [14], ADASYN [15], and SMOTE-ENN [16] — against each other across multiple ensemble classifiers within a single unified experimental framework.

This cross-cutting gap — no study jointly varies both classifier and oversampling choice within a controlled pipeline — limits evidence-based model selection for clinical stunting screening. To address this, the present study benchmarks XGBoost, Random Forest, and LightGBM across five oversampling configurations (Original, SMOTE, ADASYN, Borderline-SMOTE, SMOTE-ENN) on a compiled regional pediatric health records dataset (n = 10,000). Three objectives are pursued: (1) identify the optimal model-sampling combination maximizing F1-Score and Recall; (2) characterize model-specific sensitivity to oversampling strategy; and (3) provide actionable configuration recommendations for high-sensitivity stunting surveillance systems.

2. METHODS

This study follows a systematic experimental pipeline comprising five sequential phases: (1) Dataset Acquisition — retrospective aggregation and curation of posyandu health records; (2) Data Preprocessing — encoding, train-test splitting (80:20 stratified), and feature standardization; (3) Oversampling Application — five configurations applied exclusively to the training set; (4) Model Training and Evaluation — 15 model-sampling combinations (3 classifiers × 5 strategies) trained with 5-fold cross-validation and evaluated on the held-out test set; and (5) Comparative Performance Analysis — ranking by Recall, F1-Score, Precision, Accuracy, and ROC-AUC, as shown in Figure 1. Each phase is detailed in the subsections as follow.

Figure 1. Research Workflow

2.1. Dataset Description

The dataset used in this study is a compiled regional pediatric health records dataset derived from child health monitoring activities (posyandu) in Bangka Belitung Province, Indonesia, aggregated and curated for this research. Records were collected retrospectively from monthly posyandu measurement activities spanning 2020–2023, with data aggregation conducted under institutional coordination of ISB Atma Luhur. The

dataset represents a curated analytic sample: records with missing anthropometric values were excluded prior to analysis, and no personally identifiable information is retained in the analytical dataset. It comprises 10,000 records and seven features: Gender (categorical), Age (months), Birth Weight (kg), Birth Length (cm), Body Weight at measurement (kg), Body Length at measurement (cm), and Breastfeeding status (binary). The target variable is a binary stunting label (Stunted = 1, Non-Stunted = 0) derived from WHO height-for-age z-score criteria. Dataset characteristics are summarized in Table 1.

TABLE 1. Dataset Characteristics

Attribute	Description
Total Records	10,000
Total Features	7 (Gender, Age, Birth Weight, Birth Length, Body Weight, Body Length, Breastfeeding)
Target Classes	Binary (Stunted = 1 / Non-Stunted = 0)
Stunted Cases (Class 1)	7,955 (79.55%)
Non-Stunted Cases (Class 0)	2,045 (20.45%)
Missing Values	None
Class Imbalance Ratio	≈ 3.89 : 1

The dataset exhibits a pronounced class imbalance (ratio ≈ 3.89:1), which reflects the high prevalence of stunting in the source population — consistent with Bangka Belitung Province reporting some of the highest provincial stunting rates in Indonesia [3]. This imbalance is therefore a realistic characteristic of the target screening population rather than an artifact of sampling, and motivates the systematic evaluation of oversampling strategies to prevent classifier bias toward the majority class [17].

2.2. Data Preprocessing

Data preprocessing was conducted in four sequential stages. First, Categorical Encoding was applied using scikit-learn's LabelEncoder to transform the Gender feature into a numerical representation [18]. Second, Feature-Target Separation was performed to isolate the target variable prior to splitting, preventing target leakage. Third, Stratified Train-Test Splitting partitioned the dataset into 80% training (n = 8,000) and 20% test (n = 2,000) subsets using stratified random sampling with random seed 42 to preserve class distributions [19], as detailed in Table II. Fourth, Feature Standardization was applied using

StandardScaler fitted exclusively on the training set and applied to both subsets, strictly preventing data leakage [20]. Although tree-based ensemble models (Random Forest, LightGBM) do not inherently require feature scaling, StandardScaler was uniformly applied to all models to ensure consistency of input scale across the comparative pipeline and to support the distance-based components of SMOTE and ADASYN oversampling, which depend on Euclidean nearest-neighbor distances. All oversampling was applied exclusively to the training set after splitting and scaling; the test set was never resampled to ensure unbiased evaluation on the original data distribution.

Table 2. Train-Test Split Distribution

Subset	Total	Stunted (Class 1)	Non-Stunted (Class 0)
Training Set	8,000	6,364 (79.55%)	1,636 (20.45%)
Test Set	2,000	1,591 (79.55%)	409 (20.45%)
Total	10,000	7,955	2,045

Feature standardization follows Equation (1), where x is the original feature value, μ is the training set mean, and σ is the standard deviation, as shown in Equation 1.

$$z = (x - \mu) / \sigma \quad (1)$$

2.3. Oversampling Strategy Application

Five oversampling configurations were applied solely to the standardized training set, as summarized in Table 3. (1) Original — no resampling, serves as baseline. (2) SMOTE [13] generates synthetic minority samples by linear interpolation between a minority instance and one of its k nearest minority neighbors per Equation (2). (3) ADASYN [15] adaptively generates more synthetic samples in harder-to-learn boundary regions proportional to local majority-class density. (4) Borderline-SMOTE [14] restricts synthesis to minority instances near the decision boundary, directly targeting the most informative classification region. (5) SMOTE-ENN [16] combines SMOTE oversampling with Edited Nearest Neighbors cleaning to remove potentially noisy samples from both classes after synthesis.

$$x\_syn = x\_i + \lambda \cdot (x\_zi - x\_i), \quad \lambda \in [0, 1] \quad (2)$$

TABLE 3. Oversampling Strategy Results on Training Set

Strategy	Non-Stunted (Class 0)	Stunted (Class 1)	Total Samples	Samples Added
Original	1,636	6,364	8,000	0 (Baseline)
SMOTE	6,364	6,364	12,728	+4,728
ADASYN	6,458	6,364	12,822	+4,822
Borderline-SMOTE	6,364	6,364	12,728	+4,728
SMOTE-ENN	4,316	4,076	8,392	+392 (after cleaning)

2.4. Ensemble Classification Models

Three state-of-the-art ensemble classifiers were implemented with hyperparameters selected through a systematic grid-search procedure, then fixed for all 15 model-sampling combination experiments to ensure fair comparison, as detailed in Table 4. Grid search used 5-fold cross-validation on the training set with macro F1-Score as the optimization criterion. For XGBoost, the search space covered n_estimators ∈ {100, 200}, learning_rate ∈ {0.01, 0.05, 0.1}, and max_depth ∈ {4, 6, 8}. For Random Forest, n_estimators ∈ {100, 200} and max_depth ∈ {10, 15, None} were evaluated. For LightGBM, num_leaves ∈ {31, 63} and learning_rate ∈ {0.01, 0.05, 0.1} were searched. Final selected configurations are reported in Table 4.

XGBoost [20] minimizes the regularized objective in Equation (3), where l is a convex loss function and $\Omega(f_k)$ penalizes tree complexity. The learning rate (0.05) and max depth (6) were selected to balance generalization and computational cost; scale_pos_weight was set proportionally to the class frequency ratio to implicitly handle class imbalance, as shown in Equation 3.

$$L(\varphi) = \Sigma_i \, l(\hat{y}_i, y_i) + \Sigma_k \, \Omega(f_k) \quad (3)$$

Random Forest [21] constructs decorrelated decision trees through bootstrap aggregation. n_estimators = 200 and max_depth = 15 were configured to ensure sufficient model capacity, with class_weight = 'balanced' applying Equation (4) to inversely

scale sample weights by class frequency, providing inherent imbalance correction, as shown in Equation 4.

$$w_j = n / (k \cdot n_j) \quad (4)$$

LightGBM [22] employs leaf-wise tree growth with Gradient-based One-Side Sampling (GOSS), which retains instances with large gradients and randomly down-samples small-gradient instances, substantially reducing training time while preserving information. num_leaves = 31 and learning rate = 0.05 were selected following the LightGBM documentation guidance for medium-sized datasets, with class_weight = 'balanced' applied for imbalance handling.

TABLE 4. Ensemble Model Hyperparameter Configurations

Parameter	XGBoost	Random Forest	LightGBM
n_estimators	200	200	200
learning_rate	0.05	—	0.05
max_depth	6	15	8
subsample	0.8	—	0.8
colsample_bytree	0.8	—	0.8
num_leaves	—	—	31
max_features	—	sqrt	—
min_samples_split	—	5	—
class_weight	scale_pos_weight	balanced	balanced
random_state	42	42	42

2.5. Model Evaluation Metrics

Models were evaluated using five complementary metrics, computed on the held-out test set. Recall (as shown in Equation 7) and F1-Score (as shown in Equation 8) are the primary metrics, reflecting the clinical priority of minimizing false negatives — missed stunting cases carry greater consequence than false positive referrals in pediatric screening [17]. Accuracy (as shown in Equation 5), Precision (as shown in Equation 6), and ROC-AUC (as shown in Equation 7) provide complementary perspectives.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

$$\text{Precision} = TP / (TP + FP) \quad (6)$$

$$\text{Recall} = TP / (TP + FN) \quad (7)$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \quad (9)$$

Additionally, five-fold cross-validation F1-Score was computed on the training set for each combination to assess generalization stability and detect potential overfitting [23].

2.6. Experimental Design

The experimental framework evaluated all 15 model-sampling combinations (3 × 5) in a fully factorial design. All 15 combinations were trained and evaluated on identical train-test partitions, ensuring valid cross-combination comparison. Implementation used Python 3.10 with scikit-learn 1.3.0, XGBoost 1.7.6, LightGBM 3.3.5, imbalanced-learn 0.11.0, and Pedregosa et al.'s scikit-learn framework [24]. A fixed random seed of 42 was applied to all stochastic components — splitting, oversampling, and model training — for full reproducibility [23].

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation

Table 5 presents the comprehensive performance metrics for all 15 model-sampling combinations evaluated on the independent test set. Figure 2 provides a visual overview via F1-Score heatmap, revealing that LightGBM is the most sensitive model to sampling strategy choice, while XGBoost maintains consistently strong performance regardless of oversampling.

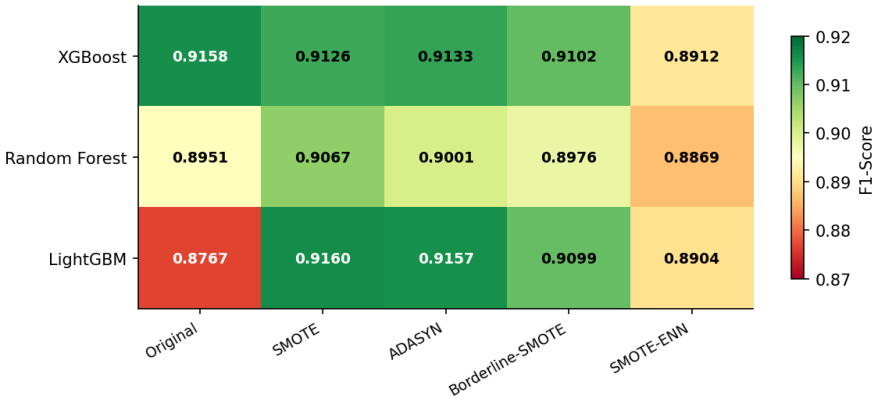

Figure 2. F1-Score Heatmap: Ensemble Model × Sampling Strategy Performance Matrix

Table 5. Comparative Performance of All 15 Model-Sampling Combinations

Model	Sampling	Accuracy	Precision	Recall	F1-Score	ROC-AUC
XGBoost	Original	0.8600	0.8778	0.9573	0.9158	0.8435
XGBoost	SMOTE	0.8570	0.8877	0.9390	0.9126	0.8424
XGBoost	ADASYN	0.8590	0.8940	0.9334	0.9133	0.8468
XGBoost	Borderline-SMOTE	0.8545	0.8939	0.9271	0.9102	0.8429
XGBoost	SMOTE-ENN	0.8295	0.9053	0.8774	0.8912	0.8399
Random Forest	Original	0.8325	0.8920	0.8982	0.8951	0.8293
Random Forest	SMOTE	0.8485	0.8884	0.9258	0.9067	0.8307
Random Forest	ADASYN	0.8400	0.8945	0.9057	0.9001	0.8313
Random Forest	Borderline-SMOTE	0.8370	0.8970	0.8982	0.8976	0.8284
Random Forest	SMOTE-ENN	0.8235	0.9046	0.8699	0.8869	0.8309
LightGBM	Original	0.8120	0.9164	0.8404	0.8767	0.8395
LightGBM	SMOTE	0.8620	0.8879	0.9459	0.9160	0.8431
LightGBM	ADASYN	0.8615	0.8878	0.9453	0.9157	0.8430
LightGBM	Borderline-SMOTE	0.8530	0.8876	0.9334	0.9099	0.8431
LightGBM	SMOTE-ENN	0.8275	0.9004	0.8806	0.8904	0.8343

Note: Highlighted cells indicate top-performing values per metric. Best F1: LightGBM+SMOTE (0.9160). Best Recall: XGBoost+Original (0.9573).

Figure 3 shows the Recall comparison across all configurations. XGBoost with Original data achieves the highest Recall of 0.9573. This can be attributed to XGBoost's scale_pos_weight parameter, which implicitly penalizes false negatives by assigning higher misclassification cost to the minority class during gradient computation. Because this mechanism already incorporates class-imbalance correction at the loss level, introducing additional synthetic samples through oversampling adds redundant and potentially noisy training signals, which explains the monotonic decline in XGBoost's Recall under all oversampling strategies. In contrast, LightGBM with class_weight='balanced' achieves a Recall of only 0.8404 on original data but improves dramatically to 0.9459 with SMOTE (+12.6%), suggesting that LightGBM's GOSS sampling

mechanism benefits from the boundary clarification provided by synthetic minority samples.

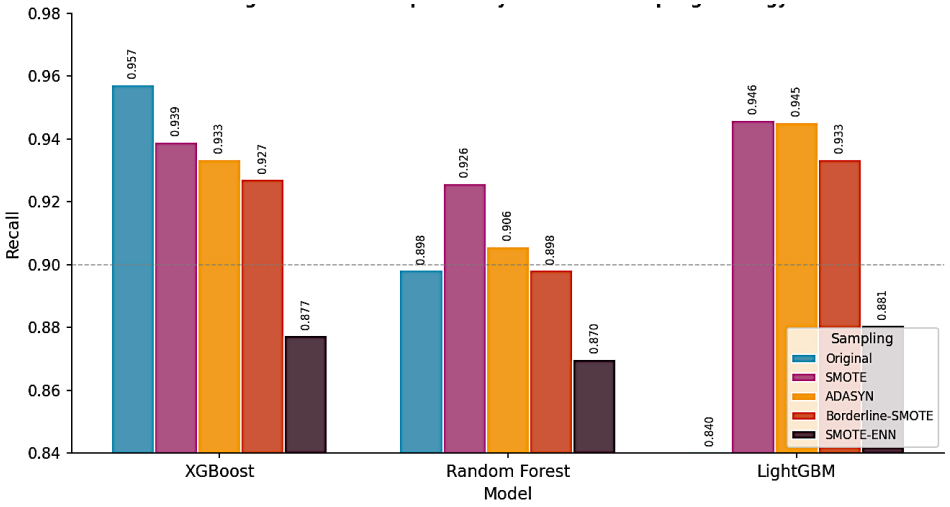

Figure 3. Recall Comparison by Model and Sampling Strategy

Figure 4 presents the ROC-AUC comparison. ROC-AUC values are consistent across strategies (range: 0.8284–0.8468), indicating that discriminative ability at the population level is relatively insensitive to oversampling choice. ADASYN achieves the highest individual ROC-AUC for XGBoost (0.8468), consistent with its adaptive boundary-focused generation creating a cleaner decision surface. The narrow ROC-AUC range across all configurations suggests that the primary differentiator between strategies is their impact on the Recall–Precision trade-off rather than overall discriminative capacity.

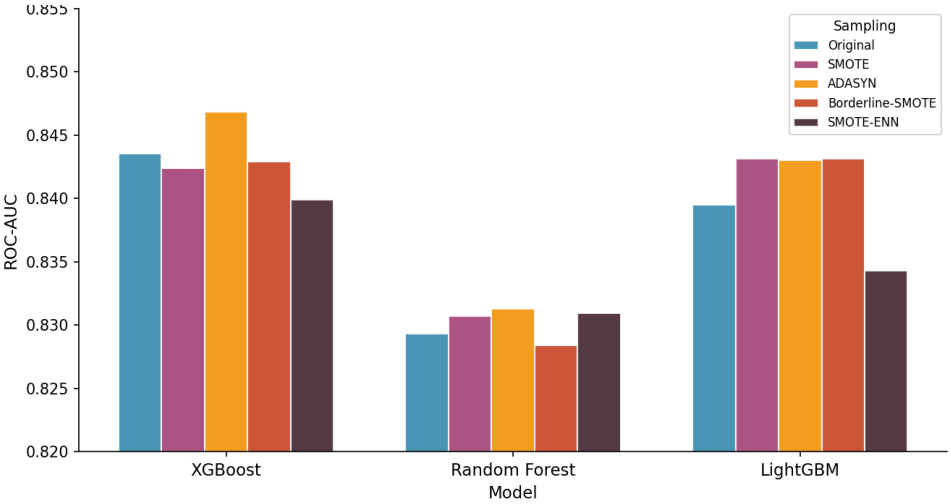

Figure 4. ROC-AUC Comparison by Model and Sampling Strategy

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

3.1.1. F1-Score Trends Across Sampling Strategies

Figure 5 reveals model-specific sensitivity patterns to oversampling strategies. XGBoost exhibits a monotonically declining F1-Score from Original (0.9158) to SMOTE-ENN (0.8912), confirming that its built-in class weighting renders external oversampling counterproductive. Random Forest shows modest but consistent improvement with SMOTE (+1.3% F1) and ADASYN (+0.6% F1), indicating that its balanced weighting alone provides incomplete boundary correction, and limited synthetic samples at boundaries help the model generalize. LightGBM demonstrates the most dramatic sensitivity: its F1-Score improves by 4.5% from Original (0.8767) to SMOTE (0.9160), but then declines with Borderline-SMOTE (0.9099) and SMOTE-ENN (0.8904). This pattern suggests that LightGBM's GOSS mechanism, which down-samples low-gradient instances, may create boundary gaps that global oversampling (SMOTE) compensates for, but that focused boundary oversampling creates conflicting gradients that interfere with GOSS's selection process. These model-specific behaviors have not been previously reported in the stunting detection literature, constituting a novel empirical finding of this study. Compared with Sugihartono et al. [11] who applied SMOTE to a single set of classifiers, this study demonstrates that sampling strategy effects are model-dependent and cannot be generalized without multi-model evaluation.

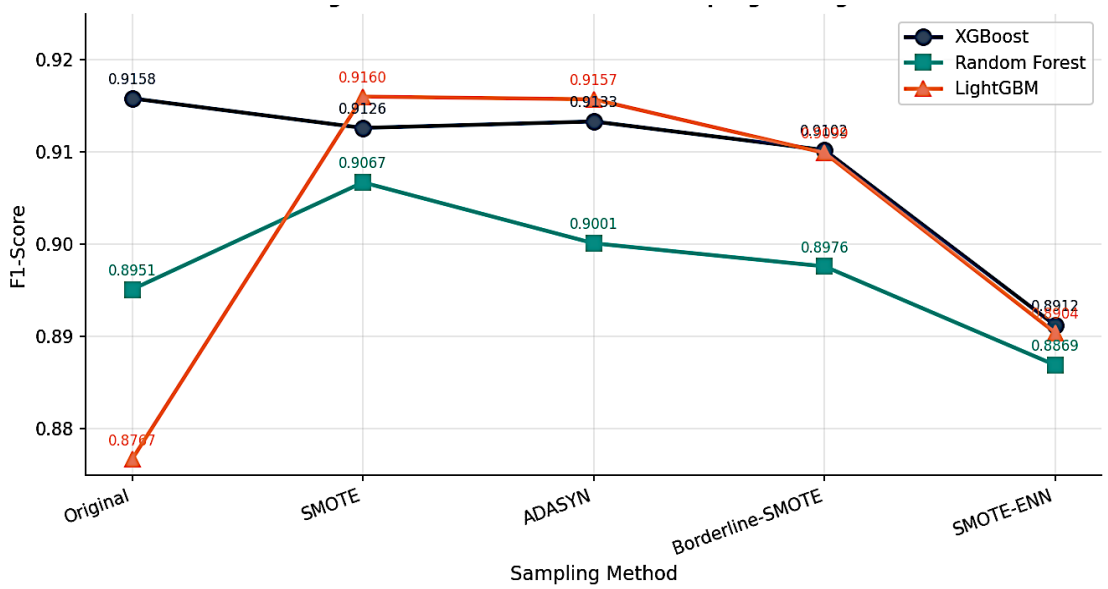

Figure 5. F1-Score Trends Across Sampling Strategies per Model

3.1.2. Sampling Strategy Effectiveness

Table 6 and Figure 6 summarize mean performance across all three models for each sampling strategy. SMOTE achieves the highest mean F1-Score (0.9118) and Recall (0.9369), confirming its general-purpose effectiveness. ADASYN and Borderline-SMOTE produce similar but slightly lower mean performance, suggesting that adaptive and boundary-focused variants do not consistently improve over standard SMOTE for this dataset. SMOTE-ENN consistently underperforms, achieving the lowest mean F1-Score (0.8895) and Recall (0.8760) across all models. This is likely because the ENN cleaning step removes a disproportionate number of borderline minority samples that are informationally valuable for boundary learning, rather than true noise, in this dataset with high within-class anthropometric variability near the stunting threshold. The Original (no sampling) condition achieves the highest Precision (0.8954) but substantially lower Recall than SMOTE conditions, confirming the expected precision-recall trade-off when imbalance is uncorrected.

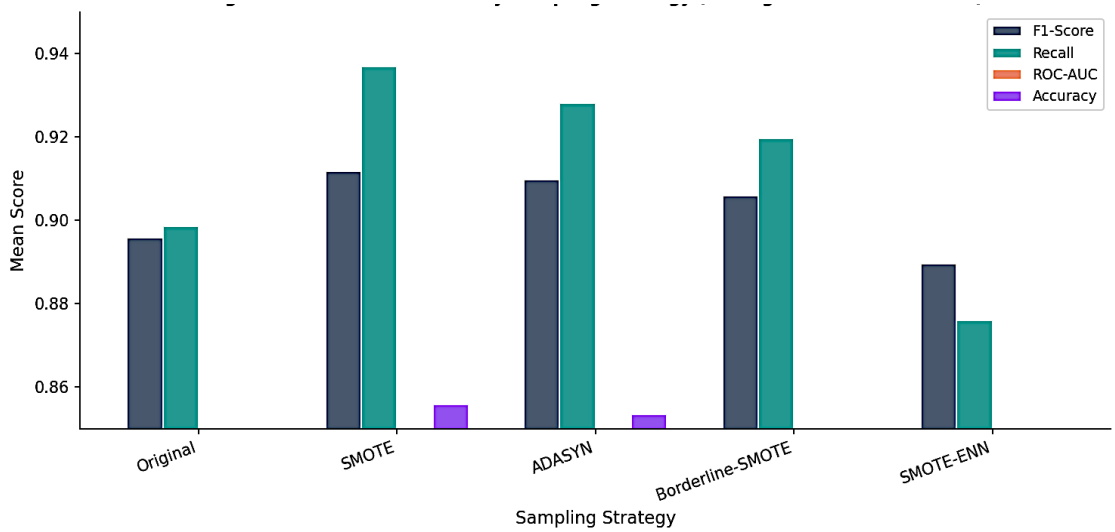

Figure 6. Mean Performance by Sampling Strategy Averaged Across All Models

Table 6. Mean Performance by Sampling Strategy (Averaged Across All Models)

Sampling Method	Accuracy	F1-Score	Recall	Precision	ROC-AUC
SMOTE	0.8558	0.9118	0.9369	0.8880	0.8388
ADASYN	0.8535	0.9097	0.9281	0.8921	0.8404
Borderline-SMOTE	0.8482	0.9059	0.9195	0.8929	0.8382

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Sampling Method	Accuracy	F1-Score	Recall	Precision	ROC-AUC
Original	0.8348	0.8959	0.8986	0.8954	0.8375
SMOTE-ENN	0.8268	0.8895	0.8760	0.9034	0.8350

3.1.3. Confusion Matrix Analysis

Figure 7 shows confusion matrices for the best configuration of each model. XGBoost with Original data achieves only 68 False Negatives out of 1,591 stunted cases (4.3% miss rate), the lowest across all configurations. LightGBM with SMOTE achieves 86 False Negatives (5.4% miss rate) with higher overall accuracy (0.862), representing the best balance for deployment scenarios where both sensitivity and specificity are prioritized. Random Forest with SMOTE achieves 118 False Negatives (7.4% miss rate), the highest among the three best configurations, though its False Positive rate is the lowest. In clinical stunting screening, the asymmetric cost of missed cases (false negatives) typically outweighs the cost of unnecessary follow-up (false positives), strongly favoring configurations that maximize Recall even at a moderate Precision cost. This consideration recommends XGBoost Original for high-sensitivity primary screening and LightGBM+SMOTE for balanced resource-constrained deployment. Compared to Pramana et al. [10], who reported competitive F1-Scores with SMOTE+ensemble, this study provides model-specific confusion analysis that better informs deployment trade-offs.

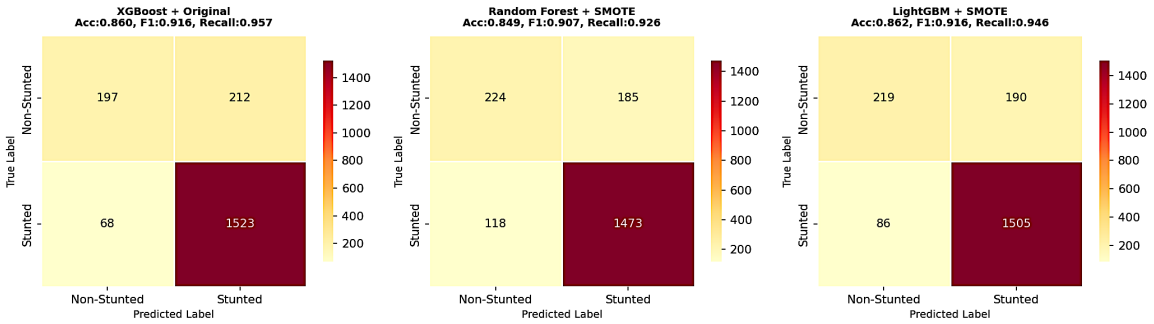

Figure 7. Confusion Matrices for Best-Performing Configuration per Model

3.1.4. Precision-Recall Trade-off

Figure 8 illustrates the precision-recall trade-off across all 15 configurations. Configurations achieving Recall above 0.93 generally show Precision values of 0.877–0.894, while SMOTE-ENN configurations cluster in the lower-Recall, higher-Precision region (Recall: 0.87–0.88, Precision: 0.90–0.91). This trade-off is clinically relevant: in high-

throughput community screening where healthcare worker capacity for follow-up is limited, a moderate-Recall, high-Precision configuration may be operationally preferred. Conversely, in rural settings with minimal referral costs, maximizing Recall is paramount to ensure no stunted child is missed. This analysis highlights that model-sampling selection should be guided by deployment context rather than a single aggregate metric.

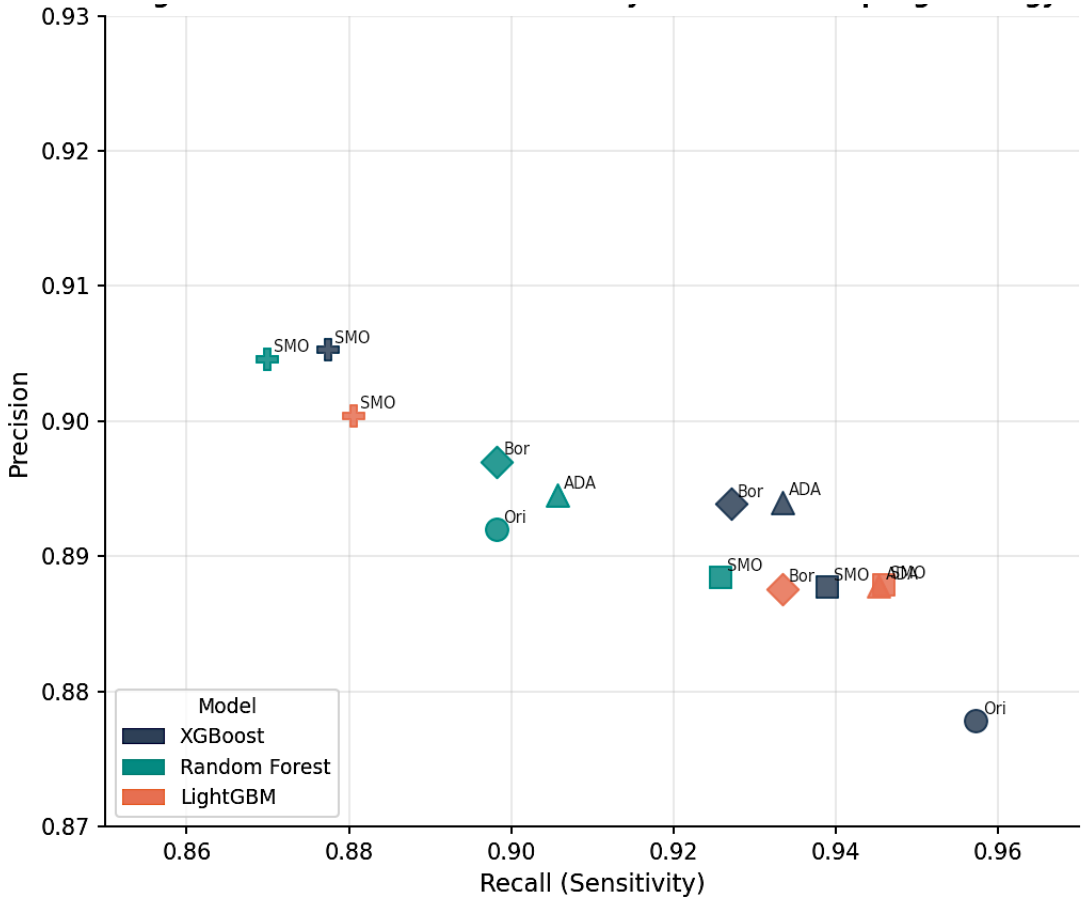

Figure 8. Precision–Recall Trade-off Across All Model-Sampling Combinations

3.2. Discussion

The results demonstrate that model–sampling interaction, rather than sampling strategy alone, determines screening performance in this study. Across the 15 evaluated configurations, two patterns are especially clear. First, XGBoost with the original imbalanced data delivered the highest Recall (0.9573), indicating the strongest ability to minimize missed stunting cases. Second, LightGBM with SMOTE achieved the best overall F1-Score (0.9160), reflecting the most balanced compromise between sensitivity and precision. This distinction is visible across Table 5 and Figures 2, 3, and 5, where XGBoost

remains comparatively stable across sampling strategies, while LightGBM shows marked sensitivity to oversampling. These findings suggest that no single resampling method can be assumed optimal across ensemble learners; instead, performance depends on how the sampling distribution interacts with the model's internal handling of class imbalance and boundary structure.

A key finding is that XGBoost did not benefit from external oversampling, and in fact showed a gradual decline in performance when SMOTE-based methods were introduced. As shown in Table 5 and Figure 4, XGBoost Recall decreased from 0.9573 under the original distribution to 0.9390 with SMOTE, 0.9334 with ADASYN, 0.9271 with Borderline-SMOTE, and 0.8774 with SMOTE-ENN. The same declining pattern appears in Figure 5 for F1-Score, where XGBoost fell from 0.9158 in the original condition to 0.8912 under SMOTE-ENN. This result strongly suggests that XGBoost's built-in imbalance correction through scale_pos_weight was already sufficient to bias learning toward the minority class, making synthetic oversampling redundant and, in some cases, harmful. Additional synthetic minority instances may have introduced marginal noise or distorted the naturally informative class structure, especially near already learnable decision regions. In practical terms, this makes XGBoost Original the most appropriate configuration when the main objective is maximum sensitivity, particularly in screening contexts where missing a stunted child is more harmful than conducting an unnecessary follow-up. Relative to prior work, the observed Recall of 0.9573 exceeds the 0.926 Recall reported by Pramana et al. [10], the 0.86–0.89 Recall range reported by Sugihartono et al. [11], and also compares favorably with the 0.872 F1-Score reported by Ndagijimana et al. [8].

By contrast, LightGBM benefited substantially from oversampling, especially from standard SMOTE. On the original data, LightGBM achieved an F1-Score of 0.8767 and Recall of 0.8404, the weakest Recall among the three original-data models. After SMOTE, however, its F1-Score increased to 0.9160 and its Recall rose to 0.9459, representing improvements of approximately 4.5% and 12.6%, respectively. This improvement is clearly visible in Figure 5, where LightGBM shows the sharpest positive response to oversampling, and in Figure 2, where the heatmap indicates that LightGBM is the most sampling-sensitive model. A plausible explanation is that LightGBM's GOSS mechanism may underrepresent low-gradient minority instances under the original class distribution, creating less stable decision boundaries for stunting cases. Standard SMOTE appears to

compensate for this by improving minority representation more evenly across the feature space. At the same time, the results show that more aggressive or selective variants such as Borderline-SMOTE and SMOTE-ENN were less effective than standard SMOTE, suggesting that boundary-focused synthetic generation or post-generation cleaning may interfere with LightGBM's gradient-based selection behavior. Thus, for settings that require a better balance between missed cases and follow-up burden, LightGBM+SMOTE emerges as a strong deployment candidate.

The performance of Random Forest occupies a middle position between these two patterns. As shown in Table 5, Random Forest improved modestly under SMOTE, with F1-Score increasing from 0.8951 to 0.9067 and Recall from 0.8982 to 0.9258. ADASYN also produced a smaller benefit, whereas Borderline-SMOTE yielded little improvement and SMOTE-ENN reduced both Recall and F1-Score. This indicates that Random Forest can benefit from oversampling, but the gains are more limited than those observed for LightGBM. Unlike XGBoost, Random Forest did not show strong resilience without resampling; unlike LightGBM, it did not show dramatic sensitivity to the choice of oversampling method. This intermediate behavior suggests that Random Forest's ensemble voting benefits from a somewhat more balanced class representation, but its boundary learning is less dependent on synthetic minority enhancement than LightGBM. From a practical standpoint, Random Forest with SMOTE may still be useful where interpretive stability and acceptable performance are preferred, but it is not the strongest option when either maximum Recall or maximum F1-Score is prioritized.

Another important result is the consistent underperformance of SMOTE-ENN across all three models. As summarized in Table 6 and illustrated in Figure 5, SMOTE-ENN produced the lowest mean F1-Score (0.8895) and lowest mean Recall (0.8760) among all sampling strategies, despite yielding the highest mean Precision (0.9034). This same pattern appears at the individual model level in Table 5, where all SMOTE-ENN configurations occupy the lower-Recall, relatively higher-Precision region also shown in Figure 7. These results suggest that the ENN cleaning step may have removed minority instances near the class boundary that were not true noise, but rather informative borderline samples necessary for distinguishing mildly stunted from non-stunted cases. In anthropometric datasets, where children close to the stunting threshold may exhibit legitimate within-class variability, aggressive neighborhood cleaning can erase clinically relevant variation.

The present findings therefore challenge the assumption that hybrid resampling methods automatically improve imbalance learning. In this dataset, cleaning after oversampling appears to reduce sensitivity, which is undesirable for early screening.

The ROC-AUC analysis adds an important nuance to the interpretation of these results. As shown in Figure 4, ROC-AUC values remained within a relatively narrow range (0.8284–0.8468) across all configurations, even when Recall and F1-Score varied substantially. This indicates that most models retained broadly similar ranking ability at the population level, regardless of sampling method. The main impact of resampling was therefore not on general separability, but on the operating trade-off between precision and recall. This is also evident in Figure 8, where the configurations spread along different points of the precision–recall space rather than separating sharply in overall discrimination. In other words, oversampling changed how aggressively the models identified positive cases, not whether the underlying signal existed. This is a practically important result, because it suggests that model selection for stunting screening should not rely on ROC-AUC alone. In a clinically imbalanced context, Recall, F1-Score, and confusion-matrix behavior provide much more actionable guidance than marginal differences in ROC-AUC.

The confusion matrix analysis in Figure 7 reinforces the clinical significance of these trade-offs. XGBoost Original produced only 68 false negatives out of 1,591 stunted cases, corresponding to a 4.3% miss rate, which is the lowest among the best-performing configurations. LightGBM+SMOTE produced 86 false negatives (5.4% miss rate) but achieved the highest overall balance in terms of F1-Score and accuracy. Random Forest+SMOTE produced 118 false negatives (7.4% miss rate), indicating a less favorable sensitivity profile despite a comparatively controlled false-positive burden. These differences matter because the costs of misclassification in stunting screening are asymmetric. A false negative represents a child who may not receive timely nutritional monitoring or intervention during a critical growth window, whereas a false positive mainly creates additional follow-up workload. For this reason, the higher-Recall configurations are clinically preferable in most early-screening scenarios. The results therefore support a context-dependent deployment recommendation: XGBoost Original is preferable for high-sensitivity primary screening, while LightGBM+SMOTE may be more appropriate in resource-constrained settings where both detection performance and referral burden must be balanced.

These findings also have implications for operational integration into community health systems. Because the model uses seven routinely collected input features, the approach could be integrated into platforms such as Posyandu information systems or SIMPUS with minimal additional data collection burden. However, deployment should be approached cautiously. The present results were obtained from retrospective test data rather than prospective screening conditions, and no clinical workflow validation was conducted. Before operational use, the system would require prospective validation, threshold calibration for local prevalence, and assessment of how model outputs affect referral practices and health worker workload. In practice, the most appropriate model configuration may differ by setting. Rural or high-risk environments, where the cost of missing a case is especially high, may justify prioritizing XGBoost Original. Urban settings with heavier caseloads and more constrained follow-up capacity may benefit more from LightGBM+SMOTE, which offers a stronger overall balance.

Several limitations should be acknowledged. The dataset was drawn from a single geographic region, Bangka Belitung Province, which limits representativeness for Indonesia's broader anthropometric and socio-demographic diversity. Model behavior may shift when applied to populations with different nutritional baselines, healthcare access, or measurement practices. In addition, the study used cross-sectional routine data, so it could identify current stunting status but not predict longitudinal growth trajectories or pre-stunting risk patterns. The analysis also focused on structured tabular features and did not incorporate explainability methods, which are important for building clinical trust and supporting decision transparency. Future research should therefore extend this work through multi-site external validation, prospective implementation studies, and the inclusion of SHAP or LIME-based explanation layers to clarify which features drive risk predictions in individual cases. Longitudinal modeling may also help shift the use case from current screening toward earlier preventive warning.

This study indicates that the main contribution lies in showing that sampling strategy effectiveness is model-dependent, not universal. Standard SMOTE was the best overall resampling method at the aggregate level, as shown in Table 6 and Figure 5, but these average masks important model-specific behavior. XGBoost performed best without oversampling, LightGBM improved dramatically with SMOTE, and SMOTE-ENN consistently reduced sensitivity across all models. These results provide a more nuanced basis for

stunting-screening model selection than studies that evaluate only a single classifier–sampling pairing [10], [11]. Rather than recommending one method universally, the present findings support selecting configurations according to the clinical objective: maximize Recall when missed cases are unacceptable, or select a more balanced F1-oriented configuration when follow-up resources are constrained.

4. CONCLUSION

This study systematically evaluated 15 model-sampling combinations across three ensemble classifiers (XGBoost, Random Forest, LightGBM) and five oversampling strategies for pediatric stunting detection, using a compiled regional dataset of 10,000 records from Bangka Belitung Province, Indonesia. The central methodological finding is that oversampling strategy effects are model-specific and cannot be generalized across architectures: XGBoost's built-in class-weighting renders external oversampling counterproductive, LightGBM benefits most substantially from SMOTE (+12.6% Recall) due to GOSS-induced boundary gaps that synthetic samples compensate for, while Random Forest shows modest consistent improvement. SMOTE-ENN consistently underperformed across all three classifiers, suggesting that hybrid cleaning strategies may be counterproductive in high-prevalence imbalanced datasets with high within-class variability. The primary study limitation is single-region data provenance; external validation on multi-regional datasets is a prerequisite before broader deployment. Future work should prioritize multi-site validation, integration of explainability techniques (SHAP/LIME), and longitudinal trajectory modeling for pre-clinical early warning.

REFERENCES

[1] UNICEF, WHO, and World Bank, *Levels and Trends in Child Malnutrition: UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates, Key Findings of the 2023 Edition*. Geneva, Switzerland: World Health Organization, 2023.

[2] C. G. Victora *et al.*, "Maternal and child undernutrition: Consequences for adult health and human capital," *Lancet*, vol. 371, no. 9609, pp. 340–357, Jan. 2008, doi: 10.1016/S0140-6736(07)61692-4.

[3] Kementerian Kesehatan Republik Indonesia, *Hasil Survei Status Gizi Indonesia (SSGI) Tahun 2022*. Jakarta, Indonesia: Kemenkes RI, 2023.

[4] T. Vaivada, N. Akseer, S. Akseer, A. Somaskandan, M. Stefopulos, and Z. A. Bhutta, "Stunting in childhood: An overview of global burden, trends, determinants, and drivers of decline," *Am. J. Clin. Nutr.*, vol. 112, suppl. 2, pp. 777S–791S, Aug. 2020, doi: 10.1093/ajcn/nqaa159.

[5] A. T. Mulyani, M. A. Khairinisa, and A. Khatib, "Understanding stunting: Impact, causes, and strategy to accelerate stunting reduction—a narrative review," *Nutrients*, vol. 17, no. 5, p. 879, Feb. 2025, doi: 10.3390/nu17050879.

[6] L. Swastina, B. Rahmatullah, A. Saad, and H. Khan, "A systematic review on research trends, datasets, algorithms, and frameworks of children's nutritional status prediction," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 13, no. 2, pp. 1868–1877, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp1868-1877.

[7] N. Novalina, I. A. A. Tarigan, F. K. Kameela, and M. Rizkinia, "Benchmarking machine learning algorithm for stunting risk prediction in Indonesia," *Bull. Electr. Eng. Inform.*, vol. 14, no. 3, pp. 2252–2263, Jun. 2025, doi: 10.11591/eei.v14i3.8997.

[8] S. Ndagijimana, I. H. Kabano, E. Masabo, and J. M. Ntaganda, "Prediction of stunting among under-5 children in Rwanda using machine learning techniques," *J. Prev. Med. Public Health*, vol. 56, no. 1, pp. 41–49, Jan. 2023, doi: 10.3961/jpmph.22.367.

[9] Y. S. Dewi, S. Hastuti, and M. Fatekurohman, "Analysis of stunting in East Java, Indonesia using random forest and geographically weighted random forest regression," *Braz. J. Biometr.*, vol. 42, no. 3, pp. 213–224, 2024, doi: 10.28951/bjb.v42i3.679.

[10] A. A. G. Y. Pramana, M. F. Maulana, M. C. Tirtayasa, and D. A. Tyas, "Enhancing early stunting detection: A novel approach using artificial intelligence with an integrated SMOTE algorithm and ensemble learning model," in *Proc. IEEE Conf. Artif. Intell. (CAI)*, Singapore, Jun. 2024, pp. 486–493, doi: 10.1109/CAI59869.2024.00098.

[11] T. Sugihartono, B. Wijaya, Marini, A. F. Alkayes, and H. A. Anugrah, "Optimizing stunting detection through SMOTE and machine learning: A comparative study of XGBoost, Random Forest, SVM, and k-NN," *J. Appl. Data Sci.*, vol. 6, no. 1, pp. 667–682, Jan. 2025, doi: 10.47738/jads.v6i1.494.

[12] M. A. Hamid and E. R. Subhiyakto, "Performance comparison of Random Forest, SVM, and XGBoost algorithms with SMOTE for stunting prediction," *J. Appl. Informat. Comput. (JAIC)*, vol. 9, no. 4, pp. 1163–1169, Aug. 2025, doi: 10.30871/jaic.v9i4.9701.

[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[14] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput. (ICIC)*, Hefei, China, Aug. 2005, pp. 878–887, doi: 10.1007/11538059_91.

[15] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Hong Kong, China, Jun. 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[16] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[17] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.

[18] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[19] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Montreal, Canada, Aug. 1995, pp. 1137–1143.

[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 3146–3154.

[23] I. Tsamardinos, E. Greasidou, and G. Borboudakis, "Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation," *Mach. Learn.*, vol. 107, no. 12, pp. 1895–1922, 2018.

[24] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.