Machine Learning-Based Stunting Classification Using Puskesmas Nutritional Data: A Comparative Study of Six Algorithms with Hyperparameter Tuning

Asmaul Husna RS*, Purnamawati2, Hendra Jaya, Anas Arfandi4
1, Doctoral Program in Engineering Vocational Education, Postgraduate Program, Makassar State University, Makassar, Indonesia
2,3,4Department of Faculty of Engineering, Makassar State University, Makassar, Indonesia

ABSTRACT

Stunting detection at Indonesia's community health centers (Puskesmas) still relies heavily on manual anthropometric assessment a process that is slow, inconsistent, and difficult to scale. This study applies machine learning to automate that classification using 40,071 real-world records from the Jeneponto Regency Health Department (2021–2024), covering three nutritional categories: Normal, Stunting, and Severe, derived from Height-for-Age Z-Scores per WHO standards. Six classifiers were compared: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting (GB). Each was tested in both default and tuned configurations, with hyperparameters optimised via GridSearchCV and stratified k-fold cross-validation. Preprocessing included invalid value removal, target leakage prevention, label encoding, StandardScaler normalisation, and stratified 80:20 splitting. Performance was measured using weighted Accuracy, Precision, Recall, and F1-Score. At baseline, RF led with 97.04% accuracy and 97.05% F1-Score, while LR trailed at 77.35% a gap that points to a non-linear classification boundary that linear models cannot handle. After tuning, GB came out ahead with 97.54% accuracy and F1-Score, overtaking RF at 97.22%. That reversal matters: it shows that GB's sequential architecture holds more capacity, but only delivers it when key parameters are properly configured. Tuned GB shows real promise as a decision-support tool for Puskesmas-level stunting screening at scale.

*Keywords:*Anthropometric Data Child Nutrition Machine Learning Stunting Classification

INTRODUCTION

Stunting defined as low height-for-age relative to WHO standards affects children's cognitive development, immune function, and long-term productivity in ways that are largely irreversible (Sadler et al., 2022; Shevaldo et al., 2024). In 2022, around 148 million children under five were stunted globally, with the heaviest burden falling on Sub-Saharan Africa and Southeast Asia (Akseer et al., 2022; Fadhilah et al., 2024). Indonesia sits among the worst-affected countries: its 21.6% prevalence that year still exceeded the WHO's 20% emergency threshold, with provinces like South Sulawesi and East Nusa Tenggara faring considerably worse than the national figure (Sudigyo et al., 2023; Lestari et al., 2025).

The consequences go well beyond physical growth. Stunted children tend to perform worse in school, earn less as adults, and are more likely to raise stunted children of their own a cycle that compounds across generations (Yosep et al., 2026; Khoirunnisa et al., 2024; Erda et al., 2024). The economic costs are substantial too, spanning healthcare expenditure, lost productivity, and reduced GDP (Janssen et al., 2025; Nur et al., 2024). Indonesia has formally recognised this, setting a 14% stunting target under the 2020–2024 RPJMN an ambition that requires not just nutritional programmes but also better detection infrastructure at the community level (Bukit et al., 2024).

[1]* Corresponding author.
E-mail address: asmaul.husna@student.unm.ac.id

Detection today depends on health workers at Puskesmas and Posyandu manually comparing Z-scores against WHO charts (Yulianto et al., 2025; Firdausi et al., 2025). The system works, but it strains under its own scale. Classification is slow, prone to inter-rater inconsistency, and unevenly staffed across a geographically fragmented country (Shi et al., 2022; Prabiantissa et al., 2024). With over 10,000 Puskesmas generating records continuously, identifying at-risk children quickly enough to intervene within the critical first 1,000 days remains a genuine operational challenge (Sundjaya et al., 2024; Gunawan et al., 2025; Kani et al., 2024).

Machine learning offers a practical path forward. Studies across Indonesia (Miranda et al., 2024), Egypt (Hendy et al., 2025), and Malawi (Mgomezulu et al., 2025) have tested classifiers ranging from RF and XGBoost to LR and SVM on anthropometric and survey data, consistently finding that ensemble methods outperform linear models and that class imbalance handling matters. RF with SMOTE, for instance, reached 88.9% accuracy and 93.0% ROC-AUC on Jakarta EMR data [20], while XGBoost topped six classifiers on Egyptian DHS data with 94.8% accuracy (Hendy et al., 2025). Work in Malawi pushed RF to near-perfect accuracy after SMOTE balancing, with LR trailing in every comparison (Mgomezulu et al., 2025). Beyond classifier benchmarks, studies have also brought in spatial data (Nduwayezu et al., 2025), satellite imagery (Arya et al., 2025), and community intervention variables (Arqam et al., 2026) broadening the feature space and reinforcing that ensemble methods paired with careful imbalance handling tend to produce the strongest results.

Most of this work, however, tests only a narrow set of algorithms, skips systematic hyperparameter tuning, or relies on data that does not reflect real Puskesmas operations. This study addresses those gaps directly. Six classifiers LR, DT, RF, SVM, KNN, and GB were benchmarked on 40,071 records from the Jeneponto Regency Health Department (2021–2024), each evaluated in both default and tuned configurations using GridSearchCV with stratified k-fold cross-validation across four weighted metrics.

Three things come out of this. First, it provides a broad algorithm-level comparison grounded in real Puskesmas data. Second, it shows that tuning changes not just scores but rankings GB overtakes RF after optimisation, something baseline benchmarking alone would not reveal. Third, it identifies tuned GB (weighted F1-Score: 97.54%) as the most suitable model for automated stunting screening, offering a concrete starting point for practitioners building decision-support tools in Indonesia's community health system.

METHODS

A. Dataset

The dataset used in this study was obtained from the Dinas Kesehatan (Health Department) of Jeneponto Regency and covers child nutritional records collected over a four-year period from 2021 to 2024. Comprising 40,071 records and 12 attributes, the dataset captures a broad range of anthropometric and nutritional information, including sequential record number, gender (encoded as 0 for female and 1 for male), age in months, body weight, body height, and three WHO-based nutritional indicators Weight for Age, Height for Age, and Weight for Height each accompanied by its corresponding Z-Score value.

Figure. 1 show that each child is classified into one of three categories: Normal, Stunting, or Severe. Making this a multi-class classification problem grounded in real-world public health data.

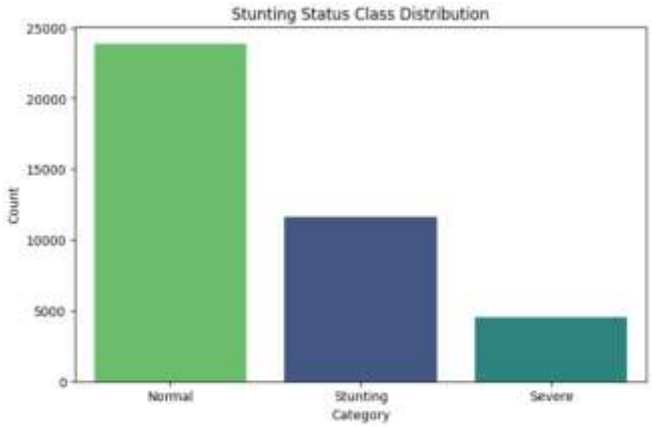

Figure 1. Class Distribution

B. Preprocessing

The raw dataset underwent the following preprocessing steps before model training:

1. Feature Selection

As shown in Figure. 2 not all columns in the raw dataset carry useful predictive signal, and including the wrong ones can actively mislead a model. The record index column (No.) was dropped immediately, as it is merely a row identifier with no semantic meaning. The three categorical Z-score label columns Weight for Age, Height for Age, and Weight for Height were also removed, since they represent discretised interpretations of the underlying numeric Z-scores rather than independent measurements, introducing redundancy without adding new information. More critically, the raw Z-Score H/A numeric column was explicitly excluded to prevent target leakage: because the Status label is derived directly and deterministically from this value, retaining it would give the model a shortcut that does not exist in real deployment scenarios, producing inflated performance metrics that would not generalise to unseen data. The final feature set retained for modelling consisted of six columns Gender, Age (Month), Weight, Height, Z-Score W/A, and Z-Score W/H representing genuinely independent anthropometric inputs.

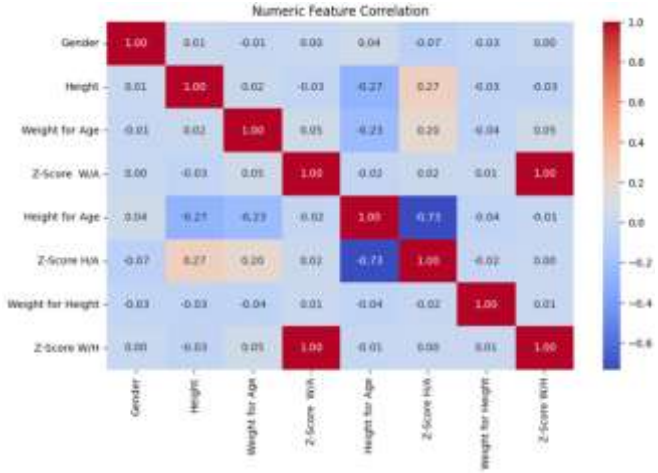

Figure 2. Feature Correlation

2. Type Coercion:

Field-collected health records frequently contain non-numeric entries resulting from data entry errors or system export artefacts. In this dataset, values such as #NUM! were encountered in several numeric columns, which pandas would otherwise read as object-type strings. To address this, all feature columns (excluding Status) were explicitly cast to numeric type using pd.to_numeric() with errors='coerce', converting any unparseable string into NaN. Rows containing these missing values were then dropped entirely, ensuring a clean, fully numeric feature matrix compatible with scikit-learn estimators.

3. Categorical Encoding:

Machine learning algorithms operate on numerical input, yet the Gender column in its original form carried string or categorical values. scikit-learn's LabelEncoder was applied to map these categories to integer representations (0 and 1), preserving the binary nature of the variable without introducing the additional dimensionality that one-hot encoding would create for a two-class feature.

4. Feature Scaling:

The six retained features span substantially different numerical ranges age is measured in months (0–59), weight in kilograms, height in centimetres, and Z-scores in standardised units. Without scaling, distance-based and gradient-based algorithms such as KNN, SVM, and LR would be disproportionately influenced by features with larger absolute values. StandardScaler was applied to transform all features to zero mean and unit variance, placing every variable on a common scale and ensuring that each contributes proportionally to the model's decision boundary. The scaled feature matrix is reflected in the sample output above, where all values are centred around zero regardless of their original measurement unit.

5. Dataset Splitting:

The preprocessed dataset was partitioned into training (80%) and test (20%) subsets using train_test_split with stratify=y and random_state=42. Stratified splitting ensures that the proportional representation of Normal, Stunting, and Severe cases is preserved in both partitions a critical consideration given the class imbalance observed in the dataset. Without stratification, random sampling could inadvertently under-represent the minority Severe class in the test set, leading to unreliable evaluation of the model's performance on the most clinically significant category.

C. Classification Models

Six supervised classification algorithms were selected to represent a diverse spectrum of model families, ranging from linear parametric methods to non-parametric instance-based learners and ensemble approaches. This deliberate breadth allows for a rigorous and systematic comparison of how different inductive biases interact with the anthropometric feature space inherent in stunting classification.

1. Logistic Regression

Despite its name, LR is a discriminative probabilistic classifier that models the posterior probability of class membership by applying the softmax function to a linear combination of input features in the multi-class setting. It serves as an interpretable linear baseline, allowing the study to quantify the degree to which non-linear decision boundaries are necessary for this classification task. Regularisation is applied via the inverse strength parameter C, with

both L1 (Lasso) and L2 (Ridge) penalty types included in the hyperparameter search to assess the impact of sparsity-inducing versus weight-shrinkage regularisation on generalisation performance.

2. Decission Tree

DT constructs an axis-aligned, hierarchical partition of the feature space by recursively selecting the split that maximises information gain at each internal node. While highly interpretable, DTs are prone to overfitting when grown without constraint, as they can memorise training samples by creating arbitrarily fine partitions. To govern this trade-off between bias and variance, the hyperparameter search encompassed the split criterion (Gini impurity and Information Gain entropy), maximum tree depth, and the minimum number of samples required to split an internal node or constitute a leaf, collectively controlling model complexity.

3. Random Forest

RF is a bagging-based ensemble that constructs a collection of decorrelated DTs, each trained on a bootstrapped sample of the training data with a randomly selected subset of features considered at each split. Predictions are aggregated by majority voting across all trees, which substantially reduces the variance that characterises individual DTs while maintaining low bias. The key hyperparameters tuned in this study number of estimators, maximum depth, and minimum samples per split govern the ensemble's capacity, individual tree complexity, and the degree of regularisation imposed on each learner (Kanz et al., 2024).

4. Support Vector Machine

SVM identifies the maximum-margin hyperplane that separates classes in a potentially high-dimensional feature space induced by a kernel function. In the non-linearly separable case, the Radial Basis Function (RBF) kernel implicitly maps inputs into an infinite-dimensional Hilbert space, enabling the construction of non-linear decision boundaries without explicitly computing the transformation. The regularisation parameter C controls the penalty assigned to misclassified training points, governing the bias-variance trade-off, while the kernel coefficient gamma determines the effective radius of influence of individual training samples. Both linear and RBF kernels were included in the search grid to evaluate whether the stunting classification problem requires non-linear separation (Azis et al., 2024).

5. K-Nearest Neighbors

KNN is a lazy, instance-based learner that defers all computation to inference time, classifying a query point based on a plurality vote among its k nearest neighbours in the training set. Its performance is highly sensitive to the choice of neighbourhood size k, the distance metric used to define proximity, and the weighting scheme applied to neighbouring votes. In this study, the grid search covered odd values of k from 3 to 11 to avoid tie-breaking ambiguity, uniform and distance-based weighting schemes, and Euclidean versus Manhattan distance metrics the latter being more robust to outliers in anthropometric measurements that may arise from field recording errors (Armando Sibuea et al., 2024).

6. Gradient Boosting

GB constructs an additive ensemble of weak learners typically shallow DTs in a sequential, stage-wise manner. At each iteration, a new tree is fitted to the negative gradient of the loss function with respect to the current ensemble's predictions, effectively performing gradient descent in function space. This mechanism allows the model to progressively correct residual

errors, capturing complex, higher-order interactions among anthropometric features that simpler models cannot represent. The three tuned hyperparameters number of estimators, learning rate, and maximum tree depth jointly control the ensemble's capacity, the contribution of each successive tree, and the risk of overfitting, requiring careful co-optimisation to achieve peak generalisation performance (Rahutomo et al., 2023).

D. Hyperparameter Tuning

Hyperparameter optimisation was performed for each of the six classifiers using scikit-learn's GridSearchCV with stratified k-fold cross-validation, ensuring proportional representation of the three stunting classes across all folds. A fold count of k = 5 was applied to RF, DT, KNN, GB, and LR, while k = 3 was used for SVM due to its quadratic kernel optimisation cost. Weighted accuracy was used as the scoring criterion to account for class imbalance (Fannany et al., 2024).

Search spaces were designed to cover a meaningful range of complexities for each model, as detailed in Table III. RF was tuned over estimator count, maximum depth, and minimum samples per split; DT over split criterion, depth, and leaf constraints; KNN over odd neighbour values from 3 to 11, weighting schemes, and distance metrics; GB over learning rate, estimator count, and tree depth; SVM over kernel type, regularisation strength, and gamma; and LR over penalty type and solver at max_iter = 5,000. The best configuration per model was evaluated on the held-out test set to obtain an unbiased estimate of generalisation performance.

Table 1. Hyperparameter Search Space Per Model

Model	Hyperparameter Search Space
Random Forest	n_estimators: [50,100,200], max_depth: [None,10,20], min_samples_split: [2,5,10]
DT	criterion: [gini,entropy], max_depth: [None,10,20,30], min_samples_split: [2,5,10], min_samples_leaf: [1,2,4]
KNN	n_neighbors: [3,5,7,9,11], weights: [uniform,distance], metric: [euclidean,manhattan]
Gradient Boosting	n_estimators: [50,100,200], learning_rate: [0.01,0.1,0.2], max_depth: [3,5,7]
SVM	C: [0.1,1,10], kernel: [linear,rbf], gamma: [scale,auto]
LR	C: [0.1,1,10], penalty: [l1,l2], solver: [liblinear,saga], max_iter: 5000

E. Evaluation Metrics

Four evaluation metrics were computed with weighted averaging to handle class imbalance across Normal, Stunting, and Severe categories. Accuracy measures the proportion of correctly classified instances overall, though it can be inflated in imbalanced settings, so it was reported alongside more informative metrics. Weighted Precision captures how reliably a model's positive predictions are correct on a per-class basis, penalising excessive false positives for minority classes. Weighted Recall reflects the model's ability to identify all true cases of each class a particularly important consideration here, since missing a severely stunted child carries far greater clinical consequences than an unnecessary referral. Finally, the weighted F1-Score, the harmonic mean of Precision and Recall, served as the primary summary metric. By penalising both false positives and false negatives simultaneously, it

provides a balanced measure of practical classification quality under real-world screening conditions.

RESULTS AND DISCUSSIONS

A. Baseline Model Performance

Table II summarises baseline performance for all six models on the held-out test set. RF came out on top with 97.04% accuracy and 97.05% F1-Score, a result that reflects how well bootstrap aggregation and random feature subsampling control variance in multi-class settings. GB followed at 95.38%, showing that its sequential boosting mechanism generalises reasonably well even without explicit tuning.

Interestingly, DT and SVM landed at almost identical accuracy 93.66% and 93.64% despite operating on entirely different principles. That convergence suggests the anthropometric feature space has enough inherent structure for both approaches to exploit. KNN was close behind at 93.50%, though its reliance on local density made it somewhat vulnerable to the wide age range (0–59 months) in the data.

LR sat well below the rest at 77.35% accuracy and 76.61% F1-Score a gap of nearly 20 percentage points from RF. That margin is hard to explain away as a tuning issue; it more likely reflects the fact that the boundaries separating Normal, Stunting, and Severe cases are genuinely non-linear, and no linear model can adequately capture them regardless of regularisation.

Table 2. Bestline Model Evaluation Results

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.970427	0.970533	0.970427	0.970453
DT	0.953831	0.954024	0.953831	0.953917
KNN	0.936611	0.936151	0.936611	0.936311
Gradient Boosting	0.936361	0.937587	0.936361	0.935754
SVM	0.934989	0.934981	0.934989	0.934295
LR	0.773521	0.767327	0.773521	0.766111

B. Tuned Model Performance

Table III presents after GridSearchCV optimisation, all six models improved but the more interesting story is the ranking shift at the top. GB, which trailed RF at baseline, emerged as the best overall model after tuning.

Table 3. Tuned Model Evaluation Results

Model	Accuracy	Precision	Recall	F1-Score
GB (Tuned)	0.975418	0.975454	0.975418	0.975412
RF (Tuned)	0.972174	0.972223	0.972174	0.972183
DT (Tuned)	0.953831	0.954024	0.953831	0.953917
KNN (Tuned)	0.950586	0.950503	0.950586	0.950538
SVM (Tuned)	0.942101	0.941818	0.942101	0.941883
LR (Tuned)	0.787746	0.780678	0.787746	0.756548

GB reached 97.54% accuracy and F1-Score after tuning a 2.16-point gain from baseline. The reversal over RF is telling: GB's sequential architecture holds more latent capacity, but only realises it when learning rate, tree depth, and estimator count are properly calibrated together. RF, being more self-regularising by design, needed less coaxing its tuned accuracy of 97.22% represents a solid but modest gain over an already strong baseline.

DT improved by 1.72 points to 95.38%, confirming the default unconstrained tree was overfitting. KNN gained 1.56 points to 95.06%, with a better neighbourhood size and distance metric making a tangible difference across the 0–59-month age range. SVM's improvement was smaller at 0.77 points, though the RBF kernel consistently outperformed linear again consistent with a non-linear classification boundary.

LR tells a different story. Its accuracy edged up to 78.77%, but its F1-Score dropped from 76.61% to 75.65%. That divergence suggests tuning pushed the model toward the dominant Normal class, trading minority-class sensitivity for marginal overall accuracy gains. It is a poor trade in a clinical context, and it underlines the core problem: no regularisation strategy can compensate for a model that is architecturally incapable of capturing non-linear decision boundaries.

C. Base vs. Tuned Accuracy Comparison

Figure. 3 plots baseline and tuned accuracy side by side for all six models. The pattern is clear: tuning helped every model, but not equally and in one case, it changed the ranking entirely.

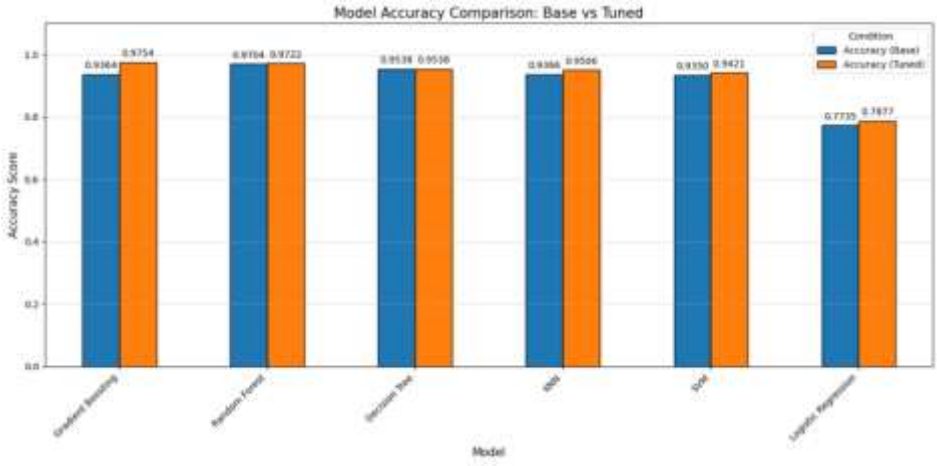

Figure 3. Model Accuracy Comparison

The standout observation is the GB–RF inversion. At baseline, RF led by 1.66 points (97.04% vs. 95.38%). After tuning, GB pulled ahead by 0.32 points (97.54% vs. 97.22%). That flip is not a statistical quirk it reflects a genuine architectural difference. RF's parallel bagging is self-stabilising, so it performs reasonably well out of the box. GB's sequential boosting, on the other hand, is more parameter-sensitive; once learning rate, tree depth, and ensemble size are properly set, it unlocks capacity that default settings leave on the table.

The middle-tier models saw the largest proportional gains. DT jumped 1.72 points to 95.38% the default unconstrained tree was clearly overfitting, and depth and leaf constraints fixed that cleanly. KNN gained 1.56 points to 95.06%, with a better neighbourhood size and distance metric making a real difference across the dataset's 0–59-month age range. SVM improved more modestly at 0.77 points, with the RBF kernel outperforming linear in every search iteration consistent with the non-linear boundary seen throughout.

LR is the outlier. Its accuracy nudged up from 77.35% to 78.77%, but its F1-Score actually fell from 76.61% to 75.65% the only model to go backwards on that metric. Tuning shifted predictions toward the Normal majority, buying marginal accuracy at the cost of minority-class balance. The gap between LR and the top performers remained at roughly 18–

19 points in both configurations. That kind of persistent deficit cannot be tuned away; it reflects a structural mismatch between a linear model and a classification boundary that is fundamentally non-linear.

Three things follow from these results. Tuning is worth doing for every model family, with gains between 0.32 and 1.72 points for the five non-linear models. Rankings are not fixed baseline benchmarking alone would have identified RF as the best model and missed the fact that GB overtakes it after optimisation. And for deployment, tuned GB at 97.54% F1-Score is the clear recommendation for Puskesmas-level stunting screening.

The results point to three consistent patterns: ensemble methods outperform linear classifiers, tuning matters more than often assumed, and the stunting classification boundary is genuinely non-linear each with practical consequences for deployment.

1. Superiority of Ensemble Methods

RF and GB dominated across all conditions, but for different reasons. RF's parallel bagging is inherently stable, delivering strong performance even with default settings. GB, by contrast, relies on sequential correction and is far more sensitive to how its parameters are configured which explains why it trailed RF at baseline (95.38% vs. 97.04%) but overtook it after tuning (97.54% vs. 97.22%). In practice, this means RF is the safer choice when tuning resources are limited, while GB is worth the extra effort when full optimisation is feasible.

2. Non-Linearity of the Classification Boundary

LR's consistent underperformance 78.77% accuracy and 75.65% F1-Score after tuning, roughly 19 points below GB is not a calibration issue. No amount of regularisation can fix the fact that stunting status depends on interactions between age, height, weight, and Z-scores that a linear model cannot capture. What makes this finding more telling is that LR's F1-Score dropped after tuning: the model became more biased toward the majority Normal class, gaining surface accuracy at the cost of sensitivity for Severe cases precisely the wrong trade-off in a clinical setting.

3. Clinical Implications for Minority Class Performance

In stunting screening, a missed Severe case is not equivalent to a false alarm. Missing a severely stunted child during the critical developmental window can have irreversible consequences, while an unnecessary follow-up is a manageable inconvenience. This asymmetry means aggregate accuracy is an insufficient selection criterion. Future work should explore class-weighted loss functions, SMOTE, or threshold calibration specifically targeting Severe class recall.

4. Value of Hyperparameter Optimisation

Tuning improved every model without exception, with gains ranging from 0.18 percentage points for RF to 2.16 for GB. The variation in improvement is itself informative: models with broader, more interactive search spaces benefited most, while self-regularising architectures like RF gained less. More importantly, the GB–RF ranking inversion would have been invisible under baseline-only benchmarking a reminder that comparing algorithms by their default configurations can lead to the wrong conclusions.

5. Positioning Within the Literature

The tuned GB F1-Score of 97.54% sits above the 80–95% range commonly reported in comparable studies. This likely reflects the combination of dataset scale (40,071 records), leakage-free preprocessing, and the breadth of the comparative framework. That said, cross-study comparisons should be treated with care differences in feature sets, class definitions, and evaluation protocols make direct comparisons unreliable without careful alignment.

6. Limitations and Future Directions

A few honest caveats apply. The dataset covers only Jeneponto Regency, so how well these models generalise to other regions with different demographic or nutritional profiles remains an open question. The feature set is also cross-sectional; adding longitudinal growth trajectories across multiple visits would likely improve predictive power, especially for borderline cases. Class imbalance was handled at the metric level through weighted averaging, but the training data itself remains skewed resampling strategies like SMOTE or ADASYN deserve systematic evaluation. Finally, none of the models currently produce explanations alongside their predictions. Integrating SHAP or LIME would be a meaningful step toward clinical trust, giving health workers a transparent basis for acting on the model's output rather than treating it as a black box.

CONCLUSIONS

This study compared six machine learning algorithms for stunting classification using 40,071 child anthropometric records from Jeneponto Regency collected between 2021 and 2024, covering three nutritional categories: Normal, Stunting, and Severe. Before training, the data went through careful preprocessing including removal of the Z-Score H/A column to prevent target leakage, StandardScaler normalisation, and stratified splitting to ensure the results would hold up on unseen data. In the baseline setting, RF came out on top with 97.04% accuracy, while LR trailed significantly at 77.35%. That roughly 20-point gap was not incidental; it reflects how poorly linear models handle the non-linear boundaries that separate stunting categories. After hyperparameter tuning with GridSearchCV, GB overtook RF with 97.54% accuracy a reversal that would have gone undetected had only default configurations been tested.

Three takeaways stand out. Ensemble models, particularly GB and RF, consistently outperformed LR by capturing complex interactions among age, weight, height, and Z-scores that a linear model simply cannot represent. Tuning made a real difference across all six models, with accuracy gains ranging from 0.18 to 2.16 percentage points. And LR's persistent weakness throughout confirms that for this kind of clinical classification task, non-linear model families are not optional they are necessary. Looking ahead, tuned GB shows genuine promise as a practical screening tool at the Puskesmas level. That said, several limitations remain open. Future work could benefit from incorporating SHAP-based explanations to support clinical trust, adding longitudinal growth data, addressing Severe class underrepresentation more directly, and exploring federated learning to allow deployment across Indonesia's distributed health infrastructure without centralising sensitive patient data.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to all Puskesmas staff for their cooperation and assistance in providing access to the nutritional data used in this study. The authors also thank Universitas Negeri Makassar for its support throughout the course of this research.

FUNDING

This research received no external funding.

AUTHOR CONTRIBUTIONS

Conceptualization, A.H.R.S., P., H.J. and A.A.; methodology, H.J.; validation, A.A.; formal analysis, A.H.R.S.; investigation, A.H.R.S.; data curation, A.H.R.S.; writing original draft preparation, A.H.R.S.; writing review and editing, P., H.J. and A.A.; supervision, P. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Akseer, N., Tasic, H., Nnachebe Onah, M., Wigle, J., Rajakumar, R., Sanchez-Hernandez, D., Akuoku, J., et al., Economic Costs of Childhood Stunting to the Private Sector in Low- and Middle-Income Countries, *EClinicalMedicine*, vol. **45**, p. 101320, from https://www.sciencedirect.com/science/article/pii/S2589537022000505, 2022. DOI: https://doi.org/10.1016/j.eclinm.2022.101320

Armando Sibuea, A. T., Harry Gunawan, P., and Indwiarti, Classifying Stunting Status in Toddlers Using K-Nearest Neighbor and Logistic Regression Analysis, *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 6–11, 2024.

Arqam, Mhd. L., Firdaus, A. A., Atmojo, A. M., Saputri, G. Z., Furizal, Palahuddin and Sirnopati, R., Integrating Education-Based Interventions and Machine Learning for Stunting Prevention: A Case Study in East Lombok, Indonesia, *Dialogues in Health*, vol. **8**, p. 100264, from https://www.sciencedirect.com/science/article/pii/S2772653325000620, 2026. DOI: https://doi.org/10.1016/j.dialog.2025.100264

Arya, P. K., Sur, K., Kundu, T., Dhote, S. and Singh, S. K., Unveiling Predictive Factors for Household-Level Stunting in India: A Machine Learning Approach Using NFHS-5 and Satellite-Driven Data, *Nutrition*, vol. **132**, p. 112674, from https://www.sciencedirect.com/science/article/pii/S089990072400323X, 2025. DOI: https://doi.org/10.1016/j.nut.2024.112674

Azis, D. M., Fauzi, R., and Suakanto, S., Development of Stunting Prediction Features to Prevent Stunting Using Support Vector Machine (SVM) Algorithm, *2024 International Conference on Digital Business and Technology Management (ICONDBTM)*, pp. 1–6, 2024.

Bukit, D. S., Lydia, M. S., Nainggolan, P. I., Mahmud, H. I., Sembiring, R. M., Hasibuan, R. M., Athirah, D., Mufida, F. M., and Rahmat, R. W. B. O. K., Leveraging Machine Learning Techniques for Stunting Detection and Height Growth Prediction in Children Aged 0-5 Years, *2024 8th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, pp. 130–34, 2024.

Erda, G., Yolanda, A. M., Adnan, A., Alika, E. R., and Erda, Z., Advanced Machine Learning Techniques for Stunting Classification Using a Stacking Ensemble Approach, *2024 4th International Conference on Electrical Engineering and Informatics (ICon EEI)*, pp. 79–84, 2024.

Fadhilah, D. N., and Gunawan, P. H., Support Vector Machine-Based Classification of Toddler Stunting in Bandarharjo, *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pp. 263–67, 2024.

Fannany, C., Gunawan, P. H., and Aquarini, N., Machine Learning Classification Analysis for Proactive Prevention of Child Stunting in Bojongsoang: A Comparative Study, *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 1–5, 2024.

Firdausi, L., and Gunawan, P. H., Stunting Analysis of Toddlers in Kota Baru, West Bekasi Using K-Nearest Neighbor and Naive Bayes, *2025 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 400–405, 2025.

Gunawan, R., Pratama, R., Impron, A., Rahmatulloh, A., Darmawan, I., and Rizal, R., Optimization of the Support Vector Machine Method with Forward Selection for Stunting Disease Detection, *2025 Tenth International Conference on Informatics and Computing (ICIC)*, pp. 1–6, 2025.

Hendy, A., Abdelaliem, S. M. F., Sultan, H. M., Alahmedi, S. H., Ibrahim, R. K., Abdelrazek, E. M. E., Elmahdy, M. A. A. and Hendy, A., Unlocking Insights: Using Machine Learning to Identify Wasting and Risk Factors in Egyptian Children under 5, *Nutrition*, vol. **131**, p. 112631, from https://www.sciencedirect.com/science/article/pii/S0899900724002806, 2025. DOI: https://doi.org/10.1016/j.nut.2024.112631

Janssen, S. M. W., Bouzembrak, Y., Yalcin, N. and Tekinerdogan, B., Machine Learning Models for Predicting Malnutrition in NICU Patients: A Comprehensive Benchmarking Study, *Computers in Biology and Medicine*, vol. **192**, p. 110326, from https://www.sciencedirect.com/science/article/pii/S0010482525006778, 2025. DOI: https://doi.org/10.1016/j.compbiomed.2025.110326

Kani, R., and Gunawan, P. H., Classification of Stunting in Toddlers from Bandarharjo Using K-Nearest Neighbors and Random Forest Algorithms, *2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pp. 258–62, 2024.

Kanz, A. F., Johanes, L. P., Alam, I. N., and Wulandhari, L. A., Stunting Prediction in Children Using Random Forest Algorithm, *2024 6th International Conference on Cybernetics and Intelligent System (ICORIS)*, pp. 1–4, 2024.

Khoirunnisa, K., and Gunawan, P. H., Analysis of Stunting Prediction for Toddlers in Bekasi Regency Using the K-Nearest Neighbors and Random Forest Algorithms, *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, pp. 932–36, 2024.

Lestari, A. D., Harry Gunawan, P., and Darwiyanto, E., Comparison of Naive Bayes and Support Vector Machine Performance in Classification of Child Stunting Status in Pemalang Regency, *2025 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pp. 545–51, 2025.

Mgomezulu, W. R., Thangata, P., Mkandawire, B. and Amoah, N., Advancing Predictive Analytics in Child Malnutrition: Machine, Ensemble and Deep Learning Models with Balanced Class Distribution for Early Detection of Stunting and Wasting, *Human Nutrition & Metabolism*, vol. **42**, p. 200340, from https://www.sciencedirect.com/science/article/pii/S2666149725000441, 2025. DOI: https://doi.org/10.1016/j.hnm.2025.200340

Miranda, E., Aryuni, M., Zakiyyah, A. Y., Kurniawati, Y. E., Sano, A. V. D. and Kumbangsila, M., An Early Prediction Model for Toddler Nutrition Based on Machine Learning from Imbalanced Data, *Procedia Computer Science*, vol. **245**, pp. 263–71, from https://www.sciencedirect.com/science/article/pii/S187705092403059X, 2024. DOI: https://doi.org/10.1016/j.procs.2024.10.251

Nduwayezu, G., Zhao, P., Pilesjö, P., Bizimana, J. P. and Mansourian, A., Multilevel Small-Area Childhood Stunting Risk Estimation: Insights from Spatial Ensemble Learning, Agro-Ecological and Environmentally Remotely Sensed Indicators, *Environmental and Sustainability Indicators*, vol. **27**, p. 100822, from https://www.sciencedirect.com/science/article/pii/S2665972725002430, 2025. DOI: https://doi.org/10.1016/j.indic.2025.100822

Nur, Y. S. R., Aldo, D., Sa'Adah, A., and Faizah, Implementation of PC Algorithm to the Incidence Factor of Stunting Disease, *2024 International Conference on Information Technology Research and Innovation (ICITRI)*, pp. 93–98, 2024.

Prabiantissa, C. N., Yamani, L. N., Hakimah, M., Puspitasari, I., and Rozi, N. F., Implementation of Artificial Neural Network (ANN) to Construct Model for Stunting in Toddlers, *2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, pp. 1–5, 2024.

Rahutomo, R., Elwirehardja, G. N., Isnan, M., Asadi, F., and Pardamean, B., Machine Learning Implementations in Childhood Stunting Research: A Systematic Literature Review, *2023 International Conference on Information Management and Technology (ICIMTech)*, pp. 229–34, 2023.

Sadler, K., James, P. T., Bhutta, Z. A., Briend, A., Isanaka, S., Mertens, A., Myatt, M., et al., How Can Nutrition Research Better Reflect the Relationship Between Wasting and Stunting in Children? Learnings from the Wasting and Stunting Project, *The Journal of Nutrition*, vol. **152**, no. 12, pp. 2645–51, from https://www.sciencedirect.com/science/article/pii/S0022316623086522, 2022. DOI: https://doi.org/10.1093/jn/nxac091

Shevaldo, G., and Gunawan, P. H., Improving Stunting Detection in Toddlers with Boosted KNN and Boosted Naïve Bayes Techniques, *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, pp. 326–31, 2024.

Shi, H., Yang, D., Tang, K., Hu, C., Li, L., Zhang, L., Gong, T. and Cui, Y., Explainable Machine Learning Model for Predicting the Occurrence of Postoperative Malnutrition in Children with Congenital Heart Disease, *Clinical Nutrition*, vol. **41**, no. 1, pp. 202–10, from https://www.sciencedirect.com/science/article/pii/S0261561421005070, 2022. DOI: https://doi.org/10.1016/j.clnu.2021.11.006

Sudigyo, D., Hidayat, A. A., Nirwantono, R., Rahutomo, R., Trinugroho, J. P. and Pardamean, B., Literature Study of Stunting Supplementation in Indonesian Utilizing Text Mining Approach, *Procedia Computer Science*, vol. **216**, pp. 722–29, from https://www.sciencedirect.com/science/article/pii/S1877050922022670, 2023. DOI: https://doi.org/10.1016/j.procs.2022.12.189

Sundjaya, T., Djuwita, R., Adisasmita, A. C., Tanjung, C., Massi, N., Fikri, B., Pradnyaparamitha, D. A. and Basrowi, R. W., Gut Microbiome Changes among

Undernutrition and Stunting Infants and Children under 2 Years: A Scoping Review, *The Open Public Health Journal*, vol. **17**, from https://www.sciencedirect.com/science/article/pii/S1874944524001138, 2024. DOI: https://doi.org/10.2174/0118749445319116240729045056

Yosep, I., Kurniawan, K., Rafiyah, I., Ramdhani, M. R. and Hikmat, R., The Need for Chatbot-Based Emotional Counseling for Parents with Stunting Children: A Qualitative Descriptive Study, *International Journal of Africa Nursing Sciences*, p. 101076, from https://www.sciencedirect.com/science/article/pii/S2214139126001034, 2026. DOI: https://doi.org/10.1016/j.ijans.2026.101076

Yulianto, S. A., Solimun, S., Efendi, A., Alim, V. I. A., Jauhar, H. S. Al, Rejeki, S. W. S. and Rinaldo Fernandes, A. A., Predicting Nutritional and Physical Stunting in Malang District:, *International Journal of Reliable and Quality E-Healthcare*, vol. **14**, no. 1, from https://www.sciencedirect.com/science/article/pii/S2160955125000051, 2025. DOI: https://doi.org/10.4018/IJRQEH.395752