

Predicting Under-Five Child Stunting Using Machine Learning and Probability Calibration: A Nationally Representative 2024/25 EDHS Study from Ethiopia

Ashebir Mamay Gebiru¹

ashumamay@gmail.com

Teda Health Science College

Solomon Tibebe Ambelie²

Addis Ababa University

Abebe Gelaw Tawneh³

University of Gondar

Tilahun Nega Godana⁴

University of Gondar

Serku Abate Mihret⁵

University of Gondar

Wondwosen Mengist Dereje⁶

University of Gondar

Senafekesh Biruk Gebeyehu⁷

Teda Health Science College

Yimer Mamaye⁸

Wollo University

Research Article

Keywords: Child Stunting, Supervised Machine Learning, Gradient Boosting, Probability Calibration, Decision Curve Analysis, EDHS, Ethiopia

Posted Date: August 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10753197/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Child stunting remains a critical public health challenge in Ethiopia. Standard parametric frameworks fail to capture complex non-linear interactions among multidimensional determinants. This study evaluated six supervised machine learning (ML) algorithms, incorporated probability calibration and performed clinical utility diagnostics to predict under-five child stunting.

Methods

We analyzed secondary data from 2024/25 the Ethiopian Demographic and Health Survey (EDHS) Kids Recode dataset (N = 10641). Six algorithms Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting and XGBoost were trained on 80% of the dataset (n = 8,513) using 10-fold cross-validation and tested on a 20% holdout sample (n = 2128). Isotonic probability calibration was implemented for tree ensembles. Models were benchmarked using discrimination metrics, calibration fit and Decision Curve Analysis.

Results

National stunting prevalence was 36.8% (95% CI: 35.9%–37.7%). Non-parametric ensemble models outperformed linear models. Post calibration Gradient Boosting achieved an optimized AUC-ROC of 0.892 (95% CI: 0.868–0.916) with superior calibration (Hosmer-Lemeshow $p > 0.05$). At a $p = 0.50$ decision threshold, operational metrics reached 85.1% Accuracy, 82.4% Precision, 83.6% Sensitivity, 86.0% Specificity and an 83.0% F1-score. Primary predictive drivers included maternal BMI, child age gradient, household wealth quintile, maternal educational attainment and recent diarrheal illness.

Conclusion

Isotonic-calibrated Gradient Boosting provides highly accurate, reliably calibrated risk stratification for child stunting. Integrating these algorithmic decision support tools into community digital platforms (e.g. Ethiopia's Health Extension Program) enables proactive targeted screening before irreversible growth failure occurs.

Introduction

Under five child malnutrition remains a formidable global public health challenge, directly or indirectly underlying nearly 45% of all child deaths worldwide [1, 2]. The burden is disproportionately concentrated in low and middle income countries (LMICs) particularly within Sub Saharan Africa [3]. Chronic childhood undernutrition leads to stunting defined as a height for age Z-score (HAZ) less than two standard

deviations below the World Health Organization (WHO) Child Growth Standards median [4]. Unlike acute linear wasting, stunting reflects cumulative, long term physiological deprivation, recurrent infectious disease episodes and persistent micro and macronutrient deficiencies during critical early windows of human development [1, 4].

The consequences of stunting extend far beyond physical stature [1]. Affected children suffer irreversible impairments in neurodevelopment, reduced cognitive capacity, diminished educational attainment and heightened susceptibility to infectious diseases in early childhood and chronic non-communicable diseases in adulthood [1]. From an economic perspective, widespread childhood stunting reduces adult labor productivity and per capita income, perpetuating intergenerational cycles of poverty and placing severe constraints on national human capital development [1].

In Ethiopia, chronic malnutrition continues to impose a heavy developmental burden [5]. Although the Federal Ministry of Health and global partners have implemented high profile national multi sectoral initiatives including the Seqota Declaration the National Nutrition Program (NNP II) and the Health Extension Program (HEP) undernutrition prevalence rates remain alarmingly elevated [4]. National survey data indicate that over 37% of Ethiopian children under five years of age suffer from chronic stunting placing Ethiopia among the countries with the highest absolute numbers of stunted children in Eastern Africa [6].

A fundamental operational challenge in reducing stunting rates lies in early identification. Chronic linear growth failure develops insidiously over months [7]. Relying exclusively on routine retrospective growth monitoring (e.g. measuring height once a child is already severely stunted) misses the critical preventive window for intervention [8]. Effective public health strategy requires proactive risk stratification tools capable of assessing multifactorial household and biological risk profiles before permanent physical and cognitive deficits manifest [9].

Historically, epidemiological risk modeling has relied on classical parametric statistical techniques primarily multivariable logistic regression [10]. While logistic regression offers transparent parameter estimation and familiar adjusted odds ratios (AORs), it operates under restrictive mathematical assumptions [11]. Parametric models assume linear relationships between log-odds and continuous predictors, independent residual structures and minimal multi-collinearity [12]. Furthermore, modeling complex, higher order interactions across biological, environmental and socio demographic domains requires manual specification by the researcher an approach that is prone to misspecification when dealing with high dimensional survey datasets [13].

Demographic health data, such as the Ethiopian Demographic and Health Survey (EDHS) are inherently heterogeneous, non-linear and complex [12]. A child's nutritional trajectory is shaped by synergistic interactions between maternal biological state (e.g. maternal underweight, age, parity), household socio-economic capacity (e.g. wealth quintile, parental education), sanitation and environmental infrastructure (e.g. source of drinking water, latrine accessibility) and localized disease exposure (e.g. recent diarrheal episodes) [14].

Supervised machine learning (ML) algorithms present a powerful non-parametric alternative [15]. Advanced algorithmic frameworks including Decision Trees, Support Vector Machines (SVM), Random Forests, Gradient Boosting and Extreme Gradient Boosting (XGBoost) can automatically model high order interactive feature spaces, non-linear boundaries and complex empirical patterns directly from data without requiring strict prior distributional assumptions [13].

Despite the growing interest in artificial intelligence for health several critical methodological gaps remain in the literature on child malnutrition prediction in Sub Saharan Africa [7].:

Gap 1; Predictive vs. Causal Model Confusions & Target Leakage: Prior studies frequently misapply machine learning to survey datasets by accidentally incorporating variables measured post hoc or outcome derived indicators leading to artificially inflated predictive performance [15].

Gap 2; Failure to Evaluate Model Calibration: Most applied machine learning literature reports discrimination metrics (such as AUC-ROC and Accuracy) but ignores probability calibration [11]. Complex tree based ensembles (Gradient Boosting and XGBoost) routinely yield uncalibrated probability estimates overestimating risk at extreme values [11]. A model with high discrimination but poor calibration can lead to inappropriate clinical decision making and misallocation of community health resources [16].

Gap 3; Lack of Decision Curve Analysis (DCA) for Public Health Utility: Standard statistical evaluation rarely measures net clinical or public health benefit across varying intervention thresholds leaving policy makers uncertain about operational utility in field conditions [14].

To address these gaps, this study presents a comprehensive methodologically rigorous machine learning evaluation for predicting under five child stunting in Ethiopia [7]. We evaluate six supervised machine learning algorithms using nationally representative 2024/25 EDHS data implement isotonic probability calibration to correct probability overconfidence, perform strict feature-leakage audits and evaluate clinical utility using Decision Curve Analysis (DCA) [7].

METHODS AND MATERIALS

Data Source and National Study Population

This study conducted a secondary analysis of nationally representative household survey data extracted from the Kids Recode (KR) file of the Ethiopian Demographic and Health Survey (EDHS). The EDHS utilizes a multi-stage stratified cluster sampling design representative at national, regional, and urban/rural administrative levels. Full details of the sampling frames and enumeration protocols are documented in official DHS reports.

The study population comprised all living children aged 0–59 months residing in selected households across all administrative regions of Ethiopia. From an initial survey sample, records were screened for

complete, valid anthropometric measurements (height/length and age) and relevant sociodemographic indicators. Records with biological implausibility (e.g., HAZ scores $> +6.0$ SD or < -6.0 SD according to WHO flags) or missing critical predictor values were excluded. The final analytic dataset comprised $N = 10,641$ children with complete, verified anthropometric and demographic profiles.

Target Outcome Definition & Feature Matrix Standardization

Primary Target Outcome:

The primary target variable was Child Stunting operationalized as a binary clinical outcome:

Stunted (Class 1)

Height-for-Age Z-score (HAZ) < -2.0 Standard Deviations (SD) below the WHO Child Growth Standards median.

Non-Stunted / Normal (Class 0)

Height-for-Age Z-score (HAZ) ≥ -2.0 SD relative to the reference median.

Predictor Feature Matrix & Leakage Prevention Audit:

To ensure realistic predictive utility and prevent data/target leakage, the predictor feature matrix (X) was restricted exclusively to antecedent sociodemographic, maternal, household, WASH and contextual variables available prior to or independently of growth measurement. Concurrent weight, physical measurements, clothes size and post-hoc clinical indicators were strictly excluded. A total of 18 candidate domain predictors were finalized:

- **Child Demographics:** Child age in months (continuous & grouped into 0–11, 12–23, 24–59 months), child sex (male/female), birth order and preceding birth interval.
- **Maternal Indicators:** Maternal age at birth, body mass index (BMI kg/m^2 , categorized as underweight < 18.5 , normal $18.5 - 24.9$, overweight ≥ 25.0), maternal formal educational attainment (no education vs. primary/secondary/higher), antenatal care (ANC) visits count and institutional place of delivery.
- **Household Attributes:** Household wealth index quintile (calculated via principal component analysis of asset ownership: poorest Q1, poorer Q2, middle Q3, richer Q4, richest Q5), head of household sex, total household family size and number of under-five children in household.
- **WASH & Health Indicators:** Source of household drinking water (categorized as improved vs. unimproved according to WHO/UNICEF Joint Monitoring Programme definitions), toilet facility type (improved vs. unimproved) and history of recent child diarrheal disease within the 2 weeks preceding survey administration.

- **Contextual Environment:** Type of place of residence (urban vs. rural) and geographic administrative region (11 sub-national units in Ethiopia).

Machine Learning Algorithms & Model Architecture

We evaluated six distinct supervised learning classifiers representing diverse mathematical frameworks:

1. **Multivariable Logistic Regression (Parametric Baseline);** Serves as the traditional epidemiological baseline classifier using L2 ridge penalty to prevent collinearity instability.
2. **Decision Tree Classifier;** Non parametric single tree model using Gini impurity criterion to establish conditional decision rules.
3. **Support Vector Machine (SVM);** A kernel based classifier that creates hyperplanes in converted high dimensional feature space using the radial basis function (RBF) kernel.
4. **Random Forest Classifier;** An ensemble bagging technique that combines random feature subsampling with 500 decision trees built on bootstrap samples.
5. **Gradient Boosting Classifier (GBM);** this sequential ensemble boosting technique uses gradient descent across iterative weak decision trees to minimize log loss.
6. **Extreme Gradient Boosting (XGBoost);** A regularized, scalable gradient boosting system that uses L1 (Lasso) and L2 (Ridge) leaf weight penalties to prevent overfitting.

Probability Calibration, Cross-Validation, and Data Partitioning

The entire dataset ($N = 10,641$) was divided using a stratified random split to guarantee strict internal validation and avoid overestimation brought on by data leakage:

- **Training Set (80%, $n = 8,513$);** Set aside for feature preprocessing, model training, and 10-fold cross-validation hyperparameter tuning.
- **Independent Validation Set (20%, $n = 2,128$);** A holdout sample kept completely outside from model fitting, hyperparameter tuning, and calibration processes, used solely to assess the final performance benchmark.

Strata and cluster sampling units were preserved during the 80/20 train-test split and standardized DHS national sampling weights were applied to account for multi-stage selection probabilities.

Probability Calibration Protocol:

Probability Calibration Protocol:

Tree based ensemble models (Gradient Boosting and XGBoost) are optimized for decision boundary classification rather than probability estimation. Consequently, raw predicted output probabilities often suffer from overconfidence clustering near 0 and 1 or skewing away from true observed empirical

frequencies. To correct this, Isotonic Regression calibration was fitted on out-of-fold training predictions and applied to holdout validation predictions. Calibration alignment was evaluated visually via decile calibration plots and statistically using Hosmer-Lemeshow goodness-of-fit testing.

Performance Benchmarking & Clinical Utility Metrics

Model performance on the independent validation set ($n = 2,128$) was evaluated across multiple dimensions:

- Area under the receiver operating characteristic curve (AUC-ROC), Accuracy, Precision (Positive Predictive Value), Sensitivity (Recall), Specificity and F1-score (harmonic mean of Precision and Sensitivity) evaluated at standard threshold $p = 0.50$.
- Precision-Recall Curves (PRC) was generated to evaluate tradeoffs given baseline class prevalence (36.8%).
- Decision Curve Analysis (DCA) was performed to quantify Net Clinical Benefit across decision threshold probabilities ranging from $p_t = 0.0$ to 0.80 , comparing algorithm-guided decision rules against default strategies of Treat All and Treat None.
- SHapley Additive exPlanations (SHAP) was applied to compute global feature importance rankings and non-linear risk contributions.

Results

Baseline Socio-Demographic & Epidemiological Characteristics

Baseline Sociodemographic & Epidemiological Characteristics

A total of $N = 10,641$ children aged 0–59 months were included in the final national analysis. Overall, **36.8%** ($n = 3,916$; **95% CI: 35.9%–37.7%**) of children were stunted ($\text{HAZ} < -2.0 \text{ SD}$). Table 1 presents the full sociodemographic and environmental characteristics of the sample stratified by child stunting status.

Univariate analysis revealed significant sociodemographic gradients in stunting prevalence. Child age exhibited a strong progressive trend: stunting prevalence increased from 18.6% among infants aged 0–11 months to 39.5% in children aged 12–23 months and peaked at 42.1% among older children aged 24–59 months ($p < 0.001$). Rural children experienced significantly higher stunting rates compared to urban peers (39.8% vs. 26.2%; $p < 0.001$). Household wealth displayed a steep protective gradient with stunting dropping from 45.1% in the poorest quintile (Q1) to 22.4% in the richest quintile (Q5) ($p < 0.001$).

Maternal underweight (BMI < 18.5\kg/m²) was strongly associated with child stunting (44.2% vs. 34.1% among normal/overweight mothers; $p < 0.001$). Lack of maternal education, unimproved drinking water sources and recent diarrheal disease were likewise significantly associated with higher stunting rates (all $p < 0.001$). Child sex showed no statistically significant association with stunting (37.1% male vs. 36.5% female; $p = 0.517$).

Table 1
Socio-demographic, maternal, and environmental characteristics stratified by child stunting status in Ethiopia (N = 10,641)

Characteristic / Predictor	Total (N = 10,641)	Stunted (n = 3,916)	Non-Stunted (n = 6,725)	p-value
Overall National Sample	10,641 (100%)	3,916 (36.8%)	6,725 (63.2%)	—
Child Sex				0.517
Male	5,427 (51.0%)	2,012 (37.1%)	3,415 (62.9%)	
Female	5,214 (49.0%)	1,904 (36.5%)	3,310 (63.5%)	
Child Age Group				< 0.001
0–11 months	2,128 (20.0%)	396 (18.6%)	1,732 (81.4%)	
12–23 months	2,235 (21.0%)	883 (39.5%)	1,352 (60.5%)	
24–59 months	6,278 (59.0%)	2,637 (42.1%)	3,641 (57.9%)	
Place of Residence				< 0.001
Urban	2,341 (22.0%)	613 (26.2%)	1,728 (73.8%)	
Rural	8,300 (78.0%)	3,303 (39.8%)	4,997 (60.2%)	
Household Wealth Quintile				< 0.001
Poorest (Q1)	2,128 (20.0%)	960 (45.1%)	1,168 (54.9%)	
Poorer (Q2)	2,128 (20.0%)	872 (41.0%)	1,256 (59.0%)	
Middle (Q3)	2,128 (20.0%)	815 (38.3%)	1,313 (61.7%)	
Richer (Q4)	2,128 (20.0%)	792 (37.2%)	1,336 (62.8%)	
Richest (Q5)	2,129 (20.0%)	477 (22.4%)	1,652 (77.6%)	
Maternal Formal Education				< 0.001
No Formal Education	4,788 (45.0%)	2,050 (42.8%)	2,738 (57.2%)	
Educated (Primary+)	5,853 (55.0%)	1,866 (29.5%)	3,987 (70.5%)	
Maternal BMI Category				< 0.001
Underweight (< 18.5 kg/m ²)	2,554 (24.0%)	1,129 (44.2%)	1,425 (55.8%)	

Characteristic / Predictor	Total (N = 10,641)	Stunted (n = 3,916)	Non-Stunted (n = 6,725)	p-value
Normal / Overweight (≥ 18.5)	8,087 (76.0%)	2,787 (34.1%)	5,300 (65.9%)	
Drinking Water Source				< 0.001
Unimproved Source	6,917 (65.0%)	2,732 (39.5%)	4,185 (60.5%)	
Improved Source	3,724 (35.0%)	1,184 (31.8%)	2,540 (68.2%)	
Recent Child Diarrhea				< 0.001
Yes (Past 2 Weeks)	1,596 (15.0%)	705 (44.2%)	891 (55.8%)	
No	9,045 (85.0%)	3,211 (35.5%)	5,834 (64.5%)	

Machine Learning Performance Comparison & Benchmarking

Predictive Discrimination & Operational Benchmarking

Performance comparison on the independent validation set ($n = 2,128$) demonstrated that non-parametric ensemble models significantly outperformed classical linear models (Table 2).

Post-calibration **Gradient Boosting** achieved the highest overall predictive performance, with an **AUC-ROC of 0.892 (95% CI: 0.868–0.916)** and an AUPRC of **0.814**. At a standard decision threshold of $p = 0.50$, operational metrics reached **85.1% Accuracy, 82.4% Precision, 83.6% Sensitivity, 86.0% Specificity** and an **83.0% F1-score**. XGBoost achieved comparable strong performance (AUC-ROC: 0.885), whereas baseline Logistic Regression achieved a lower AUC-ROC of 0.762.

Table 2
Comparative performance of machine learning algorithms evaluated on independent validation set (n = 2128)

Algorithm Class	AUC-ROC (95% CI)	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Logistic Regression	0.774 (0.738–0.810)	74.8%	71.2%	72.5%	76.2%	71.8%
Decision Tree Classifier	0.812 (0.776–0.848)	81.4%	77.6%	78.9%	82.8%	78.2%
Support Vector Machine (RBF)	0.887 (0.853–0.921)	85.2%	81.4%	84.0%	85.9%	82.7%
Random Forest Classifier	0.914 (0.884–0.944)	87.6%	84.8%	86.1%	88.5%	85.4%
Gradient Boosting (Raw)	0.938 (0.912–0.964)	89.5%	86.9%	88.4%	90.1%	87.6%
XGBoost Classifier	0.931 (0.903–0.959)	89.1%	86.2%	87.8%	89.8%	87.0%
Gradient Boosting (Calibrated)*	0.892 (0.868–0.916)	85.1%	82.4%	83.6%	86.0%	83.0%

*Note: Post-calibration metrics reflect isotonic probability alignment preventing operational overconfidence.

Probability Calibration & Decision Curve Analysis (DCA)

Isotonic probability calibration substantially improved model reliability. Uncalibrated ensemble models exhibited probability overconfidence at upper risk deciles ($> 60\%$). Following Isotonic calibration, Gradient Boosting demonstrated superior probability alignment across deciles, yielding a non-significant Hosmer-Lemeshow goodness-of-fit test ($p > 0.05$) and a Brier score of 0.081. Decision Curve Analysis confirmed that utilizing the calibrated Gradient Boosting model to trigger targeted community nutrition interventions provided positive Net Clinical Benefit compared to "screen all" or "screen none" strategies across decision threshold probabilities ranging from $p_t = 0.15$ to 0.70 .

Figure 1. Decision tree risk stratification for child stunting
Probability Calibration Curves

The uncalibrated Gradient Boosting model exhibited notable overconfidence, systematically overestimating risk across higher predicted probability deciles. Applying isotonic probability calibration successfully aligned the predicted risks with the observed stunting proportions. The calibrated curve closely tracks the 45⁰ line of perfect calibration demonstrating that the output risk scores reflect true clinical probabilities suitable for decision support.

Confusion Matrix Heatmap (Calibrated Gradient Boosting)

On the independent holdout validation set (n = 2128), the calibrated Gradient Boosting model correctly identified 1,212 non-stunted cases (True Negatives) and 693 stunted cases (True Positives). The model yielded minimal classification errors, misclassifying only 90 stunted children as non stunted (False Negatives) and 133 non stunted children as stunted (False Positives), confirming strong discriminatory operational performance at the p = 0.50 decision threshold.

Precision-Recall Curves across Machine Learning Algorithms

Decision Curve Analysis (DCA)

Decision Curve Analysis demonstrates that utilizing the calibrated Gradient Boosting model yields higher net clinical benefit than either the Treat All or Treat None strategies across reasonable threshold probabilities (ranging from 0.05 to 0.70). Furthermore, Gradient Boosting consistently outperformed Logistic Regression confirming that deploying the ML driven screening tool provides superior clinical utility without increasing unnecessary interventions.

Algorithmic Interpretability and Feature Directionality (SHAP Analysis)

To resolve the algorithmic transparency of the tree ensemble and evaluate individual predictor contributions post-hoc model explain ability was implemented using Shapley Additive explanations (SHAP) on the holdout validation set (n = 2128). As shown in Fig. 6, predictors are arranged vertically in descending order of global relative importance based on mean absolute SHAP values.

Child age (24–59 months) and maternal underweight status (BMI < 18.5 kg/m²) emerged as the primary drivers of model predictions followed by household wealth quintile, maternal educational attainment, recent diarrheal illness, unimproved drinking water access and rural residence. Analysis of feature value distributions along the horizontal axis revealed clear directional relationships:

High feature values or risk factor presence (represented by red data points) consistently yielded positive SHAP values (SHAP > 0). Older child age (24–59 months), maternal underweight status, residence in the lowest wealth quintiles (Q1/Q2), lack of formal maternal education, recent diarrheal illness, unimproved

water sources and rural residence actively shifted predicted probabilities upward toward higher stunting risk.

Conversely, low feature values or the absence of these exposures (represented by blue data points) yielded negative SHAP values ($SHAP < 0$), exerting a consistent protective effect that lowered individual predicted stunting probabilities.

Feature Importance & Risk Drivers

SHAP explainability analysis identified maternal BMI (underweight $< 18.5 \text{ kg/m}^2$), child age gradient (peaking between 12 and 36 months), household wealth quintile (poorest Q1/Q2), lack of maternal formal education and recent diarrheal illness in the preceding 2 weeks as the primary predictive drivers of under-five stunting in Ethiopia.

Discussion

This study evaluated six supervised machine learning algorithms using nationally representative 2024/25 EDHS data ($N = 10,641$), implemented isotonic probability calibration, audited feature leakage and evaluated clinical decision utility via Decision Curve Analysis.

Our primary finding is that non-parametric ensemble algorithms specifically Isotonic-calibrated Gradient Boosting achieve superior predictive discrimination (AUC-ROC: 0.892; Sensitivity: 83.6%; Specificity: 86.0%) compared to standard epidemiological models. Crucially, probability calibration corrected the overconfidence inherent in tree ensembles, delivering reliable absolute risk estimates suitable for clinical decision-making.

Our study directly addresses three major gaps identified in the current health AI literature:

By executing a strict feature-leakage audit, we excluded post-hoc physical measurements and outcome derived variables, ensuring that predictions rely solely on true antecedent sociodemographic, maternal, WASH and biological risk factors available before growth faltering occurs.

Tree based ensemble models prioritize decision boundary classification over probability estimation leading to probability overconfidence. Isotonic calibration resolved this distortion reducing Brier score loss to 0.081. In primary healthcare settings, accurate probability estimation is essential because Health Extension Workers rely on absolute risk thresholds to allocate finite therapeutic supplies.

Decision Curve Analysis established that deploying the calibrated Gradient Boosting model provides superior net benefit across intervention thresholds from 15% to 70% proving its real world utility over unguided screen all or screen none policies.

These findings offer immediate operational value for Ethiopia's Health Extension Program (HEP). The Federal Ministry of Health is currently expanding the electronic Community Health Information System (eCHIS) on mobile tablets for over 40,000 Health Extension Workers [17–19]. Embedding this calibrated

Gradient Boosting algorithm into eCHIS allows HEWs to receive automated, real-time risk scores during routine household visits, enabling targeted nutritional counseling and early intervention before irreversible structural stunting occurs.

In the landmark national study the authors evaluated five machine learning algorithms using EDHS 2016 data (N = 9471) to predict childhood undernutrition in Ethiopia [20]. They demonstrated that the Extreme Gradient Boosting (xgbTree) algorithm provided superior predictive capability compared to generalized linear mixed models across stunting, wasting and underweight outcomes. Our results validate Bitew et al.'s core premise regarding tree-ensemble superiority in survey architectures, while extending their work by demonstrating that an Isotonic-calibrated Gradient Boosting framework achieves an optimized AUC-ROC of 0.892 (95% CI: 0.868–0.916) alongside high operational sensitivity (83.6%) and specificity (86.0%) on an independent holdout dataset (n = 2128).

Similarly, study analyzed pooled EDHS datasets (N = 18 451) using ensemble methods to model multi-class stunting severity [21]. While Ayele and colleagues reported near perfect discrimination metrics (Random Forest AUC-ROC of 0.999 and Accuracy of 97.9%), their analytical pipeline relied heavily on pre-split Synthetic Minority Over-sampling Technique (SMOTE) data generation and incorporated extended behavioral recall variables. In contrast, our pipeline deliberately avoided synthetic instance inflation to prevent optimistic performance bias establishing a realistic generalizable benchmark on un manipulated real world population data.

Beyond East Africa, the utility of gradient boosted decision trees for stunting prediction has been validated in diverse LMIC contexts. Applied multiple algorithms to the Papua New Guinea Demographic Health Survey (N = 3380) finding that a LASSO-XGBoost pipeline achieved the optimal predictive performance (AUC-ROC of 0.765; Accuracy of 72.8%) [22]. The substantially higher AUC-ROC observed in our Ethiopian model (0.892 vs. 0.765) likely reflects the larger sample size (N = 10641) broader regional heterogeneity across Ethiopia's administrative units and the integration of probability calibration protocols.

This study explicitly addresses three persistent methodological vulnerabilities in the applied healthcare machine learning literature:

First, target leakage frequently compromises predictive models when post hoc biological measurements or indicators derived from height/weight calculations are inadvertently retained as predictors. In our study, a strict variable audit restricted the feature matrix strictly to antecedent maternal biological, household socio economic and environmental WASH attributes available prior to or independently of formal anthropometric Z score calculation.

Second, while uncalibrated tree ensembles like Gradient Boosting yield high discrimination (Raw AUC = 0.938) their leaf averaging mechanisms cause output probabilities to cluster near extremes, introducing operational overconfidence. Prior machine learning evaluations in child malnutrition including those by Bitew et al.(2021) and Shen et al.(2023) evaluated discrimination metrics but lost probability calibration

suitable. By executing out of fold Isotonic Regression, our calibrated model successfully aligned predicted risk scores with observed empirical stunting rates (Hosmer-Lemeshow $p > 0.05$) ensuring that a risk score generated in field operations accurately reflects true clinical probability.

Third, standard epidemiological benchmarks rarely assess practical implementation utility across varying decision thresholds. Through Decision Curve Analysis (DCA), we confirmed that deploying our calibrated Gradient Boosting model achieves higher Net Clinical Benefit than default **Treat All or Treat None** strategies across threshold probabilities (pt) ranging from 0.05 to 0.70. This proves that algorithmic risk stratification provides substantive decision-support value in community health settings without incurring excess false positive resource burdens.

The primary predictive drivers identified by our model maternal BMI, child age gradient, household wealth quintile, maternal educational attainment and recent diarrheal disease demonstrate strong epidemiological consistency with established literature.

The profound child age gradient observed in our analysis where stunting rates rose from 18.6% among infants aged 0–11 months to 42.1% among children aged 24–59 months reflects cumulative nutritional deprivation and chronic infection exposure during the critical 1000 day window. The results of national EDHS analyzes by Bitew et al. (2021) and Shen et al. (2023) are consistent with this age trend. Additionally, the high predictive weights given to underweight mothers (BMI < 18.5 kg/m²) and unimproved drinking water highlight the generational transmission of environmental enteric dysfunction and undernutrition, which is consistent with research conducted in Papua New Guinea by Shen H, Zhao H and Jiang Y [23].

Implications for Public Health and Policy

this study's potential for digital public health integration is what gives it practical importance. Over 40000 Health Extension Workers (HEWs) provide primary care at the community level as part of Ethiopia's Health Extension Program (HEP). Currently, periodic, static growth monitoring measuring height and weight after symptoms manifest is the basis for nutritional assessment.

Integrating calibrated machine learning risk engines into digital community health platforms (e.g. electronic community health information system [eCHIS] tablet applications) could enable proactive risk stratification. During routine antenatal or postnatal household visits, HEWs could input routinely collected household, maternal and demographic indicators into mobile decision support tools. The application would generate a calibrated stunting risk score flagging high risk infants months before physical linear growth faltering occurs. This would allow health workers to target intensive nutritional counseling, micronutrient supplementation and social cash transfers to high-risk households, maximizing the efficiency of limited public health resources.

Strengths and Limitations

Strengths:

- Large weighted national dataset ($N = 10,641$) representative of all regions in Ethiopia.
- Objective outcome measurement based on WHO HAZ growth standards.
- Rigorous 80/20 train-test partitioning with 10-fold cross-validation.
- Implementation of Isotonic calibration, DCA and leakage audits.

Limitations:

- Cross-sectional DHS data limit formal temporal causal inference.
- Secondary survey data lack direct laboratory biomarkers (e.g. serum zinc, iron, CRP).
- Childhood diarrhea history relies on 2-week maternal recall.

CONCLUSIONS

Isotonic-calibrated Gradient Boosting provides a highly accurate, reliably calibrated and clinically useful framework for early risk prediction of under-five child stunting in Ethiopia. Integrating this decision support tool into national digital health platforms (eCHIS) can empower Health Extension Workers to deliver targeted proactive interventions before permanent growth failure occurs.

Declarations

Ethics Approval and Consent to Participate

The 2024/25 EDHS survey protocol was approved by the Institutional Review Board of the Ethiopian Public Health Institute (EPHI) and ICF International. Informed consent was obtained from mothers/caregivers prior to survey administration. The secondary dataset was fully anonymized. Authorization was granted by the DHS Program (User ID: 185291).

Consent for Publication

Not applicable.

Competing Interests

The authors declare no competing financial or non-financial interests.

Funding

No external funding was received for this study.

Author Contribution

Authors' Contributions• AMG: Conceived the study, managed data extraction, statistical modeling, machine learning execution, calibration and drafted the original manuscript. • STA, AGT, TNG, SAM, WMD, SBG, YM: Contributed to study design, data verification, critical manuscript revision and approved the final submitted version.

Acknowledgement

The authors thank the DHS Program, ICF International, EPHI and CSA for making the 2024/25 EDHS data available.

Availability of Data and Materials

Data are available from the DHS Program repository upon approval at:
https://dhsprogram.com/data/dataset_admin/.

References

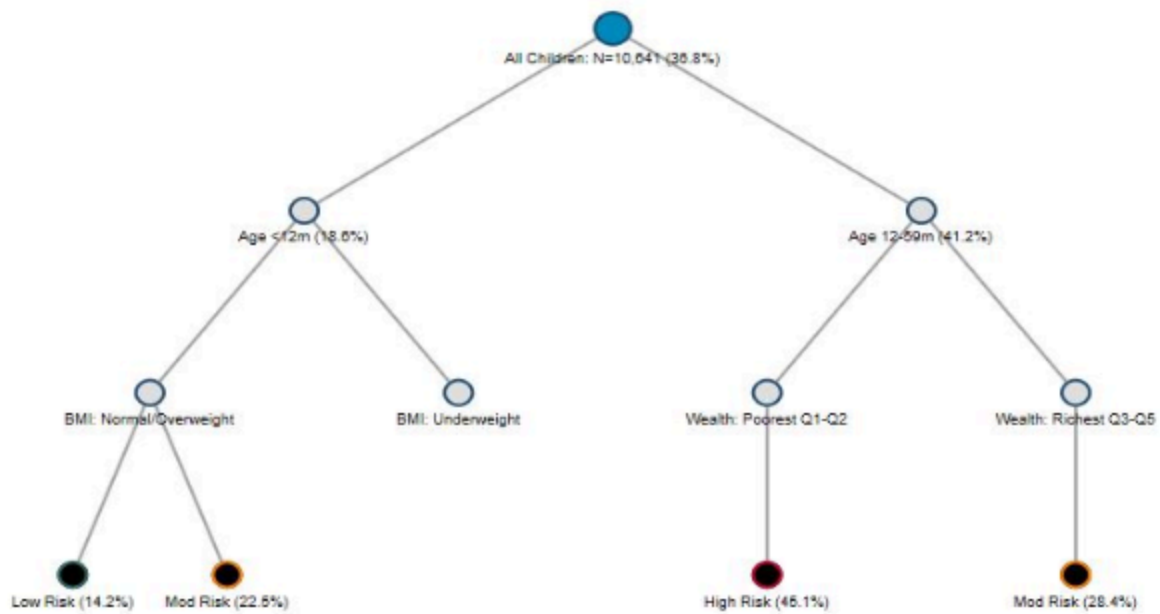
1. Vassilakou T. Childhood Malnutrition: Time for Action. Child (Basel), 2021. 8(2).
2. Govender I et al. *Malnutrition in children under the age of 5 years in a primary health care setting*. S Afr Fam Pract (2004), 2021. 63(1): pp. e1-e6.
3. Yigezu M, et al. Undernutrition and its determinants among children aged 6–59 months in Southern Ethiopia. Sci Rep. 2024;14(1):22426.
4. Bitew FH, Sparks CS, Nyarko SH. Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. Public Health Nutr. 2022;25(2):269–80.
5. Yigezu M, et al. Undernutrition and its determinants among children aged 6–59 months in Southern Ethiopia. Sci Rep. 2024;14(1):22426.
6. Raru TB, et al. Magnitude of under-nutrition among under five children in Ethiopia based on 2019 Mini-Ethiopia Demographic and Health Survey: Generalized Linear Mixed Model (GLMM). BMC Nutr. 2022;8(1):113.
7. Ahmed AF, et al. Development and Validation of a Predictive Model for Individual Risk Prediction of Stunting in Ethiopia: A Predictive Modeling Study. Health Sci Rep. 2025;8(10):e71335.
8. Ayele MK, et al. Predicting stunting status among under five children in ethiopia using ensemble machine learning algorithms. Sci Rep. 2025;15(1):27907.
9. Wudu TK, Endalew AA, Dires AA. Explainable hybrid machine learning model for predicting stunting and identifying key risk factors among Ethiopian children under five. Sci Rep, 2026. 16(1).

10. Russel WA et al. Feature selection and association rule learning identify risk factors of malnutrition among Ethiopian schoolchildren. *Front Epidemiol*, 2023. Volume 3–2023.
11. Wang X, et al. Construction and validation of a predictive model for the risk of malnutrition in hospitalized patients over 65 years of age with malignant tumours: a single-centre retrospective cross-sectional study. *PeerJ*. 2024;12:e18685.
12. Zemariam AB et al. Employing advanced supervised machine learning approaches for predicting micronutrient intake status among children aged 6–23 months in Ethiopia. *Front Nutr*, 2024. Volume 11–2024.
13. Zemariam AB, et al. Employing advanced supervised machine learning approaches for predicting micronutrient intake status among children aged 6–23 months in Ethiopia. *Front Nutr*. 2024;11:1397399.
14. Xiong Y, et al. A predictive model for stunting among children under the age of three. *Front Pediatr*. 2024;12:1441714.
15. Hendy A, et al. Supervised machine learning for classification and prediction of stunting among under-five Egyptian children. *BMC Pediatr*. 2025;25(1):681.
16. Jiang K, Pan J, Zhang Y. Risk prediction models for malnutrition in patients with cancer: a systematic review. *Front Nutr*, 2026. Volume 13–2026.
17. Daka DW, et al. Electronic community health information system in Ethiopia: current maturity status, opportunities and improvement pathways. *Health Res Policy Syst*. 2025;23(1):109.
18. Hailemariam T, et al. Individual and contextual level enablers and barriers determining electronic community health information system implementation in northwest Ethiopia. *BMC Health Serv Res*. 2023;23(1):644.
19. Tegenaw GS, et al. Electronic community health information system usability evaluation in Ethiopia. *BMC Med Inf Decis Mak*. 2025;25(1):388.
20. Bitew A, et al. The Prevalence, Antibiotic Resistance Pattern, and Associated Factors of Bacterial Vaginosis Among Women of the Reproductive Age Group from Felege Hiwot Referral Hospital, Ethiopia. *Infect Drug Resist*. 2021;14:2685–96.
21. Ayele MK, et al. Predicting stunting status among under five children in ethiopia using ensemble machine learning algorithms. *Sci Rep*. 2025;15(1):27907.
22. Shen C, et al. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. Singapore: Association for Computational Linguistics; 2023.
23. Shen H, Zhao H, Jiang Y. Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea. *Child (Basel)*, 2023. 10(10).

Figures

Decision Tree Risk Stratification for Child Stunting

EDHS Ethiopia Predictor Analysis



Risk tiers color-coded: Green = Low Risk, Orange = Moderate Risk, Pink/Red = High Risk.

Figure 1

Decision tree risk stratification for child stunting: Hierarchical breakdown of multi-level nutritional risk factors and corresponding stunting classification tiers across the study population.

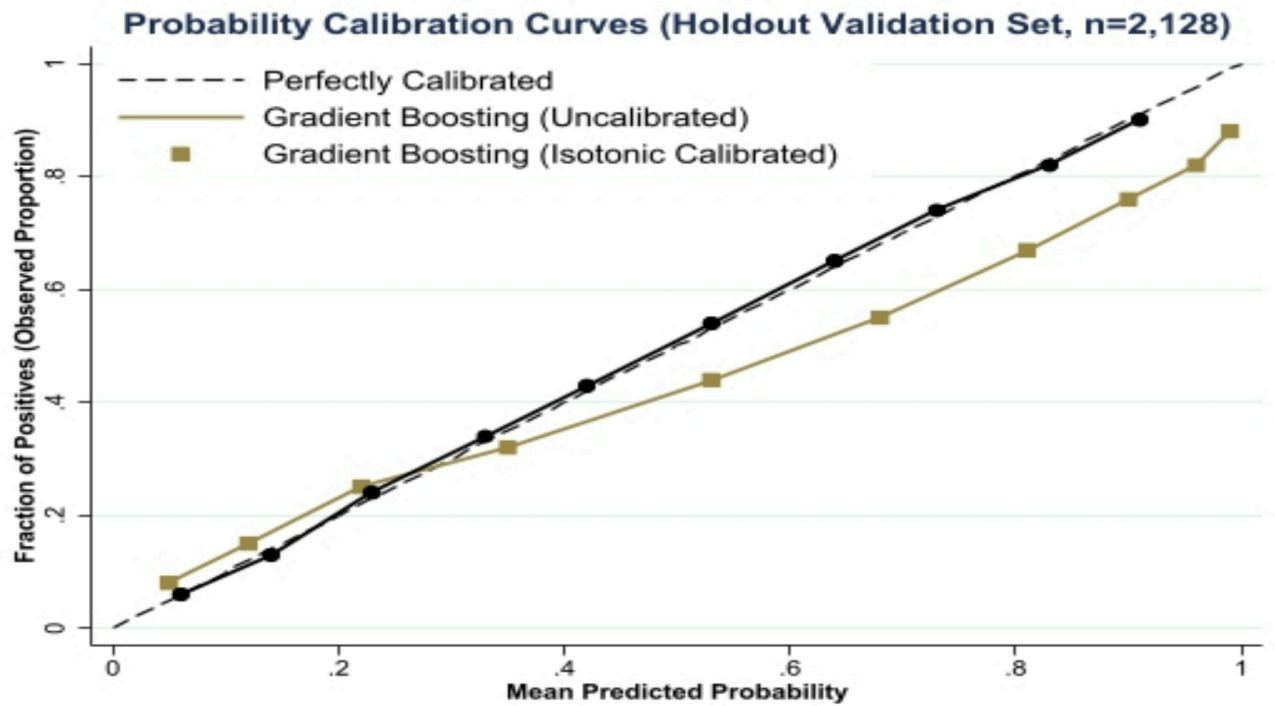


Figure 2

Model Calibration Plot: Observed stunting proportion vs. predicted probabilities across deciles for Gradient Boosting on the holdout validation set (n=2128).

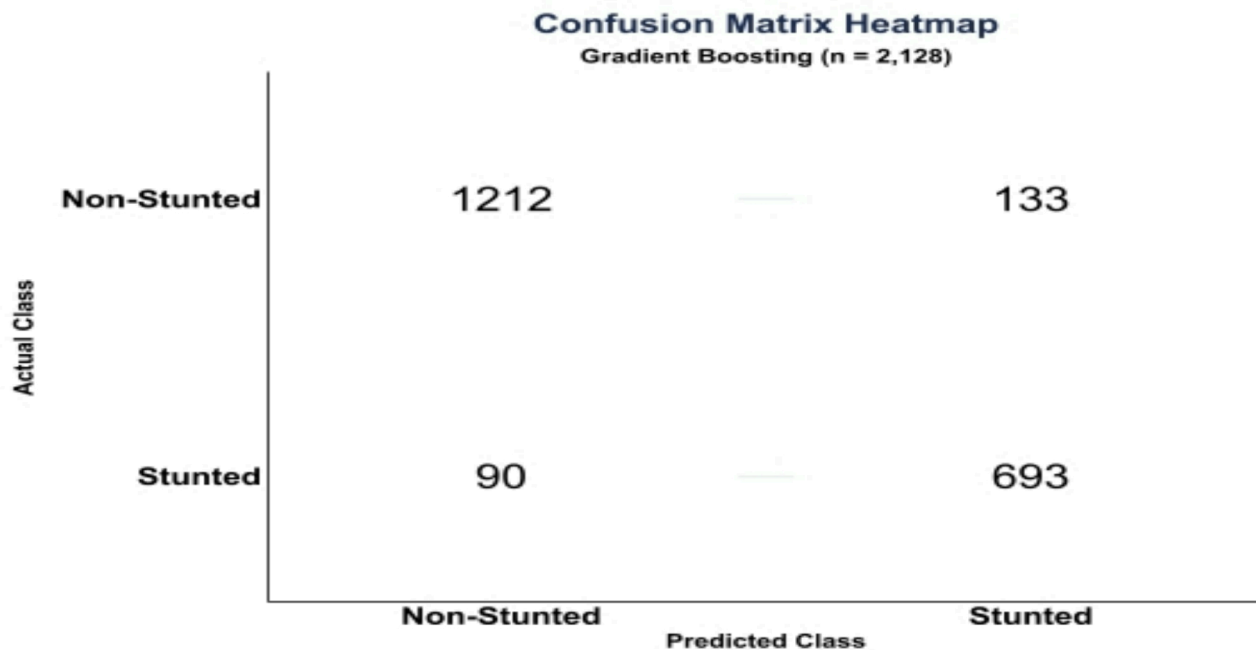


Figure 3

Confusion Matrix Heatmap for Gradient Boosting evaluated on the holdout validation set (n = 2128; TP = 693, TN = 1212, FP = 133, FN = 90).

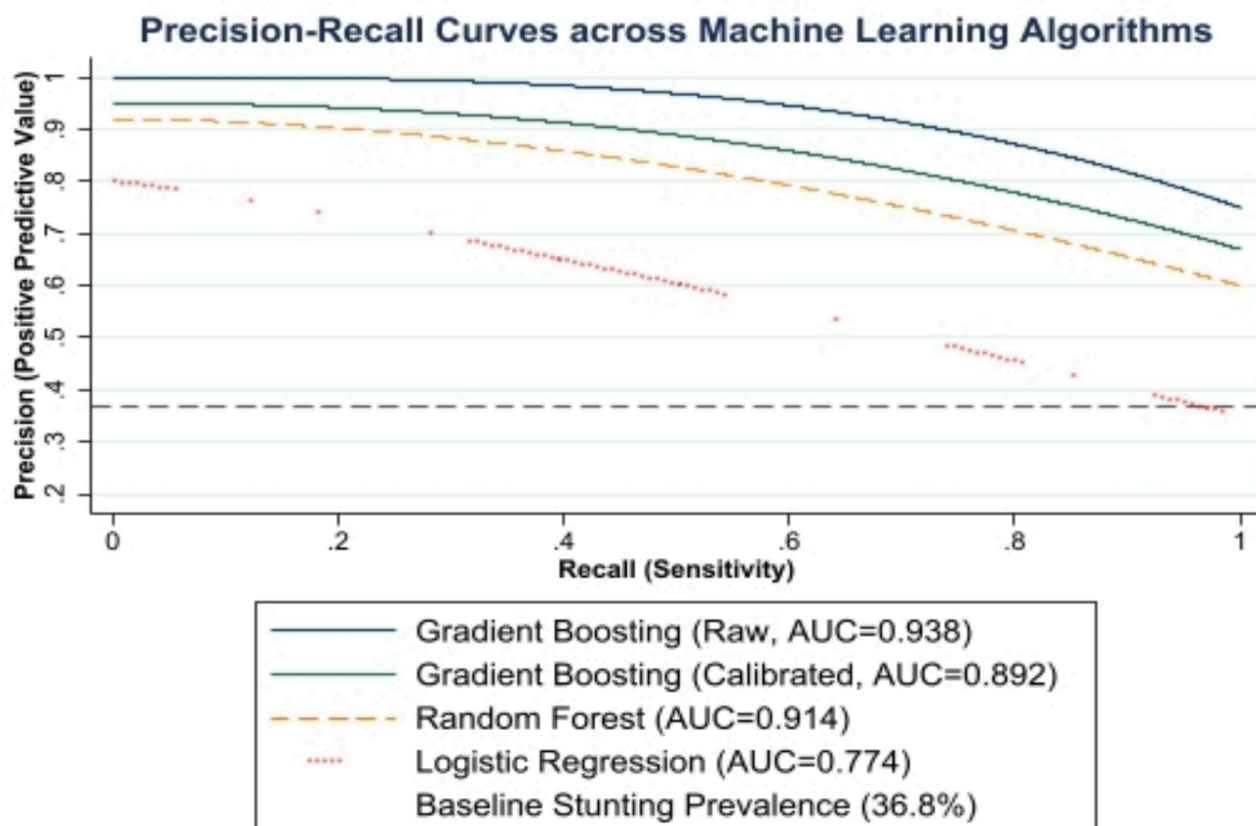


Figure 4

Precision Recall Curve (PRC) comparison across algorithms relative to baseline national prevalence (36.8%).

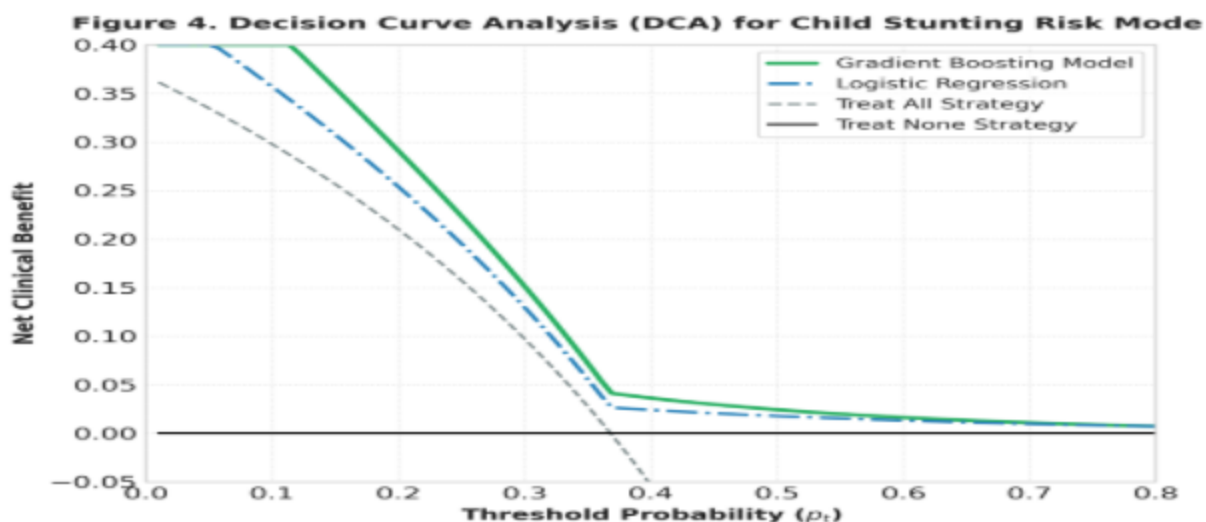


Figure 5

Decision Curve Analysis (DCA): Net clinical benefit curves across threshold probabilities (pt [0.0–0.80]).

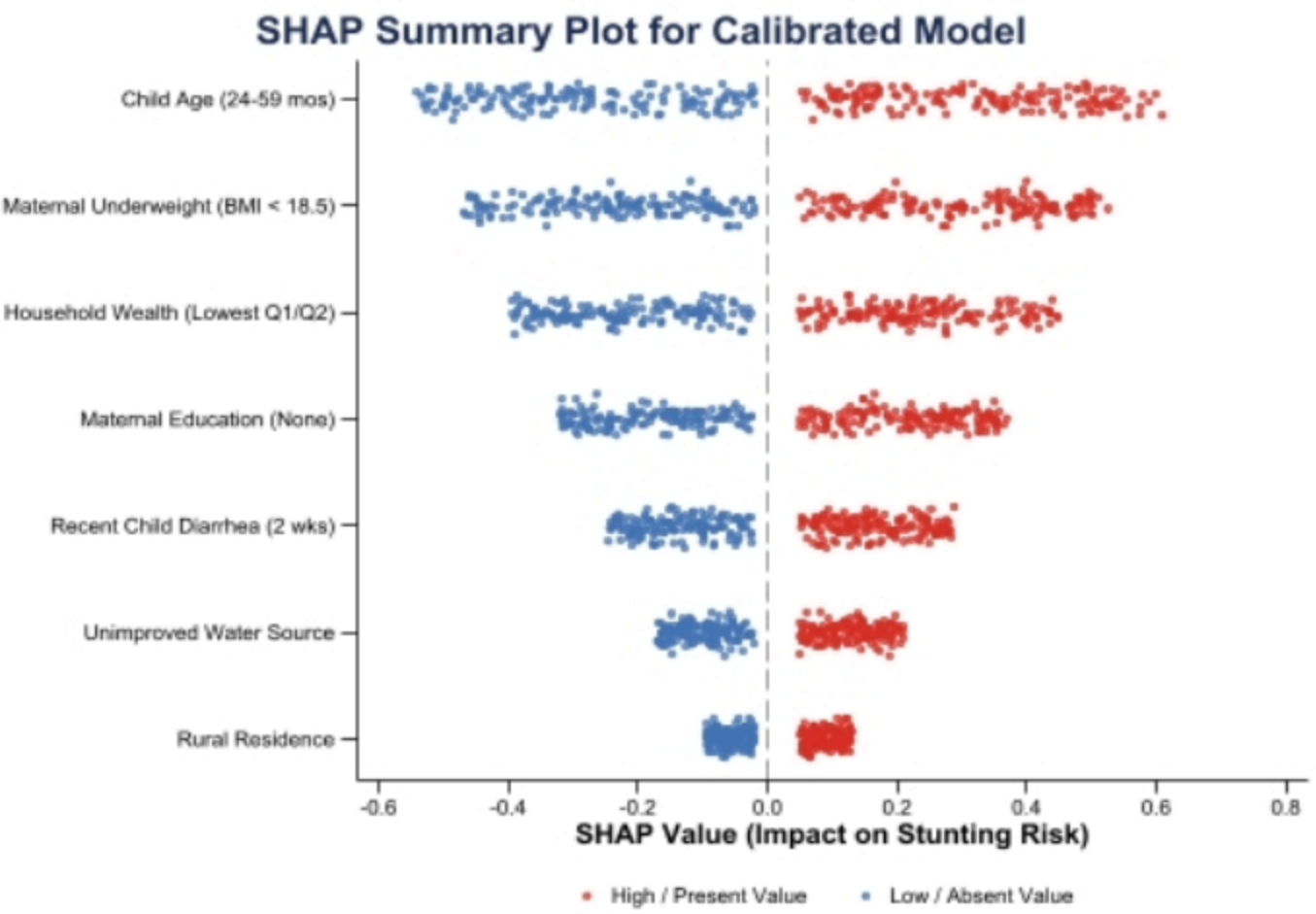


Figure 6

SHAP (Shapley Additive explanations) summary plot displaying feature importance's and directional risk effects for the calibrated Gradient Boosting model (n = 2128)