UNS UNIVERSITAS SEBELAS MARET
PLACENTUM
Laman resmi: http://jurnal.uns.ac.id/placentum
ISSN 2303-3746
e-ISSN 2620-9969
PLACENTUM Jurnal Ilmiah Kesehatan dan Aplikasinya

Stunting Classification and Prediction in Children Under Five: A Systematic Review of Machine Learning-Based Studies

Fanny Kartika Fajriyani[1], Ropitasari[2], Riska Fajar Fatony[3], Nahdiyah Karimah[4]

[1] Faculty of Psychology and Health, Walisongo State Islamic University, Semarang, Indonesia

[2,4]D3 Midwifery, Vocational School, Universitas Sebelas Maret, Surakarta, Indonesia

[3]Health Information Management Program, Faculty of Health, Politeknik Negeri Jember, Jember, Indonesia

** Corresponding author*

E-mail: fanny.kartika.fajriyani@walisongo.ac.id

ABSTRACT

Background: Childhood stunting remains a major public health problem, particularly in low- and middle-income countries. Machine learning may support the identification and prediction of stunting by integrating multidimensional information related to maternal, child, household, socioeconomic, environmental, and nutritional factors. This systematic review aimed to synthesize evidence on machine learning-based models for stunting classification and prediction among children under five, including predictor domains, algorithm performance, model interpretability, and potential applications in child health programs.
Method: A systematic review was conducted according to PRISMA 2020 guidelines. Studies published from January 2021 to January 2026 were searched in PubMed, Scopus, Google Scholar, Web of Science, and ScienceDirect. Eligible studies involved children under five, applied machine learning or deep learning algorithms to stunting-related outcomes, and reported at least one model-performance measure. Data were synthesized narratively, and methodological quality and applicability were assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST).
Results: Fifteen studies met the inclusion criteria. Frequently reported predictor domains included maternal education, household socioeconomic status, birth weight, child age, maternal characteristics, sanitation, birth-related factors, and geographical conditions. Random Forest, Gradient Boosting, and Extreme Gradient Boosting were commonly evaluated and demonstrated competitive performance across different datasets. However, comparisons were limited by heterogeneity in populations, predictors, preprocessing procedures, outcome definitions, validation strategies, and performance metrics.
Conclusion: Machine learning-based models show potential for identifying and predicting childhood stunting. Further research should emphasize clear outcome definitions, external validation, transparent reporting, and evaluation in routine child health services before widespread implementation.

Keywords: stunting; classification; prediction; machine learning; children under five; systematic review

INTRODUCTION

Stunting remains a major public health problem among children under five, particularly in low- and middle-income countries (LMICs). It is defined as a height-for-age Z-score below −2 standard deviations according to the World Health Organization (WHO) Child Growth Standards and reflects the cumulative effects of chronic undernutrition, recurrent infections, inadequate caregiving, and adverse environmental conditions during early life.[1,4,8] Approximately 148 million children under five were affected by stunted growth globally in 2024, indicating that progress toward Sustainable Development Goal (SDG) 2.2 remains insufficient.[6,30] Beyond impaired linear growth, stunting is associated with adverse cognitive, immune, educational,

and economic outcomes, with potential consequences extending across the life course.[3,8,29,35]

The development of stunting is multifactorial, involving interactions among maternal, child, household, environmental, and socioeconomic factors. Maternal nutritional status, birth weight, infant feeding practices, infectious diseases, household food insecurity, sanitation, environmental conditions, and socioeconomic circumstances may jointly influence child growth trajectories.[8,13–15] Because growth faltering can develop gradually before becoming clinically apparent, identifying children only after growth impairment has occurred may limit opportunities for preventive intervention.[1,8] Therefore, approaches that can identify children at increased risk before substantial growth deficits are established are important for strengthening early prevention and child health programs.

Current stunting surveillance relies largely on anthropometric assessment, cross-sectional surveys, and conventional epidemiological analyses. These approaches are essential for monitoring nutritional status and population-level burden but are primarily oriented toward identifying existing growth deficits rather than estimating future risk.[4,6] Conventional statistical models may also have limitations in representing complex and nonlinear relationships among the multiple determinants of child growth.[7,8] These challenges have increased interest in machine learning (ML), which can incorporate multidimensional data and identify complex relationships among predictors. Various ML algorithms, including Random Forest, Extreme Gradient Boosting (XGBoost), Support Vector Machine, and neural network-based approaches, have been applied to maternal, child health, and nutritional outcomes.[2,9,20] The increasing availability of routinely collected demographic, anthropometric, clinical, environmental, and socioeconomic data further provides opportunities for developing risk prediction models to support preventive strategies.[2,10,18]

However, evidence on ML-based prediction of childhood stunting remains heterogeneous. Existing studies differ in population characteristics, data sources, predictor selection, algorithms, validation strategies, and performance metrics, limiting direct comparison of reported model performance. Predictor sets also vary considerably, ranging from child and maternal characteristics to household socioeconomic conditions, birth history, sanitation, feeding practices, and geographical factors.[2,8,16] In addition, predictive performance alone does not fully determine the usefulness of a model for child health practice. Model interpretability, explainability, methodological quality, validation, and feasibility of integration into routine child health services are also important considerations for translating predictive models into practice.[17,18,23]

Although previous reviews have examined artificial intelligence and ML applications in healthcare and neonatal medicine more broadly [9,19,20], a focused synthesis of ML approaches for early prediction of stunting among children under five remains needed. In particular, evidence has not been systematically brought together across predictor domains, algorithm performance, model explainability, and considerations for implementation within child health programs. Addressing these dimensions may help clarify which predictors are repeatedly identified across studies, how reported algorithm performance varies across methodological settings, and what factors should be considered when translating prediction models into public health practice.

Therefore, this systematic review aimed to synthesize evidence on ML approaches for the early prediction of stunting among children under five.

Specifically, the review examined the predictor domains used in prediction models, compared the reported performance of ML algorithms, assessed model interpretability and explainability, and considered the implications of these models for integration into child health programs. The synthesis also considered methodological challenges and areas requiring further validation to support the development and implementation of clinically and publicly relevant prediction models.

METHODS

This systematic review was conducted in January 2026. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines and aimed to synthesize evidence on machine learning-based classification and prediction of stunting among children under five years of age. The review focused on predictor domains, machine learning algorithms, model performance, explainability approaches, and potential applications in child health programs.

A comprehensive literature search was conducted in January 2026 using PubMed, Scopus, Google Scholar, Web of Science, and ScienceDirect. The search strategy combined terms related to stunting, children under five, and machine learning using Boolean operators AND and OR. The main terms included "stunting," "linear growth faltering," "height-for-age," "child," "children," "under-five," "infant," "toddler," "machine learning," "artificial intelligence," "deep learning," "predictive model," "prediction," "classification," "Random Forest," "XGBoost," "Support Vector Machine," "Gradient Boosting," and "Artificial Neural Network." The search strategy was adapted to the indexing system and syntax of each database. The complete database-specific search strategies, search dates, and search results are provided in Supplementary Table S1. Reference lists of eligible studies and relevant review articles were also manually screened.

Studies were eligible if they were peer-reviewed original research articles published in English between January 2021 and January 2026, involved children under five years of age, investigated childhood stunting or a directly relevant linear growth outcome, applied machine learning or deep learning algorithms for classification or prediction, and reported at least one model-performance measure, including accuracy, sensitivity, specificity, precision, F1-score, area under the receiver operating characteristic curve (AUC), or calibration. Studies examining general malnutrition were included only when stunting outcomes were reported separately. Reviews, conference abstracts, editorials, letters, commentaries, protocols, animal studies, and studies without adequate information on model development, validation, or predictive performance were excluded.

Retrieved records were exported to reference management software, and duplicates were removed. Two reviewers independently screened titles, abstracts, and full-text articles. Disagreements were resolved through discussion, with adjudication by a third reviewer when necessary. Data were extracted using a standardized form covering author, publication year, country, study design, data source, sample size, participant characteristics, outcome definition, predictor variables, preprocessing, algorithms, validation methods, performance metrics, explainability techniques, and potential programmatic applications.

Risk of bias and applicability were independently assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST), covering participants, predictors, outcomes, and statistical analysis. Studies with high or unclear risk of bias were retained but interpreted cautiously. Because of

substantial heterogeneity in populations, datasets, predictor sets, algorithms, preprocessing procedures, validation strategies, outcome definitions, and performance metrics, quantitative meta-analysis was not performed. Findings were synthesized narratively according to three predefined dimensions: machine learning algorithms, predictor domains, and explainability approaches. Predictors were categorized into maternal, child, household socioeconomic, environmental/geographical, birth-related, nutritional, and healthcare-related domains. Explainability methods such as SHAP and feature-importance analysis were interpreted as approaches to improve model transparency rather than as evidence of causal relationships. No power calculation was required because the review involved previously published literature and no primary data collection. Ethical approval and informed consent were not required because no human participants or identifiable personal data were directly involved.

RESULT

The study selection process was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. A total of 703 records were identified through a comprehensive search of five electronic databases, including Google Scholar (n = 467), PubMed (n = 106), Scopus (n = 50), ScienceDirect (n = 53), and Web of Science (n = 27). After removing 18 duplicate records, 685 unique records remained for title and abstract screening (Figure 1).

During the initial screening, 621 records were excluded because they did not meet the predefined eligibility criteria. The main reasons for exclusion included studies that were not relevant to machine learning-based prediction of childhood stunting (n = 177), studies that did not apply machine learning algorithms (n = 180), studies involving populations other than children under five years of age (n = 105), and publications categorized as review articles, conference proceedings, editorials, letters, or study protocols (n = 159). Following title and abstract screening, 64 full-text articles were assessed for eligibility. After full-text evaluation, 49 studies were excluded for several reasons, including failure to develop or validate a machine learning prediction model (n = 16), use of conventional statistical analyses without machine learning approaches (n = 18), evaluation of outcomes other than childhood stunting, such as wasting, underweight, obesity, or neonatal outcomes (n = 9), and insufficient reporting of machine learning methodology or model performance metrics (n = 6)

Finally, 15 studies fulfilled all eligibility criteria and were included in the qualitative synthesis. These studies investigated a wide range of machine learning algorithms, including conventional classifiers, ensemble learning methods, deep learning models, and explainable artificial intelligence approaches, for the early prediction of stunting among children under five years of age. Collectively, the included studies provided comprehensive evidence regarding the role of maternal, child, household socioeconomic, environmental, and nutritional predictors in improving early identification of children at risk of stunting and supporting data-driven child health programs. Risk of bias and applicability were assessed using PROBAST across the domains of participants, predictors, outcomes, and statistical analysis. The assessment results are presented in Table 2.

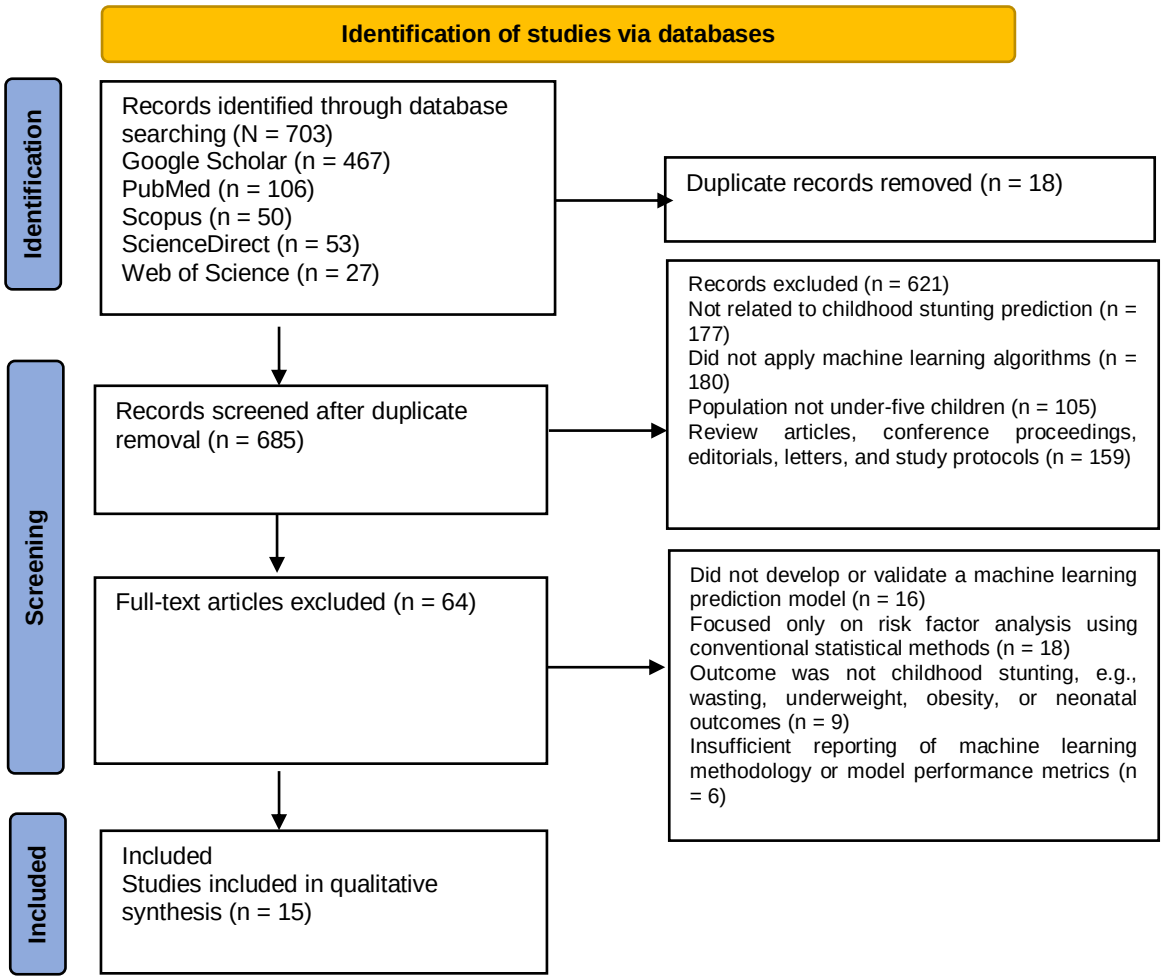

Figure 1. Prisma Flow Diagram

The main concerns involved participant selection, predictor measurement, outcome definition, missing-data handling, model development, overfitting control, and model validation. All eligible studies were retained in the qualitative synthesis. Due to differences in study populations, datasets, predictors, algorithms, validation procedures, and performance metrics, the findings were synthesized narratively rather than through quantitative meta-analysis.

Table 1. Risk of Bias and Applicability Assessment Using PROBAST

No.	Study	Participants	Predictors	Outcome	Analysis	Overall RoB	Applicability
1	Khan et al. (2021)	Low	Low	Low	High	High	Low
2	Shen et al. (2023)	Low	Low	Low	Low	Low	Low
3	Novalina et al. (2025)	Low	Low	Low	Unclear	Unclear	Low
4	Mkungudza et al. (2024)	Low	Low	Low	Unclear	Unclear	Low
5	Ndagijimana et al. (2023)	Low	Low	Low	Low	Low	Low
6	Hendy et al. (2025)	Low	Low	Low	Low	Low	Low
7	Mgomezulu et al. (2025)	Unclear	Low	Low	High	High	Low
8	Wicaksono et al. (2025)	Low	Low	Low	Unclear	Unclear	High
9	Anhar et al. (2025)	High	Unclear	Unclear	High	High	High
10	Mulyani et al. (2025)	Low	Low	Low	Unclear	Unclear	High
11	Sutarmi et al. (2023)	Unclear	Low	Unclear	High	High	High
12	Sahamony et al. (2024)	Low	Low	Low	Unclear	Unclear	Low
13	Ayele et al. (2025)	Low	Low	Low	Low	Low	Low
14	Anku & Duah (2024)	Low	Low	Low	Unclear	Unclear	Low
15	Islam et al. (2024)	Low	Low	Low	Low	Low	Low

Table 2. Characteristics of Included Studies on Machine Learning-Based Prediction of Childhood Undernutrition

No.	Study	Study Design and Dataset	Sample Size	Outcome	Algorithms Evaluated	Main Findings
1	Khan et al. (2021) Country: Bangladesh	Retrospective secondary data analysis using the Bangladesh Demographic and Health Survey (BDHS) 2014.	6,044 children	Stunting	Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and other algorithms	Child, maternal, household, and socioeconomic characteristics were used to predict stunting.
2	Shen et al. (2023) Country: Papua New Guinea.	Cross-sectional secondary data analysis using a national survey dataset from Papua New Guinea.	3,380 children	Stunting	Random Forest, XGBoost, Decision Tree, and other algorithms	Child age, sex, anthropometric characteristics, maternal factors, and household conditions were relevant predictors.
3	Novalina et al. (2025) Country: Indonesia.	Retrospective secondary data analysis using the Indonesian Family Life Survey (IFLS).	1,561 children	Stunting	Logistic Regression, Random Forest, XGBoost, Support Vector Machine, and other algorithms	Demographic, socioeconomic, maternal, and child-related variables contributed to stunting prediction.
4	Mkungudza	Cross-sectional secondary data	5149 children	Undernutr	Random Forest, Decision	Household socioeconomic status,

No.	Study	Study Design and Dataset	Sample Size	Outcome	Algorithms Evaluated	Main Findings
	et al. (2024) Country: Malawi.	analysis using the Malawi Demographic and Health Survey (MDHS).		ition	Tree, Logistic Regression, and other algorithms	maternal characteristics, child characteristics, and environmental factors were considered important predictors.
5	Ndagijimana et al. (2023) Country: Rwanda.	Cross-sectional secondary data analysis using a national survey dataset from Rwanda.	3,814 children	Stunting	Random Forest, XGBoost, Decision Tree, and other algorithms	Child age, sex, maternal education, household wealth, residence, and access to health services were influential predictors.
6	Hendy et al. (2025) Country: Egypt.	Retrospective secondary data analysis using repeated Demographic and Health Survey datasets from Egypt in 2005, 2008, and 2014.	NR	Stunting and undernutrition	Random Forest, XGBoost, Logistic Regression, and other algorithms	Predictor importance varied according to survey year, nutritional outcome, and available covariates.
7	Mgomezulu et al. (2025) Country: Malawi.	Secondary analysis of household survey data from the Malawi Living Standards Measurement Study (LSMS).	15,638 observations; 5,941 for stunting and 6,155 for wasting in the training data	Stunting and wasting	Random Forest, XGBoost, Decision Tree, and other algorithms	Ensemble models demonstrated competitive performance. Child and household characteristics were important predictors.
8	Wicaksono et al. (2025) Country: Indonesia.	Retrospective observational study using routine health records from Jeneponto, Indonesia.	40,071 children	Stunting	Random Forest, XGBoost, Decision Tree, and other algorithms	Routine anthropometric and demographic data were used to identify children at risk of stunting.
9	Anhar et al. (2025) Country: Indonesia.	Secondary data analysis using a publicly available Kaggle dataset.	10,000 records	Stunting	Random Forest, XGBoost, Logistic Regression, and other algorithms	Child age, sex, anthropometric variables, and household characteristics were used as predictors.
10	Mulyani et al. (2025) Country: Indonesia.	Retrospective secondary data analysis using local health data and a Kaggle dataset.	121,043 records	Stunting	Random Forest, XGBoost, Decision Tree, and other algorithms	Large-scale data and ensemble algorithms supported high predictive performance, although results depended on the dataset and validation procedure.

No.	Study	Study Design and Dataset	Sample Size	Outcome	Algorithms Evaluated	Main Findings
11	Sutarmi et al. (2023) Country: Indonesia.	Observational analytical study using a primary dataset collected from infants in Indonesia.	192 infants	Stunting	Random Forest, Support Vector Machine, Naive Bayes, and other algorithms	Machine learning was applied to classify infant stunting status. The reported class totals require verification against the original dataset.
12	Sahamony et al. (2024) Country: Indonesia.	Retrospective observational study using child health records from Lubuk Linggau, Indonesia, collected in 2023.	400 children	Stunting	Naive Bayes and Random Forest	Naive Bayes and Random Forest showed strong classification performance using routinely collected child health data.
13	Ayele et al. (2025) Country: Ethiopia.	Cross-sectional secondary data analysis using the Ethiopian Demographic and Health Survey (EDHS).	18,451 observations before resampling; 33,495 after SMOTE	Stunting and other forms of undernutrition	Random Forest, XGBoost, Logistic Regression, Decision Tree, and other algorithms	Feature selection and SMOTE were used to address class imbalance. Ensemble methods demonstrated competitive predictive performance.
14	Anku and Duah (2024) Country: Ghana.	Cross-sectional secondary data analysis using the Ghana Demographic and Health Survey (GDHS).	8,564 children	Stunting and other forms of undernutrition	XGBoost, Random Forest, Logistic Regression, Decision Tree, and other algorithms	XGBoost demonstrated strong predictive performance. Child, maternal, household, and socioeconomic characteristics were important predictors.
15	Islam et al. (2024) Country: Bangladesh.	Cross-sectional secondary data analysis using the Bangladesh Demographic and Health Survey (BDHS) 2017–2018.	7,796 for stunting; 7,777 for wasting; and 7,838 for underweight	Stunting, wasting, and underweight	Random Forest, XGBoost, Logistic Regression, Decision Tree, and other algorithms	Predictor importance and model performance differed across stunting, wasting, and underweight outcomes.

NR = Not Reported

DISCUSSION

The present systematic review synthesized evidence from fifteen studies investigating machine learning (ML) approaches for the early prediction of stunting among children under five years of age. Overall, the findings indicate that ML has considerable potential to improve the early identification of children at risk of stunting by integrating multidimensional predictors encompassing biological, maternal, socioeconomic, environmental, and healthcare-related factors. Childhood growth faltering is influenced by multiple interacting determinants operating from pregnancy through early childhood, supporting the need for prediction approaches that consider information beyond anthropometric measurements alone.[1,3,4,8] Recent advances in artificial intelligence and precision nutrition further indicate that data-driven approaches may strengthen maternal and child health interventions, particularly in low-resource settings.[2] However, these findings should be interpreted in the context of substantial methodological heterogeneity across the included studies. Differences in study populations, sample sizes, predictor selection, feature engineering, class imbalance handling, and validation strategies may have contributed to the variation in reported predictive performance and limit direct comparisons between algorithms.

One of the most consistent findings across the included studies is that childhood stunting cannot be adequately explained by a single determinant. Instead, prediction models incorporated variables from multiple domains, including maternal characteristics, child characteristics, household socioeconomic conditions, birth history, environmental factors, and geographical context. Maternal education, household wealth, birth weight or birth size, child age, maternal nutritional status, sanitation, birth interval, and geographical residence were repeatedly identified as important predictors across different studies. These findings are consistent with evidence demonstrating that early-life growth is shaped by complex biological, nutritional, environmental, and socioeconomic processes.[1,3,8]

The importance of multidimensional predictors is also supported by studies applying ML specifically to stunting prediction. Khan et al. demonstrated that ML-based variable selection could identify important predictors of childhood stunting in Bangladesh, highlighting the potential of ML to accommodate complex predictor structures in population-based data.[5] Similarly, studies conducted in Papua New Guinea, Rwanda, Malawi, Indonesia, and Egypt identified combinations of maternal, child, socioeconomic, and environmental variables as relevant predictors of stunting.[11,22,25,27]

This multidimensional pattern is particularly relevant because early childhood growth is affected by both prenatal and postnatal exposures. Birth-related characteristics may reflect intrauterine growth and maternal conditions, whereas household socioeconomic circumstances, sanitation, nutrition, and access to healthcare may influence subsequent growth trajectories.[1,8] In addition, evidence concerning food insecurity and food price inflation indicates that broader socioeconomic conditions can influence child undernutrition through household food access and dietary quality.[13]

Nevertheless, the apparent importance of individual predictors should be interpreted cautiously because predictor availability and selection differed considerably between studies. Some models were developed using nationally representative demographic or health surveys, whereas others used community-based or regional datasets. Consequently, a predictor identified as highly influential in one population may

not have the same predictive contribution in another population with different socioeconomic, environmental, or healthcare characteristics.

Differences in predictor selection may also influence model performance independently of the algorithm used. Studies incorporating a broader range of maternal, child, socioeconomic, and environmental variables may provide the algorithm with more information to discriminate between children at higher and lower risk. Conversely, models developed using a more limited predictor set may demonstrate lower apparent predictive performance but potentially greater feasibility for implementation in routine community settings. Therefore, higher predictive performance should not automatically be interpreted as evidence of a superior algorithm when the underlying predictor sets differ substantially.[24,25]

Maternal education and household socioeconomic status were among the frequently reported predictors. These factors may operate through multiple pathways involving food security, healthcare utilization, hygiene, caregiving practices, and health literacy. Birth-related characteristics, particularly birth weight and birth size, also contributed substantially to prediction in several studies, reflecting the importance of prenatal and early-life conditions in subsequent growth trajectories.[1,8] Environmental variables, particularly sanitation and geographical residence, further emphasize the contribution of living conditions and contextual factors to stunting risk.[11,22]

Collectively, these findings reinforce that prediction of childhood stunting should consider multidimensional information rather than relying exclusively on anthropometric indicators. However, future prediction studies should provide greater consistency in defining predictor domains and clearly justify predictor selection to facilitate meaningful comparison across populations and models.

Another important finding was the frequently superior reported performance of ensemble learning algorithms, particularly Gradient Boosting, Random Forest, and XGBoost, compared with conventional statistical approaches. Gradient Boosting demonstrated strong performance in the Bangladesh study, while XGBoost showed favorable performance in studies from Papua New Guinea and Indonesia.[5,11,22] Random Forest also performed consistently well across several datasets. These findings are broadly consistent with the capacity of ensemble and other advanced ML approaches to model nonlinear relationships and complex interactions within high-dimensional biomedical data.[20]

However, the apparent advantage of ensemble algorithms should be interpreted in light of methodological heterogeneity among the included studies. Predictive performance is determined not only by algorithm selection but also by the characteristics of the population, sample size, predictor set, preprocessing procedures, feature selection, class distribution, and validation strategy. Thus, differences in reported performance cannot be attributed solely to differences between algorithms.

Study population is an important source of this heterogeneity. The included studies were conducted in Bangladesh, Papua New Guinea, Indonesia, Malawi, Rwanda, and Egypt and used different types of data sources, including national demographic surveys, national health surveys, community-based datasets, and regional datasets.[5,11,22,24,25,27] These populations differ in socioeconomic conditions, healthcare access, environmental exposures, and distributions of maternal and child characteristics. An algorithm may therefore achieve high discrimination in one population because the predictors

have strong discriminatory patterns in that setting, but its performance may decline when applied to another population with different predictor distributions.

This issue is particularly important for low- and middle-income countries, where AI applications may be affected by differences in data availability, digital infrastructure, health information systems, and population characteristics.[18,19] Therefore, the predictive performance reported by individual studies should not automatically be assumed to represent the expected performance of the same algorithm in other populations.

Sample size may also have influenced the reported performance. The included studies ranged from relatively small community-based datasets to large population-based datasets, including a study from Indonesia involving 40,071 children.[22] Larger datasets may provide greater statistical information for model development and allow more complex algorithms to learn patterns from heterogeneous observations. Conversely, models developed from smaller datasets may be more susceptible to overfitting, particularly when the number of predictors is large relative to the sample size. Evidence from ML applications to longitudinal biomedical data similarly indicates that model robustness is closely related to the characteristics and structure of the available data.[16,20]

Therefore, reported performance should be considered together with sample size and model complexity rather than interpreted as an independent characteristic of the algorithm.

Feature selection and feature engineering represent important sources of variation among the included studies. Several studies applied approaches such as LASSO, recursive feature elimination, stepwise selection, SMOTE, or undersampling.[5,11,22,25] For example, the Papua New Guinea study combined LASSO with XGBoost, whereas the Indonesian study by Novalina et al. evaluated different approaches to class imbalance, including SMOTE and undersampling.[11,22] The Malawi study also specifically investigated alternative variable-selection approaches for stunting prediction.[25]

These procedures can substantially modify the information available to the learning algorithm. Feature selection can reduce dimensionality and remove potentially redundant variables, whereas feature engineering can transform or combine predictors to improve their representation. Consequently, differences in predictive performance may partly reflect preprocessing and feature engineering rather than algorithmic superiority alone.

This consideration is important when comparing models across studies. Two algorithms evaluated in different studies may effectively be solving different prediction problems because they are trained using different predictor sets and preprocessing pipelines. For example, an XGBoost model trained after systematic variable selection and transformation cannot be directly compared with a Random Forest model trained using a larger unselected predictor set. Such comparisons may overestimate the contribution of the algorithm itself.

Class imbalance represents another important source of heterogeneity. The distribution of children with and without stunting can affect classification performance, particularly when the outcome of interest represents a minority class. Some included studies addressed this problem using SMOTE or undersampling.[22] These approaches modify the effective distribution of observations available during model training and may improve the identification of minority-class observations.

However, performance obtained after resampling should not be interpreted as directly equivalent to performance obtained from models trained on the

original class distribution. Resampling can influence sensitivity, specificity, precision, and other classification metrics. Therefore, studies should clearly report the prevalence of stunting, the method used to address class imbalance, and whether resampling was performed only within the training dataset.

This issue also has implications for reproducibility. Without clear reporting of preprocessing, feature engineering, and imbalance-handling procedures, it is difficult to determine whether differences in model performance are attributable to algorithm selection or to differences in the analytical pipeline. Future studies should therefore provide transparent descriptions of all preprocessing procedures and ensure that these procedures are appropriately separated between model development and validation datasets.

Validation methodology represents another factor that may substantially influence reported predictive performance. Differences in data partitioning, cross-validation procedures, and external validation can affect the degree to which reported performance reflects true model generalizability. Methodological reviews of ML applications in healthcare emphasize the importance of robust evaluation and transparent reporting when translating predictive models into clinical or public health settings.[17,18,20]

Performance estimated from a single training-test split may be more sensitive to the characteristics of a particular sample than performance obtained through repeated cross-validation or independent external validation. Consequently, models evaluated using different validation strategies should not be interpreted as directly comparable, even when they report the same performance metric.

The distinction between internal predictive performance and transportability is particularly important for stunting prediction. A model may demonstrate excellent discrimination within its development dataset while performing less well in a new geographical, socioeconomic, or healthcare setting. This concern is relevant because several included studies relied on country-specific or regional datasets.[11,22,24,25,27] Differences in population structure, outcome prevalence, data quality, and predictor distributions may reduce model performance when the model is applied to a new population.

The use of pediatric data in AI and ML also requires particular attention to data quality, representativeness, transparency, and responsible model development.[13] Accordingly, external validation using an independent population should be considered an essential step before implementation in routine child health services. Future studies should report not only discrimination but also calibration and clinical or public health utility to provide a more comprehensive assessment of model performance.

Taken together, these findings indicate that there is no single algorithm that can be considered universally optimal for predicting childhood stunting. Rather, model performance reflects the interaction between algorithm choice and the characteristics of the dataset, predictor architecture, preprocessing procedures, class distribution, and validation strategy. Consequently, comparisons between algorithms should ideally be conducted using the same population, predictor set, preprocessing pipeline, and validation framework. This approach would provide a more meaningful assessment of whether observed differences in performance are attributable to the algorithm itself.

An emerging theme identified in this review is the incorporation of Explainable Artificial Intelligence (XAI), particularly SHAP, into stunting prediction models. Conventional high-performing ML models may be perceived as black boxes because healthcare professionals cannot readily determine how individual predictors contribute to predictions. Lack

of interpretability may limit trust, clinical adoption, and accountability.[17,18]

Several included studies used explainability approaches to identify the relative contribution of predictors to model predictions. In the Indonesian study by Wicaksono et al., explainable ML was used to examine important predictors of stunting, including maternal and socioeconomic characteristics.[31] Such approaches provide information not only on which children may be at increased risk but also on which characteristics contribute most strongly to the predicted risk.

However, explainability should not be interpreted as evidence of causal relationships. A predictor with a high contribution to a model's prediction indicates that the variable is informative for prediction within the analyzed dataset; it does not necessarily mean that modifying that predictor will directly reduce stunting risk. This distinction is particularly important for public health decision-making, where predictive associations should not automatically be translated into causal intervention priorities.

XAI may nevertheless provide several practical advantages. Transparent explanations may increase confidence among healthcare professionals, facilitate interpretation of complex models, and support communication between technical developers and health practitioners.[17] In low- and middle-income countries, explainable and transparent AI may also support responsible integration of digital technologies into health systems.[19]

However, explainability should complement rather than replace rigorous model evaluation. A highly interpretable model that has poor calibration or limited external validity would still have limited practical value. Therefore, future studies should evaluate interpretability alongside discrimination, calibration, external validation, fairness, and implementation feasibility.[17,18,23]

The findings of this review have important implications for maternal and child health programs, particularly in low- and middle-income countries where childhood stunting remains a major public health concern. The Sustainable Development Goals emphasize the importance of reducing malnutrition and improving child health outcomes, creating a need for approaches that support earlier identification and targeted prevention.[30]

ML-based prediction provides an opportunity to complement conventional anthropometric monitoring by identifying children who may be at elevated risk before substantial growth impairment occurs. Recent advances in AI and precision nutrition indicate that integration of diverse maternal, child, nutritional, and contextual information could strengthen personalized and population-level approaches to maternal and child health.[2]

Predictive models could potentially be integrated into existing child health information systems, community nutrition surveillance platforms, and digital growth-monitoring applications. During antenatal care, postnatal care, immunization services, or community-based nutrition monitoring, routinely collected maternal and child information could potentially be used to generate individualized risk estimates. Such predictions could support earlier nutritional counselling, breastfeeding support, complementary feeding education, micronutrient interventions, sanitation improvement, and other preventive strategies.

Digital health information systems may provide an important infrastructure for such integration. Evidence from Malawi illustrates the potential role of digital health information systems in supporting the management of child stunting, although successful implementation requires attention to local organizational and health-system contexts.[10] Similarly, broader evidence indicates that AI can strengthen healthcare systems in low- and middle-income

countries when appropriate infrastructure, governance, workforce capacity, and contextual adaptation are available.[19]

However, implementation should consider the methodological heterogeneity identified in this review. A model developed from a national survey may not automatically be applicable to a specific community or primary healthcare setting. Differences in population characteristics, available predictors, data quality, healthcare infrastructure, and prevalence of stunting may affect model performance after implementation. Therefore, adaptation and external validation should precede routine deployment.

The balance between predictive accuracy and feasibility is also important. Highly complex ensemble models may provide strong predictive performance but require greater computational resources and may be more difficult to operationalize in resource-constrained settings. In contrast, simpler models such as Decision Trees may provide lower predictive performance but greater interpretability and practical feasibility in community screening contexts. This trade-off should be considered when selecting models for implementation rather than prioritizing predictive accuracy alone.

For Indonesia, this issue is particularly relevant because national and routine health datasets provide opportunities for developing context-specific prediction models. Recent Indonesian studies have demonstrated the feasibility of applying ML, including XGBoost, Random Forest, and other algorithms, to stunting risk prediction.[22,31-34,36] However, these models should undergo rigorous internal and external validation before being translated into routine child nutrition surveillance or policy applications.

A key consideration in interpreting the findings of this systematic review is the substantial methodological heterogeneity among the included studies. Although all studies applied ML to childhood stunting prediction, they differed in population characteristics, sample size, predictor availability, feature selection and engineering, handling of class imbalance, algorithms evaluated, and validation procedures.[5,11,22,24,25,27]

Differences in study populations may affect both predictor distributions and outcome prevalence. Nationally representative survey populations may capture greater socioeconomic and geographical diversity than community-based datasets, whereas regional studies may reflect more homogeneous populations. Consequently, an algorithm's performance may partly depend on the population structure in which it was developed rather than solely on its inherent predictive capability.

Differences in sample size may influence model stability and the risk of overfitting. Large datasets can support more complex models and provide greater representation of heterogeneous population characteristics, whereas smaller datasets may produce less stable estimates, particularly when numerous predictors are considered. The included evidence illustrates this variation, ranging from relatively small datasets to large population-based samples.[11,22,24,27]

Differences in predictor selection and feature engineering may further affect model performance. The included studies used different combinations of maternal, child, socioeconomic, environmental, and birth-related variables and applied different feature-selection approaches, including LASSO, recursive feature elimination, and other variable-selection procedures.[5,11,25] Therefore, two algorithms evaluated in different studies may effectively be solving different prediction problems because the available information differs between datasets.

Class imbalance handling is another potential source of performance variation. Studies using SMOTE or undersampling modify the distribution of training observations and may improve recognition

of minority outcomes.[22] However, this may also affect the resulting performance estimates and limits direct comparison with models trained using the original class distribution.

Finally, validation methodology is critical for assessing generalizability. Differences in data partitioning, cross-validation, and external validation can produce different estimates of predictive performance.[17,20] Without standardized validation procedures, a model with a high reported accuracy or discrimination in one study cannot necessarily be considered superior to a model with lower reported performance in another study.

Accordingly, the predictive performance reported across the included studies should be interpreted as context-dependent evidence rather than a universal ranking of algorithms. This is an important distinction because the apparent superiority of an algorithm may partly reflect differences in population, predictors, preprocessing, class distribution, and validation rather than the intrinsic superiority of the algorithm itself.

This methodological heterogeneity represents an important limitation of the current evidence base. Future studies should adopt standardized reporting frameworks, clearly describe predictor selection and preprocessing procedures, report class distribution and imbalance-handling methods, and prioritize robust internal validation followed by independent external validation. Where feasible, comparative studies should evaluate multiple algorithms using the same dataset, identical predictor sets, consistent preprocessing pipelines, and comparable validation frameworks. Such designs would enable more reliable assessment of the incremental value of different ML algorithms and facilitate translation into real-world child health programs.[17,18,23]

The strength of this review lies in its focus on recent evidence concerning ML-based prediction of childhood stunting and its synthesis across predictor domains, algorithms, explainability, and potential programmatic applications. The use of PRISMA 2020 and PROBAST provided a structured approach to study identification and assessment of methodological quality. The inclusion of studies from multiple geographical settings also provides a broader perspective on the applicability of ML approaches across different child health contexts.

Nevertheless, several limitations should be considered. First, substantial methodological heterogeneity across studies prevented quantitative meta-analysis and limited direct comparison of model performance. Second, differences in population characteristics, sample size, predictor sets, preprocessing, class imbalance handling, and validation procedures may have contributed to variation in reported performance. Third, several studies were based on country-specific or regional datasets, which may limit transportability to other populations. Fourth, the availability and reporting of external validation were limited, making it difficult to determine how well the models would perform in independent populations. Finally, differences in reporting of predictive performance metrics further constrain direct comparison across studies.

Therefore, the findings should be interpreted as evidence supporting the potential of ML for stunting risk prediction, rather than as definitive evidence that one algorithm is universally superior. Future research should emphasize transparent reporting, standardized validation, external validation, calibration assessment, and evaluation of clinical and public health utility alongside discrimination.[17,18,20,23]

CONCLUSION

This systematic review indicates that machine learning has the potential to support the early prediction of childhood stunting by incorporating

multidimensional predictors related to biological, maternal, environmental, and socioeconomic conditions. Across the included studies, maternal education, household socioeconomic status, birth weight, child age, maternal nutritional status, sanitation, birth interval, and geographical factors were frequently identified as important predictors. These findings suggest that childhood stunting prediction may benefit from models that consider multiple determinants rather than relying exclusively on anthropometric characteristics.

Ensemble learning algorithms, including Gradient Boosting, Random Forest, and Extreme Gradient Boosting (XGBoost), were frequently reported to demonstrate promising predictive performance. However, differences in datasets, predictor selection, outcome definitions, validation procedures, and performance metrics limit direct comparisons between models. Explainable Artificial Intelligence (XAI), particularly SHAP-based methods, may improve model interpretability, although its contribution to clinical decision-making and public health practice requires further evaluation.

Overall, machine learning may complement existing child growth monitoring and nutrition surveillance systems by supporting the identification of children who may be at risk of stunting. Nevertheless, the available evidence remains heterogeneous, and the effectiveness, generalizability, fairness, and practical utility of these models cannot yet be assumed across different populations and healthcare settings. Future research should prioritize external and prospective validation, standardized definitions of predictors and outcomes, transparent reporting, assessment of model calibration and clinical utility, and evaluation in diverse low- and middle-income country settings. Further studies should also examine the responsible integration of machine learning into routine maternal and child health programs while addressing data quality, privacy, ethical considerations, and the potential risk of algorithmic bias.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all researchers whose published studies were included in this systematic review, as their valuable contributions provided the evidence base for this synthesis. We also acknowledge the reviewers and editors for their constructive comments and suggestions, which helped improve the quality of this manuscript.

REFERENCES

1. Benjamin-Chung J, Mertens A, Colford JM, Hubbard A, van der Laan MJ, Coyle J, et al. Early-childhood linear growth faltering in low- and middle-income countries. Nature. 2023;621:550–557. doi:10.1038/s41586-023-06418-5.
2. Mehta S, Huey SL, Fahim SM, Sinha S, Rajagopalan K, Ahmed T, et al. Advances in artificial intelligence and precision nutrition approaches to improve maternal and child health in low-resource settings. Nat Commun. 2025;16:7673. doi:10.1038/s41467-025-62985-3.
3. Robertson RC, Edens TJ, Carr L, Mutasa K, Gough E, Evans C, et al. The gut microbiome and early-life growth in a population with high prevalence of stunting. Nat Commun. 2023;14:654. doi:10.1038/s41467-023-36135-6.
4. Mertens A, Benjamin-Chung J, Colford JM, Hubbard A, van der Laan MJ, Coyle J, et al. Child wasting and concurrent stunting in low- and middle-income countries. Nature. 2023;621:558–567. doi:10.1038/s41586-023-06480-z.
5. Khan JR, Tomal JH, Raheem E. Model and variable selection using machine learning methods with applications to childhood stunting in Bangladesh.

Inform Health Soc Care. 2021;46(4):425–442. doi:10.1080/17538157.2021.1904938.
6. Food and Agriculture Organization of the United Nations, International Fund for Agricultural Development, United Nations Children's Fund, World Food Programme, World Health Organization. The state of food security and nutrition in the world 2024. Rome: FAO; 2024. doi:10.4060/cd1254en.
7. De Francesco D, Reiss JD, Roger J, Tang AS, Chang AL, Becker M, et al. Data-driven longitudinal characterization of neonatal health and morbidity. Sci Transl Med. 2023;15(683):eadc9854. doi:10.1126/scitranslmed.adc9854.
8. Mertens A, Benjamin-Chung J, Colford JM, Coyle J, van der Laan MJ, Hubbard A, et al. Causes and consequences of child growth faltering in low-resource settings. Nature. 2023;621:568–576. doi:10.1038/s41586-023-06501-x.
9. Keleş E, Bağcı U. The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review. NPJ Digit Med. 2023;6:220. doi:10.1038/s41746-023-00941-5.
10. Banda C, Nyanjahia AP. Digital health information systems and the management of child stunting in Malawi: a practice-based intervention study from a social work perspective. Int J Innov Sci Res Technol. 2026;11(1). doi:10.38124/ijisrt/26jan663.
11. Shen H, Zhao H, Jiang Y. Machine learning algorithms for predicting stunting among under-five children in Papua New Guinea. Children (Basel). 2023;10(10):1638. doi:10.3390/children10101638.
12. Fitzgerald R, Manguerra H, Arndt MB, Gardner WM, Chang YY, Zigler B, et al. Current dichotomous metrics obscure trends in severe and extreme child growth failure. Sci Adv. 2022;8(20):eabm8954. doi:10.1126/sciadv.abm8954.
13. Headey D, Ruel MT. Food inflation and child undernutrition in low- and middle-income countries. Nat Commun. 2023;14:5761. doi:10.1038/s41467-023-41543-9.
14. Fontaine F, Turjeman S, Callens K, Koren O. The intersection of undernutrition, microbiome, and child development in the first years of life. Nat Commun. 2023;14:3554. doi:10.1038/s41467-023-39285-9.
15. Matos YAS, Cano J, Shafiq H, Williams C, Sunny J, Cowardin CA. Colonization during a key developmental window reveals microbiota-dependent shifts in growth and immunity during undernutrition. Microbiome. 2024;12:71. doi:10.1186/s40168-024-01783-3.
16. Javidi H, Mariam A, Khademi G, Zabor EC, Zhao R, Radivoyevitch T, et al. Identification of robust deep neural network models of longitudinal clinical measurements. NPJ Digit Med. 2022;5:106. doi:10.1038/s41746-022-00651-4.
17. Nasarian E, Alizadehsani R, Acharya UR, Tsui K. Designing interpretable machine learning systems to enhance trust in healthcare: a systematic review and proposed responsible clinician-AI collaboration framework. Inf Fusion. 2024;108:102412. doi:10.1016/j.inffus.2024.102412.
18. Dhanda SS, Panwar D, Lin CC, Sharma TK, Rastogi D, Bindewari S, et al. Advancement in public health through machine learning: a narrative review of opportunities and ethical considerations. J Big Data. 2025;12:154. doi:10.1186/s40537-025-01201-x.
19. Ciecierski-Holmes T, Singh R, Axt M, Brenner S, Barteit S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping

review. NPJ Digit Med. 2022;5:162. doi:10.1038/s41746-022-00700-y.
20. Cascarano A, Mur-Petit J, Hernández-González J, Camacho M. Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artif Intell Rev. 2023;56:1711–1771. doi:10.1007/s10462-023-10561-w.
21. Deng W, O'Brien MK, Andersen RA, Rai R, Jones E, Jayaraman A. A systematic review of portable technologies for the early assessment of motor development in infants. NPJ Digit Med. 2025;8:63. doi:10.1038/s41746-025-01450-3.
22. Novalina N, Tarigan IAA, Kameela FK, Rizkinia M. Benchmarking machine learning algorithm for stunting risk prediction in Indonesia. Bull Electr Eng Inform. 2025;14(3):2252–2263. doi:10.11591/eei.v14i3.8997.
23. Muralidharan V, Burgart AM, Daneshjou R, Rose S. Recommendations for the use of pediatric data in artificial intelligence and machine learning: ACCEPT-AI. NPJ Digit Med. 2023;6:166. doi:10.1038/s41746-023-00898-5.
24. Ndagijimana S, Kabano IH, Masabo E, Ntaganda JM. Prediction of stunting among under-5 children in Rwanda using machine learning techniques. J Prev Med Public Health. 2023;56(1):41–49. doi:10.3961/jpmph.22.388.
25. Mkungudza J, Twabi HS, Manda SOM. Development of a diagnostic predictive model for determining child stunting in Malawi: a comparative analysis of variable selection approaches. BMC Med Res Methodol. 2024;24:175. doi:10.1186/s12874-024-02283-6.
26. Hibberd ML, Webber DM, Rodionov DA, et al. Bioactive glycans in a microbiome-directed food for children with malnutrition. Nature. 2023;625:157–165. doi:10.1038/s41586-023-06838-3.
27. Hendy A, Ibrahim RK, Abdelaliem SMF, Zaher A, Alkubati SA, Abd El-kader RG, et al. Supervised machine learning for classification and prediction of stunting among under-five Egyptian children. BMC Pediatr. 2025;25:681. doi:10.1186/s12887-025-06138-x.
28. Mgomezulu WR, Thangata P, Mkandawire B, 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. Hum Nutr Metab. 2025;42:200340. doi:10.1016/j.hnm.2025.200340.
29. Carneiro P, Kraftman L, Mason G, Moore L, Rasul I, Scott M. The impacts of a multifaceted prenatal intervention on human capital accumulation in early life. Am Econ Rev. 2021;111:2506–2549. doi:10.1257/aer.20191726.
30. United Nations Department of Economic and Social Affairs. The Sustainable Development Goals report 2024. New York: United Nations; 2024. doi:10.18356/9789213589755.
31. Wicaksono A, Prasetyo D, Mar'atullatifah Y, Iswavigra DU, Mahmudah H, Hapsari A. Data analysis and explainable machine learning for stunting prediction. J Artif Intell Leg Technol. 2025;1(1):1–10.
32. Anhar MF, Soewito B. Application of XGBoost-based machine learning methods to predict stunting. Int J Artif Intell Res. 2025;9(1.1). doi:10.29099/ijair.v9i1.1.1542.
33. Mulyani H, Musawarman, Faturrochman R, Permana DH. Machine learning-based early detection of stunting and intervention recommendations. Bit-Tech. 2025;8(2):2160–2170. doi:10.32877/bt.v8i2.3213.
34. Sutarmi, Warijan, Indrayana T, Putro DP, Gunawan I. Machine learning

model for stunting prediction. J Health Sains. 2023;4(9):10–23.
35. Wijeakumar S, Forbes SH, Magnotta VA, Deoni S, Jackson K. Stunting in infancy is associated with atypical activation of working memory and attention networks. Nat Hum Behav. 2023;7:2199–2211. doi:10.1038/s41562-023-01725-3.
36. Sahamony NF, Terttiaavini, Rianto H. Analysis of performance comparison of machine learning models for predicting stunting risk in children's growth. MALCOM Indones J Mach Learn Comput Sci. 2024;4(2):413–422. doi:10.57152/malcom.v4i2.1210.
37. Anku EK, Duah HO. Predicting and identifying factors associated with undernutrition among children under five years in Ghana using machine learning algorithms. PLoS One. 2024;19(2):e0296625. doi:10.1371/journal.pone.0296625.
38. Islam MM, Kibria NMSJ, Kumar S, Roy DC, Karim MR. Prediction of undernutrition and identification of its influencing predictors among under-five children in Bangladesh using explainable machine learning algorithms. PLoS One. 2024;19(12):e0315393. doi:10.1371/journal.pone.0315393.
39. Ayele MK, Baye GA, Yesuf SH, Engda AA, Mitiku ET. Predicting stunting status among under-five children in Ethiopia using ensemble machine learning algorithms. Sci Rep. 2025;15:27907. doi:10.1038/s41598-025-03206-1.