Nusantara Journal of Multidisciplinary Informatics

Volume 1 Issue 1
https://doi.org/10.56705/njomi.v1i1.4

Research Article

Beyond Accuracy: A Data-Validity and Target-Reconstruction Audit of Machine Learning for Stunting Classification in 40,071 Indonesian Toddlers

Akhmad Hidayat[1]
[1] Universitas Hasanuddin, Jl. Perintis Kemerdekaan No.10, Tamalanrea Indah, Kec. Tamalanrea, Kota Makassar, Sulawesi Selatan 90245, Indonesia, hidayata24h@studnt.unhas.ac.id

Correspondence should be addressed to Akhmad Hidayat; hidayata24h@studnt.unhas.ac.id

Received 22 January 2026 Accepted 20 March 2026; Published 30 April 2026

Abstract:
Introduction: Machine-learning models for stunting classification often report very high accuracy, but this performance may partly result from reconstructing anthropometric rules used to define the target rather than predicting future stunting risk. This study evaluated the validity of machine-learning-based stunting classification using 40,071 toddler records from Jeneponto Regency, Indonesia. Method: Raw spreadsheet integrity, annual sampling structure, anthropometric variable representation, target consistency, duplicate profiles, and feature provenance were audited before modeling. The 2024 cohort was used for primary analysis because the 2021–2023 datasets contained only stunted cases. Random Forest models were evaluated using five feature configurations under stratified and duplicate-profile-aware grouped cross-validation, supported by cross-algorithm robustness analysis. Results: The categorical stunting label showed 99.9937% agreement with the WHO HAZ < −2 criterion. Under grouped validation, Age, Gender, and Height achieved a ROC-AUC of 0.9886, which decreased to 0.7756 after Height was removed. Adding Weight-for-Height Z-score increased ROC-AUC to 0.9970, indicating indirect reintroduction of height information. Furthermore, 83.56% of classification errors occurred within ±0.20 SD of the WHO HAZ threshold, with 30.74-fold higher odds of misclassification near this boundary. Conclusion: High machine-learning performance in this dataset primarily reflects contemporaneous anthropometric target reconstruction. Stunting studies should therefore distinguish automated classification from prospective prediction and explicitly evaluate data provenance, feature construction, and validation design.

Keywords: Stunting; Machine Learning; Target Reconstruction; Data Leakage; Anthropometry; Feature Ablation; Validation Bias.

1. Introduction

Childhood stunting remains an important public health problem worldwide. In 2024, approximately 150.2 million children under five were estimated to be stunted [1]. According to the World Health Organization (WHO), stunting is defined as length or height for age below minus two standard deviations from the WHO Child Growth Standards [2]. Beyond impaired physical growth, childhood stunting has been associated with poorer educational attainment and cognitive outcomes later in life [3], [4]. In Indonesia, the national stunting prevalence decreased from 21.5% in 2023 to 19.8% in 2024, but the condition remains a major nutritional concern [5]. Previous Indonesian studies have linked stunting with child, maternal, socioeconomic, sanitation, and environmental factors [6], while prospective studies have demonstrated that information available early in life can be used to estimate later stunting risk [7].

The increasing availability of health data has encouraged the use of machine-learning methods for stunting assessment. One publicly available dataset from Jeneponto Regency, South Sulawesi, contains 40,071 records of children measured between 2021 and 2024, including age, sex, weight, height, several anthropometric Z-scores, and

nutritional-status classifications [8]. Studies using this dataset have reported high performance from tree-based algorithms. Wicaksono et al. obtained accuracies above 96% using XGBoost, Random Forest, and Decision Tree and identified height as the most influential feature [9]. Resha and Toding similarly reported high XGBoost performance using basic anthropometric measurements after excluding explicit Z-score predictors [10].

Machine-learning classification has also been applied across diverse health-related datasets using different combinations of algorithms, preprocessing procedures, and validation strategies. Previous studies have evaluated classification performance for cardiovascular disease, anemia, obesity, mental-health disorders, and breast-cancer diagnosis using approaches such as K-Nearest Neighbor, Decision Tree, Random Forest, ensemble learning, and cross-validation [11], [12], [13]–[15]. These studies demonstrate the broad applicability of machine learning in health-related classification, while also emphasizing that performance estimates depend on the characteristics of the dataset, predictor construction, and evaluation procedure.

However, high classification accuracy does not necessarily indicate genuine prospective prediction. Under the WHO growth standard, Height-for-Age Z-score (HAZ) is calculated using a child's age, sex, and measured height, while stunting status is subsequently determined from the HAZ threshold of −2 SD [2]. Therefore, a model trained using age, sex, and height measured at the same time as the outcome may primarily approximate the anthropometric decision rule rather than identify independent predictors of future stunting. The same concern applies to derived variables such as Weight-for-Height Z-score, which indirectly retains height information. Such relationships are related to target or label leakage, which may lead to overly optimistic interpretations of machine-learning performance when predictors contain information closely associated with outcome construction [16], [17].

The importance of this distinction is also demonstrated by recent leakage-aware healthcare research, where removing potentially post-decision variables produced a substantial decline in model performance, indicating that apparently strong discrimination may partly originate from features that would not be appropriate at the intended prediction point [18].

Recent research has begun to question extremely high performance in anthropometric stunting classification. Putra and Hendrastuty showed that near-perfect machine-learning results on another large stunting dataset were strongly influenced by deterministic relationships between anthropometric predictors and nutritional labels [19]. However, existing research using the Jeneponto dataset has mainly focused on algorithm comparison, optimization, and explainability [9], [10]. The effects of dataset provenance, year-specific sampling, spreadsheet representation, repeated anthropometric profiles, and direct or indirect height-derived information have not been systematically evaluated. These issues are important because inappropriate data preparation or validation design can produce performance estimates that do not represent generalization to genuinely unseen observations [11], [20].

Accordingly, this study conducts a data-validity and target-reconstruction audit of machine-learning-based stunting classification using the Jeneponto dataset. The analysis examines raw spreadsheet integrity and annual sampling structure, evaluates the dependence of model performance on direct and indirect height information, compares conventional stratified validation with duplicate-profile-aware grouped validation, and investigates whether classification errors are concentrated around the WHO HAZ = −2 decision boundary. Similarly, recent health-informatics research has emphasized that machine-learning outputs should be interpreted according to their demonstrated predictive capability; when discrimination is limited, the resulting model may be more appropriately positioned as an exploratory assessment tool rather than a clinical diagnostic system [21]. The study addresses four research questions: RQ1: What data-integrity and sampling characteristics affect the suitability of the dataset for machine-learning-based stunting classification? RQ2: To what extent does model performance depend on direct or indirect height-derived information? RQ3: How does duplicate-profile-aware validation affect apparent model performance compared with conventional random stratified validation? RQ4: Are classification errors concentrated around the WHO stunting threshold, and what does this imply for interpreting high performance as prediction rather than target reconstruction?

2. Method

Data Source and Study Design

This study used the publicly available Dataset Stunting and Nutritional Status of Toddler from Jeneponto Regency, South Sulawesi, Indonesia, Version 4 [8]. The dataset contains 40,071 records collected between 2021 and 2024, including age, gender, weight, height, Weight-for-Age Z-score (WAZ), Height-for-Age Z-score (HAZ), Weight-for-Height Z-score (WHZ), and nutritional-status labels.

A methodological validation design was applied to examine data integrity, target reconstruction, feature dependence, and model-validation bias. The raw annual Excel files were treated as the primary data source, while the combined and preprocessed files were used only for consistency checking.

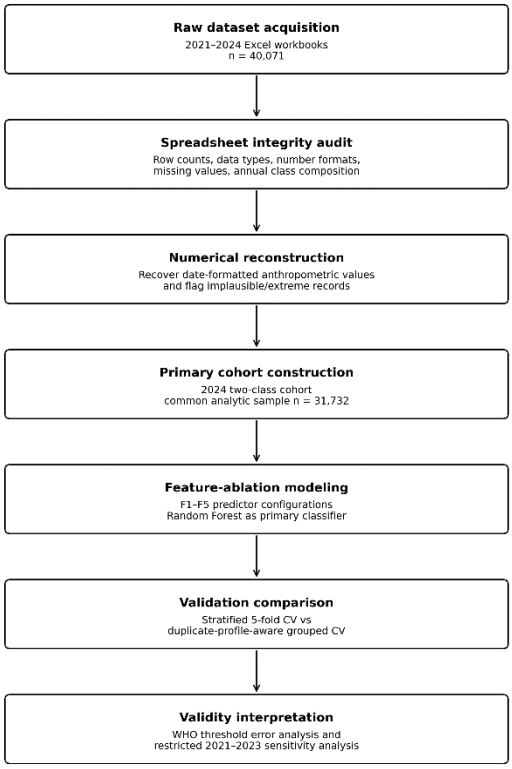

Figure 1. Data-validity and target-reconstruction audit workflow

Data Integrity and Cohort Selection

Each workbook was inspected at cell level to verify row counts, data types, number formats, missing values, class distributions, and repeated records. Date-formatted anthropometric cells were reconstructed according to their displayed numerical representation before statistical analysis.

The categorical Height for Age variable was used as the binary target, with Stunted = 1 and Not Stunted = 0. HAZ was excluded from all predictive feature sets because stunting is defined by the WHO criterion $HAZ < -2$ [2]. Agreement between the stored categorical label and this threshold was examined separately as a target-consistency audit.

The 2024 data were selected for the primary machine-learning analysis because only this year contained both outcome classes. Records with ($|HAZ| > 6$) were excluded according to anthropometric plausibility criteria [22], resulting in a common analytic cohort of 31,732 observations. Missing age values were handled by median imputation within each training fold.

Feature-Ablation Experiments

Five predictor configurations were evaluated to determine how direct and indirect height information affected model performance.

Table 1. Predictor configurations

Configuration	Predictors
F1	Age, Gender, Height
F2	Age, Gender, Weight
F3	Age, Gender, Weight, WAZ
F4	Age, Gender, Weight, WAZ, WHZ
F5	Age, Gender, Weight, Height

HAZ and the categorical Height-for-Age status were excluded from all predictor sets. WHZ was considered an indirect height-derived variable because the indicator is calculated using weight relative to height [2].

Duplicate-Profile and Validation Strategy

Repeated anthropometric profiles were identified using identical combinations of gender, age, weight, height, WAZ, HAZ, and WHZ. These records were retained but assigned to the same validation group to prevent identical profiles from appearing simultaneously in training and testing sets.

Two five-fold validation strategies were compared:

- conventional stratified cross-validation; and
- duplicate-profile-aware stratified group cross-validation.

The same analytic sample and fold assignments were used for all feature configurations to ensure fair comparison.

Machine-Learning Models

Random Forest was used as the primary classifier because tree-based methods have shown strong performance on the Jeneponto dataset [9], [10]. The model used 100 trees and a fixed random state of 42 [23].

The use of multiple algorithms as a robustness assessment is consistent with previous classification studies by Azis and colleagues. Their work has evaluated K-Nearest Neighbor with cross-validation for cardiovascular disease [11], Random Forest for anemia prediction [12], comparative classification algorithms under cross-validation [24], Decision Tree modeling with multi-metric evaluation [25], and deep-learning approaches for medical-image classification [14]. Collectively, these studies support evaluating model behavior across different learning paradigms rather than interpreting a single high accuracy value as sufficient evidence of generalizability.

To test whether the observed pattern depended on a specific algorithm, additional robustness experiments were conducted using Logistic Regression, Decision Tree, and Histogram-based Gradient Boosting. Extensive hyperparameter optimization was intentionally avoided because the objective was to evaluate model validity rather than identify the highest-performing algorithm.

Evaluation Metrics

Performance was evaluated using ROC-AUC, Precision–Recall AUC, balanced accuracy, sensitivity, specificity, F1-score, and Brier score. Predictions from all validation folds were combined into a complete out-of-fold prediction set before calculating the final metrics.

Balanced accuracy was calculated as:

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \quad (1)$$

and F1-score as:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2)$$

WHO Threshold Error Analysis

To evaluate whether model errors were concentrated near the WHO stunting boundary, the absolute distance between each HAZ value and −2 was calculated as:

$$D = |HAZ + 2| \quad (3)$$

Misclassification rates were examined within ±0.10, ±0.20, and ±0.50 SD of the threshold. Error rates near and farther from the ±0.20 SD boundary were also compared using an odds ratio.

Cross-Year Sensitivity Analysis

Models trained on the 2024 cohort were additionally applied to the 2021–2023 datasets. Because these historical files contained only stunted observations, they were not treated as conventional external validation cohorts. Only sensitivity was reported, while ROC-AUC, specificity, and balanced accuracy were omitted because these metrics require both positive and negative cases.

Software and Reproducibility

Data processing and analysis were conducted in Python using `openpyxl`, `pandas`, `NumPy`, and `scikit-learn` [26]. A random seed of 42 was used throughout the experiments. All preprocessing procedures, including imputation, were fitted only within the corresponding training folds to prevent information leakage.

3. Result and Discussion

Data Integrity and Sampling Characteristics

The four annual datasets contained 40,071 records, consisting of 3,412 observations from 2021, 1,111 from 2022, 3,808 from 2023, and 31,740 from 2024. However, the class distribution differed substantially across years. All observations from 2021–2023 were labelled as stunted, whereas the 2024 dataset contained both classes.

Table 2. Year-specific stunting distribution

Year	Total	Stunted	Not Stunted	Prevalence
2021	3,412	3,412	0	100.00%
2022	1,111	1,111	0	100.00%
2023	3,808	3,808	0	100.00%
2024	31,740	7,873	23,867	24.80%
Overall	**40,071**	**16,204**	**23,867**	**40.44%**

This structure indicates that the earlier annual files represent positive-only samples and cannot be used independently for conventional binary-classification evaluation. Therefore, the 2024 cohort was used for the primary analysis.

The spreadsheet audit also identified date-formatted numerical cells. In 2024, 20,158 Weight cells, or 63.51%, were stored internally using Excel date representations but visually displayed as decimal values. Similar formatting was found in several WAZ, HAZ, and WHZ cells. These values required reconstruction before modeling, demonstrating that spreadsheet-level integrity checks are necessary before automatic numerical conversion.

Target Consistency and Feature Dependence

The WHO criterion HAZ < −2 reproduced the categorical stunting label for 31,738 of 31,740 observations, corresponding to 99.9937% agreement. Only two records were inconsistent with the expected threshold relationship. This confirms that HAZ is effectively a target-defining variable and should not be included as a predictor.

After quality filtering, the common modeling cohort consisted of 31,732 observations. Random Forest performance under the five feature configurations is presented in **Table 3**.

Table 3. Random Forest performance under feature ablation

Features	Stratified ROC-AUC	Grouped ROC-AUC	Grouped F1
Age + Gender + Height	**0.9958**	**0.9886**	**0.944**
Age + Gender + Weight	0.8595	**0.7756**	0.5439
Age + Gender + Weight + WAZ	0.909	**0.7805**	0.5205
Age + Gender + Weight + WAZ + WHZ	**0.9992**	**0.997**	**0.9477**
Age + Gender + Weight + Height	**0.9972**	**0.9904**	0.9341

Removing Height reduced grouped ROC-AUC from 0.9886 to 0.7756, indicating that direct linear-growth information accounted for a substantial proportion of classification performance. Adding WAZ produced only a small improvement, whereas adding WHZ increased ROC-AUC to 0.9970.

This result is important because WHZ indirectly contains height information through the WHO weight-for-height calculation [2]. Consequently, removing the explicit Height variable while retaining WHZ does not constitute a genuinely height-independent prediction model. **Figure 2** illustrates the performance differences across feature configurations and validation strategies.

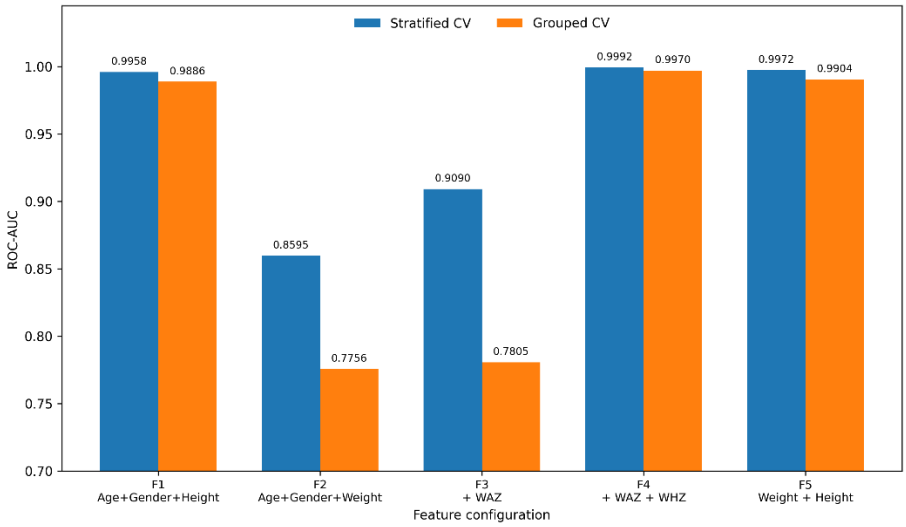

Figure 2. Comparison of ROC-AUC across feature configurations under stratified and duplicate-profile-aware grouped cross-validation.

The findings indicate that high classification performance depends strongly on variables directly or indirectly related to the anthropometric definition of stunting. This supports the methodological concern that predictor provenance should be evaluated rather than considering only predictor names or conventional feature importance.

Effect of Validation Strategy and Algorithm

The dataset contained 4,132 repeated anthropometric profile groups involving 8,320 observations. Preventing identical complete profiles from appearing simultaneously in training and validation sets had the greatest effect on height-free models. For Age + Gender + Weight, ROC-AUC decreased from 0.8595 to 0.7756, while the addition of WAZ decreased from 0.9090 to 0.7805 under grouped validation.

The effect was much smaller when Height was included because the target-related anthropometric signal remained very strong. This shows that conventional random validation can produce optimistic estimates, particularly when predictor information is weaker and repeated profiles are distributed across folds [16].

The high performance of Height-based classification was also consistent across nonlinear algorithms.

Table 4. Grouped-validation performance using Age + Gender + Height

Model	ROC-AUC	F1
Logistic Regression	0.9154	0.6307
Decision Tree	0.9707	0.9408
Random Forest	0.9886	0.944
HistGradientBoosting	0.9953	0.944

Variation across algorithms and validation folds has also been observed in previous machine-learning applications. Cross-validation studies involving obesity and mental-health classification reported measurable variation in performance across folds [26], [27], while Decision Tree-based breast-cancer classification similarly emphasized validation beyond a single train-test estimate [28]. These observations reinforce the present study's use of multiple evaluation strategies and support interpreting performance in relation to validation design rather than algorithm ranking alone.

Decision Tree, Random Forest, and HistGradientBoosting all achieved ROC-AUC above 0.97. This confirms previous findings that nonlinear tree-based models perform particularly well on the Jeneponto dataset [9], [10]. However, the present results suggest a different interpretation: these models efficiently approximate the nonlinear relationship between age, gender, height, and the WHO-defined stunting boundary rather than necessarily discovering independent future-risk factors.

Errors Around the WHO Stunting Boundary

Classification errors were strongly concentrated around HAZ = −2. Among the 870 errors generated by the grouped Random Forest model using Age + Gender + Height, 61.95% occurred within ±0.10 SD, 83.56% within ±0.20 SD, and 97.13% within ±0.50 SD of the threshold.

The error rate among observations within ±0.20 SD was 14.24%, compared with only 0.54% among observations farther from the threshold. The resulting odds ratio was 30.74 with a 95% confidence interval of 25.62–36.88.

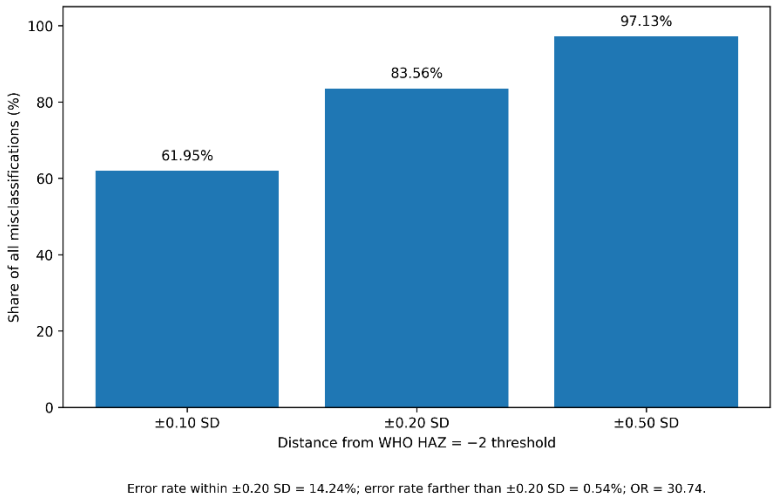

Figure 3. Concentration of classification errors around the WHO HAZ = −2 stunting decision boundary.

The concentration of errors near the exact decision boundary provides further evidence that the model primarily approximates the anthropometric classification rule. If the model were identifying broader independent risk mechanisms, errors would be expected to be distributed more widely across the anthropometric space. This finding therefore supports interpreting the high scores as target reconstruction or contemporaneous classification, rather than prospective prediction [18], [19].

Cross-Year Sensitivity and Overall Interpretation

Models trained on 2024 were additionally applied to the positive-only historical cohorts. Because these datasets contained no non-stunted observations, only sensitivity was calculated.

Table 5. Sensitivity in historical positive-only cohorts

Predictors	2021	2022	2023
Age + Gender + Height	83.62%	93.97%	92.80%
Age + Gender + Weight	35.52%	44.91%	41.62%

The Height-based model retained high sensitivity, whereas the height-free model identified fewer than half of the historical stunted cases. This again confirms the dominant role of linear-growth information. Nevertheless, these results should not be interpreted as full temporal external validation because specificity and discrimination could not be assessed in single-class historical datasets.

Overall, four findings were consistent across the analyses. First, the categorical outcome was almost completely determined by the WHO HAZ threshold. Second, machine-learning performance declined substantially when direct and indirect height information was removed. Third, duplicate-profile-aware validation reduced the apparent performance of height-free models. Fourth, most classification errors occurred close to the WHO decision boundary.

These results indicate that high machine-learning accuracy in this dataset primarily reflects automated reconstruction of current anthropometric status. Such models may still be useful for nutritional screening or automated classification, but they should not automatically be described as early stunting prediction. Genuine prospective prediction requires predictors available before outcome determination, such as maternal, birth, socioeconomic, environmental, feeding, morbidity, or longitudinal growth factors [7], [20].

4. Conclusion

This study showed that very high machine-learning performance in stunting classification should be interpreted carefully when the predictors are closely related to the anthropometric variables used to define the target. Using 40,071 records from Jeneponto Regency, the analysis identified important issues involving annual sampling structure, spreadsheet representation, repeated anthropometric profiles, and feature provenance.

The categorical stunting label showed 99.9937% agreement with the WHO HAZ < -2 criterion. Under duplicate-profile-aware validation, Random Forest using Age, Gender, and Height achieved a ROC-AUC of 0.9886, but performance decreased to 0.7756 when Height was removed. Adding WHZ increased ROC-AUC to 0.9970, indicating that derived variables may indirectly reintroduce height information. Moreover, 83.56% of classification errors occurred within ±0.20 SD of the WHO threshold, showing that the model primarily approximated the anthropometric decision boundary.

These findings indicate that high performance on this dataset is more appropriately interpreted as automated classification or target reconstruction than as prospective prediction of future stunting. Machine-learning studies should therefore evaluate data provenance, feature construction, sampling structure, and validation design before translating high accuracy into claims of early prediction. Future prospective models should prioritize predictors available before outcome determination, including maternal, birth, socioeconomic, environmental, feeding, morbidity, and longitudinal growth factors.

References

[1] World Health Organization, United Nations Children's Fund, and World Bank Group, Levels and Trends in Child Malnutrition: UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates—Key Findings of the 2025 Edition. Geneva, Switzerland: World Health Organization, 2025.

[2] World Health Organization, WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development. Geneva, Switzerland: WHO, 2006.

[3] E. Lestari, A. Siregar, A. K. Hidayat, and A. A. Yusuf, "Stunting and its association with education and cognitive outcomes in adulthood: A longitudinal study in Indonesia," PLOS ONE, vol. 19, no. 5, Art. no. e0295380, 2024, doi: https://doi.org/10.1371/journal.pone.0295380.

[4] M. A. Alam et al., "Impact of early-onset persistent stunting on cognitive development at 5 years of age: Results from a multi-country cohort study," PLOS ONE, vol. 15, no. 1, Art. no. e0227839, 2020, doi: https://doi.org/10.1371/journal.pone.0227839.

[5] Badan Kebijakan Pembangunan Kesehatan, Survei Status Gizi Indonesia (SSGI) 2024 Dalam Angka. Jakarta, Indonesia: Kementerian Kesehatan Republik Indonesia, 2025.

[6] H. Anastasia et al., "Determinants of stunting in children under five years old in South Sulawesi and West Sulawesi Province: 2013 and 2018 Indonesian Basic Health Survey," PLOS ONE, vol. 18, no. 5, Art. no. e0281962, 2023, doi: https://doi.org/10.1371/journal.pone.0281962.

[7] D. Azriani, D. Agustian, Y. Zuhairini, I. N. Yulita, and M. Dhamayanti, "Prediction models for stunting at 2-years-old from Indonesian newborn population," BMC Pediatrics, vol. 25, Art. no. 718, 2025, doi: https://doi.org/10.1186/s12887-025-06096-4.

[8] A. H. RS, P. Purnamawati, H. Jaya, A. Arfandi, I. Suhardi, S. Widodo, and S. Lonang, "Dataset Stunting and Nutritional Status of Toddler from Jeneponto Regency, South Sulawesi, Indonesia," Mendeley Data, Version 4, 2026, doi: https://doi.org/10.17632/wzwpc9j5bx.4.

[9] A. Wicaksono, D. Prasetyo, Y. Mar'atullatifah, D. U. Iswavigra, H. Mahmudah, and A. Hapsari, "Data Analysis and Explainable Machine Learning for Stunting Prediction," Journal of Artificial Intelligence and Legal Technology, vol. 1, no. 1, pp. 35–44, 2025.

[10] M. Resha and A. Toding, "Predicting Independent Z-Score Stunting Through Fundamental Anthropometric Measurements Utilizing Extreme Gradient Boosting (XGBoost)," Journal Zetroem, vol. 8, no. 2, pp. 148–154, 2026, doi: https://doi.org/10.36526/ztr.v8i2.8736.

[11] H. Azis, Purnawansyah, F. Fattah, and I. P. Putri, "Performa klasifikasi K-NN dan cross-validation pada data pasien pengidap penyakit jantung," ILKOM Jurnal Ilmiah, vol. 12, no. 2, pp. 81–86, 2020, doi: https://doi.org/10.33096/ilkom.v12i2.507.81-86.

[12] H. Azis and N. Rismayanti, "Prediksi anemia dari pixel gambar dan level hemoglobin menggunakan Random Forest Classifier," NERO (Networking Engineering Research Operation), vol. 9, no. 1, 2024, doi: https://doi.org/10.21107/nero.v9i1.27916.

[13] F. T. Admojo and N. Rismayanti, "Estimating obesity levels using Decision Trees and K-Fold Cross-Validation: A study on eating habits and physical conditions," Indonesian Journal of Data and Science, vol. 5, no. 1, pp. 37–44, 2024, doi: https://doi.org/10.56705/ijodas.v5i1.126.

[14] A. P. Wibowo, M. Taruk, T. E. Tarigan, and M. Habibi, "Improving mental health diagnostics through advanced algorithmic models: A case study of bipolar and depressive disorders," Indonesian Journal of Data and Science, vol. 5, no. 1, pp. 8–14, 2024, doi: https://doi.org/10.56705/ijodas.v5i1.122.

[15] A. Halid, I. G. N. W. Arsa, R. A. Azdy, and A. A. J. Permana, "Development of a Decision Tree classifier for breast cancer diagnosis using Fine Needle Aspirate data," Indonesian Journal of Data and Science, vol. 5, no. 3, pp. 229–236, 2024, doi: https://doi.org/10.56705/ijodas.v5i3.202.

[16] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023, doi: https://doi.org/10.1016/j.patter.2023.100804.

[17] S. E. Davis, M. E. Matheny, S. Balu, and M. P. Sendak, "A framework for understanding label leakage in machine learning for health care," Journal of the American Medical Informatics Association, vol. 31, no. 1, pp. 274–280, 2024, doi: https://doi.org/10.1093/jamia/ocad178.

[18] D. H. Zulfikar, E. S. J. Atmadji, and B. S. W. Poetro, "Leakage-aware and explainable machine learning for healthcare claim fraud detection using imbalanced medical insurance data," International Journal of Artificial Intelligence in Medical Issues, vol. 4, no. 1, pp. 19–34, 2026, doi: https://doi.org/10.56705/z3207345.

[19] T. A. Putra and N. Hendrastuty, "Evaluasi validitas model machine learning pada klasifikasi stunting berbasis data antropometri dan hubungan deterministik," Building of Informatics, Technology and Science (BITS), vol. 8, no. 1, pp. 62–72, 2026, doi: https://doi.org/10.47065/bits.v8i1.9584.

[20] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, Art. no. e078378, 2024, doi: https://doi.org/10.1136/bmj-2023-078378.

[21] M. I. Siami, A. W. Murdiyanto, and Sumiyatun, "A comparative study of machine learning models for stress level classification using social media and lifestyle data," International Journal of Artificial Intelligence in Medical Issues, vol. 4, no. 1, pp. 93–109, 2026, doi: https://doi.org/10.56705/r4d62a66.

[22] World Health Organization and United Nations Children's Fund, Recommendations for Data Collection, Analysis and Reporting on Anthropometric Indicators in Children Under 5 Years Old. Geneva, Switzerland: World Health Organization, 2019.

[23] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[24] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001, doi: https://doi.org/10.1214/aos/1013203451.

[25] H. Azis, F. T. Admojo, and E. Susanti, "Analisis perbandingan performa metode klasifikasi pada dataset multiclass citra busur panah," Techno.Com, vol. 19, no. 3, pp. 286–294, 2020, doi: https://doi.org/10.33633/tc.v19i3.3646.

[26] H. Azis and S. R. Jabir, "Chemical composition and aroma profiling: Decision Tree modeling of formalin tofu," Journal of Embedded Systems, Security and Intelligent Systems, vol. 4, no. 2, pp. 206–211, 2023, doi: https://doi.org/10.59562/jessi.v4i2.1162.

[27] H. Azis, Y. Salim, and Y. Puspitasari, "Deep learning approaches for soft tissue classification of cephalometric images in orthodontics," JOIV: International Journal on Informatics Visualization, vol. 10, no. 3, pp. 988–995, 2026, doi: https://doi.org/10.62527/joiv.10.3.3885.

[28] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, Art. no. e0118432, 2015, doi: https://doi.org/10.1371/journal.pone.0118432.

[29] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011.