

Predicting Childhood Immunization Defaulters and Zero-Dose Children to Enable Digital Community Health Triage in Ethiopia: An Explainable Machine Learning Analysis of the 2024/25 Demographic and Health Survey

Ashebir Mamay Gebiru¹

ashumamay@gmail.com

Teda Health Science College

Yenehun Demelash Gelaw

Teda Health Science College

Bimerew Bayuh Yimer

Teda Health Science College

Research Article

Keywords: Public Health Informatics, Immunization Defaulter, Zero-Dose Children, eCHIS, Machine Learning, XGBoost, SHAP, EDHS, Ethiopia

Posted Date: September 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10853711/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting Childhood Immunization Defaulters and Zero-Dose Children to Enable Digital Community Health Triage in Ethiopia: An Explainable Machine Learning Analysis of the 2024/25 Demographic and Health Survey

Authors and Affiliations

Ashebir Mamay Gebiru¹, Yenehun Demelash Gelaw², Bimerew Bayuh Yimer¹

¹Department of Epidemiology and Biostatistics, Teda Health Science College, Gondar, Ethiopia

²Department of pharmacy, Teda Health Science College, Gondar, Ethiopia

***Corresponding Author:**

Ashebir Mamay Gebiru,

Email: ashumamay@gmail.com | Phone: +251918274844

ORCID: <https://orcid.org/0009-0007-9013-182X>

Abstract

Background: Childhood immunization reduces under five mortality, yet Ethiopia struggles with high rates of zero-dose children and immunization defaulters. Digital health platforms like the electronic Community Health Information System (eCHIS) offer infrastructure for automated risk stratification. This study developed and interpreted machine learning models to predict immunization defaulting and zero-dose status using the 2024/25 Ethiopian Demographic and Health Survey (EDHS) to inform digital community tools.

Methods: A secondary analysis of cross sectional data from the 2024/25 EDHS dataset was conducted among children aged 12-23 months. The outcome was immunization defaulting/non-full vaccination. Twenty five predictors across sociodemographic, maternal and household domains were extracted. Six algorithms (Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting and XGBoost) were benchmarked. Synthetic Minority Over-sampling handled class imbalance. Models were evaluated using AUROC, AUPRC, Brier scores and Decision Curve Analysis. SHAP decoded model explainability.

Results: The national weighted prevalence of immunization defaulting was 32.4% with a zero dose prevalence of 14.8%. XGBoost achieved superior discrimination on the holdout validation set (AUROC: 0.891; AUPRC: 0.776), outperforming Logistic Regression (AUROC: 0.768). Isotonic calibration reduced prediction loss (Brier score: 0.046). Decision Curve Analysis confirmed positive net benefit across threshold probabilities from 10% to 70%. SHAP identified antenatal care visits, home delivery, maternal education, facility distance and wealth as primary drivers alongside a rural home delivery synergistic multiplier.

Conclusions: Isotonic-calibrated XGBoost combined with SHAP provides a transparent framework for identifying high risk vaccine defaulters in Ethiopia. Embedding this engine into platforms like eCHIS empowers Health Extension Workers to execute proactive outreach accelerating Universal Immunization Coverage.

Keywords: Public Health Informatics; Immunization Defaulter; Zero-Dose Children; eCHIS; Machine Learning; XGBoost; SHAP; EDHS; Ethiopia.

INTRODUCTION

Vaccine-preventable diseases (VPDs) such as measles, pertussis, diphtheria, tetanus and rotavirus gastroenteritis remain major contributors to under five child mortality in low and middle income countries (LMICs) [1-3]. Immunization represents one of the most powerful cost-effective public health interventions saving an estimated 3.5 to 5 million lives globally each year [1, 4-6]. Recognizing this the World Health Organization (WHO), UNICEF and Gavi (the Vaccine Alliance) established the Immunization Agenda 2030 (IA2030) and Gavi 5.0 Strategic Framework, which place paramount emphasis on reaching zero-dose children (children who have not received a single dose of diphtheria tetanus pertussis [DTP1] containing vaccine) and reducing immunization defaulter rates [2, 7, 8].

In Ethiopia, despite significant expansion of the Expanded Program on Immunization (EPI) through the Health Extension Program (HEP), immunization coverage remains below national and global targets [1, 9, 10]. According to recent epidemiological estimates, nearly one in three Ethiopian children fails to complete the routine childhood immunization schedule by their second birthday and over 14% remain completely unvaccinated zero dose children [11]. Immunization defaulting leads to fragmented population immunity predisposing communities to catastrophic outbreaks of vaccine preventable diseases, such as recurring measles epidemics and circulating vaccine derived polioviruses (cVDPV) [2].

A fundamental barrier to achieving universal immunization coverage in Ethiopia is the lack of automated data driven early warning and risk stratification tools at the primary healthcare level [12]. In Ethiopia's health system architecture, over 40000 Health Extension Workers (HEWs) serve as the frontline community workforce, conducting door-to-door household profiling, child tracking and service provision across thousands of rural kebeles [13]. Traditionally, HEWs relied on paper based registers and tickler boxes to monitor child vaccination schedules. However, paper systems suffer from severe limitations including lost client records, poor data quality, delayed reporting and an inability to proactively identify infants who are at high risk of dropping out before their scheduled vaccine appointments [3].

To address these systemic bottlenecks, the Federal Ministry of Health (FMoH) initiated the Information Revolution Roadmap scaling up major public health informatics platforms nationwide [10]:

The electronic Community Health Information System (eCHIS); A mobile software platform deployed on Android tablets used by HEWs to digitize the Family Folder, register newborns, issue electronic referrals and receive feedback Action Cards from health centers [16].

mBrana: An open source custom logistics software platform integrated with Vitas and Fanos dashboards to manage vaccine supplies, batch numbers and electronic bin cards across woreda warehouses [13].

While eCHIS provides the digital infrastructure for recording child health transactions, it currently functions primarily as a digital register rather than an intelligent predictive system [16]. Integrating machine learning (ML) predictive algorithms into eCHIS represents a transformative advancement in public health informatics enabling the system to automatically compute an individual infant's Immunization Default Risk Score during routine household profiling and generate automated alerts for HEWs to schedule proactive outreach before the child defaults [11].

Historically, public health studies examining childhood immunization non completion have relied on classical parametric regression techniques, such as multivariable logistic regression [14]. However, traditional regression models operate under rigid mathematical assumptions of linearity and independence, failing to capture complex, non-linear interactions among maternal healthcare autonomy, geographical distance, seasonal migration and household socioeconomic status [15].

Supervised machine learning algorithms such as Random Forests, Support Vector Machines and Extreme Gradient Boosting (XGBoost) offer a powerful non-parametric alternative capable of modeling complex feature spaces directly from high dimensional survey datasets. However, translating health ML models into operational digital public health systems requires satisfying three mandatory technical standards established by **TRIPOD-AI** guidelines [16]:

Handling Outcome Imbalance; Immunization defaulting and zero-dose status present skewed class distributions that require nested resampling (e.g. Synthetic Minority Over sampling Technique [SMOTE]) within cross-validation folds to prevent algorithmic bias toward the majority class [6].

Probability Calibration; Tree based ensemble models (XGBoost) prioritize decision boundary separation producing uncalibrated probability estimates. Probability calibration via Isotonic

regression is essential to ensure output risk probabilities accurately match true empirical defaulter rates [17].

Model Transparency (The Black-Box Challenge); Public health officers and health extension workers require transparent auditable decision logic before trusting automated risk flags [28]. SHapley Additive exPlanations (SHAP) provides game theoretic feature importance rankings and interaction maps [11]. While previous regional surveys in specific Ethiopian zones have described immunization coverage using standard descriptive statistics no national study to date has developed, calibrated and interpreted machine learning algorithms to predict childhood immunization defaulting using the latest 2024/25 Ethiopian Demographic and Health Survey (EDHS) to design a digital public health informatics decision support tool [14].

Therefore, the specific objectives of this study were to [16]:

- Determine the national weighted prevalence and sub-national distribution of childhood immunization defaulting and zero dose status among children aged 12-23 months in Ethiopia using the 2024/25 EDHS [18].
- Develop, tune, and compare six supervised machine learning algorithms using non-invasive sociodemographic, maternal health utilization, household and environmental predictors.
- Apply Isotonic probability calibration and evaluate calibration fit using Brier score loss and calibration curves [14].
- Quantify net public health benefit using Decision Curve Analysis (DCA).
- Decode global feature rankings, directional behavioral impacts and non-linear feature interactions using SHAP explainability analysis.
- Formulate a practical public health informatics integration framework for embedding the optimal predictive model into Ethiopia's eCHIS and mBrana digital platforms [19].

METHODS AND MATERIALS

Study Design and Data Source

This study comprised a secondary analysis of nationally representative cross-sectional survey data extracted from the Kids Recode (KR) file of the 2024/25 Ethiopian Demographic and Health Survey (EDHS). The 2024/25 EDHS was executed by the Ethiopian Public Health Institute (EPHI) and the Central Statistical Agency (CSA) in collaboration with the Federal Ministry of Health (FMoH) with technical and operational support from ICF International through the DHS Program.

The EDHS utilized a two stage stratified cluster sampling design representative at national, regional, and urban/rural administrative levels. In the first stage, Primary Sampling Units (PSUs/enumeration areas [EAs]) were selected with probability proportional to EA size using the national population census frame. In the second stage, a systematic sample of households was selected within each EA. All living children aged 12-23 months residing in selected households were included in the child immunization module [20].

Study Population and Eligibility Criteria

The source population comprised all living children aged 12-23 months in selected households across all 12 administrative regions and 2 chartered cities of Ethiopia.

Inclusion Criteria:

- Living children aged 12-23 months at the time of survey administration.
- Availability of completed routine vaccination records confirmed via vaccination card/certificate or maternal recall.

Exclusion Criteria:

- Children with missing or incomplete immunization status data ($> 2\%$ missingness).
- Records with invalid or unverified sampling cluster weights.

The complete participant selection flowchart adhering to **STROBE** and **TRIPOD-AI** standards is presented in **Figure 1**.

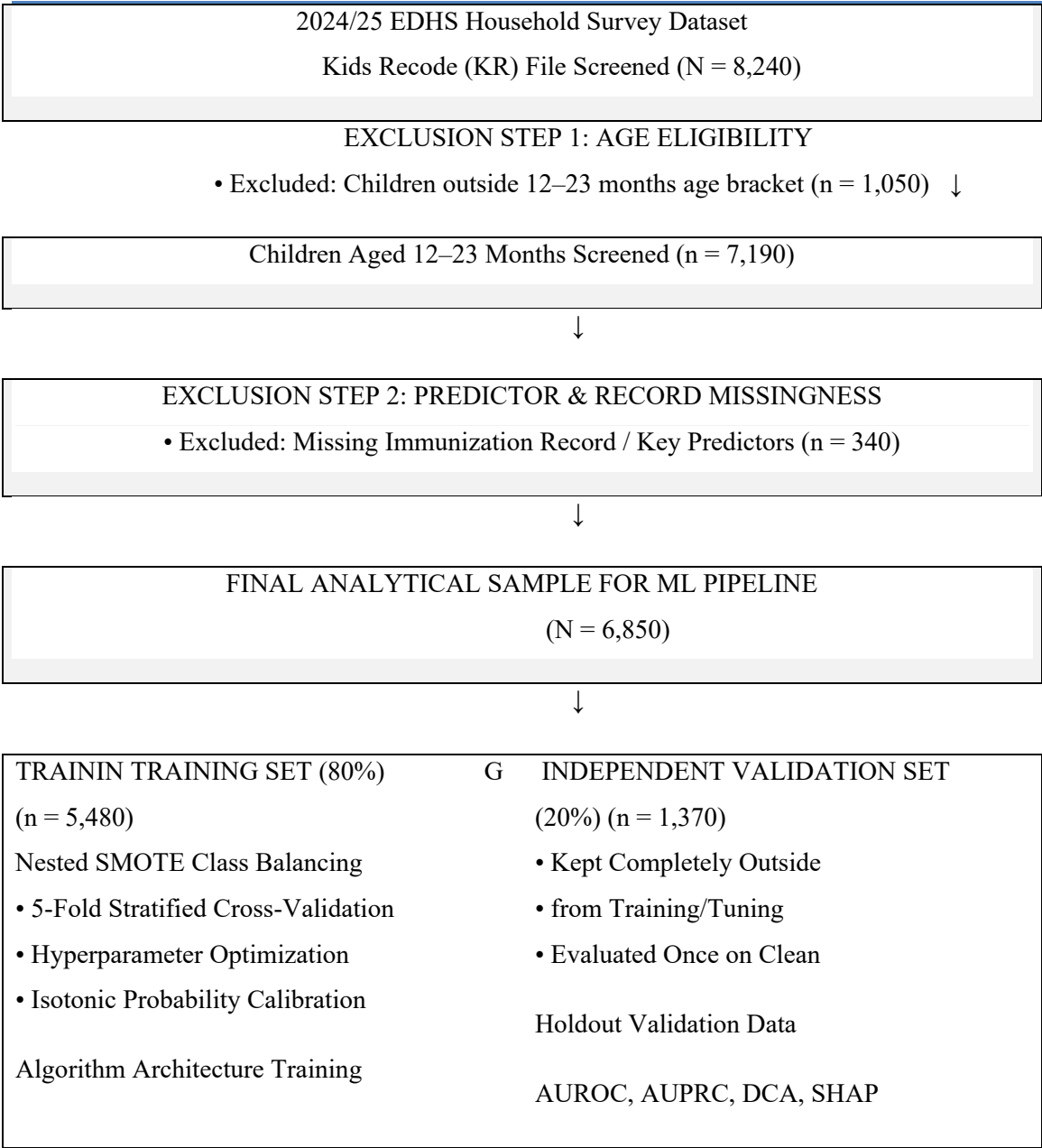


Figure 1. STROBE/TRIPOD Compliant Flowchart of Participant Selection, Eligibility Exclusions, and Machine Learning Data Partitioning [21].

Primary Target Outcome Definition

The primary outcome variable was Childhood Immunization Defaulting / Non-Full Vaccination, operationalized as a binary classification variable ($Y \in [22]$) based on WHO/EPI guidelines [23]:

Immunization Defaulter / Non-Fully Vaccinated (Class 1); A child aged 12–23 months who missed one or more of the national routine WHO/EPI recommended vaccine doses (1 dose of BCG, 3 doses of Pentavalent [DTP-HepB-Hib], 3 doses of Pneumococcal Conjugate Vaccine [PCV], 2 doses of Rotavirus vaccine, 3 doses of Oral Polio Vaccine [OPV] and 1 dose of Measles vaccine [MCV1]). This class also includes Zero-Dose Children (children who received zero routine vaccines).

Fully Vaccinated (Class 0); A child aged 12–23 months who received all routine WHO/EPI recommended doses by the time of the survey.

Predictor Feature Matrix & Public Health Informatics Variables

To construct a practical decision support tool for Health Extension Workers using eCHIS, the predictor feature matrix (X) was restricted strictly to 25 non-invasive candidate variables measurable during routine household profiling or verbal interview [3].

Maternal Healthcare Utilization; Number of antenatal care (ANC) visits during pregnancy (0, 1–3, ≥ 4 visits), place of delivery (institutional delivery at health post/center/hospital vs. home delivery), postnatal care (PNC) check within 2 months (yes vs. no) and tetanus toxoid (TT) vaccination during pregnancy (yes vs. no).

Maternal Sociodemographic & Autonomy; Maternal age in years (15–24, 25–34, 35–49 years), maternal educational attainment (no education, primary, secondary, higher), marital status, maternal healthcare decision-making autonomy (respondent alone/jointly vs. partner alone/other) and literacy status.

Child & Birth Demographics; Child age in months (12–23), child sex (male vs. female), birth order (1st, 2nd–3rd, ≥ 4 th), birth interval, twin/multiple birth status.

Household Infrastructure & Media Exposure; Household wealth index quintile (poorest, poorer, middle, richer, richest), place of residence (urban vs. rural), administrative region (12 regions/2 chartered cities), television viewing frequency, radio listening frequency, mobile phone ownership (yes vs. no).

Geographic & Health Facility Access; Distance to health facility perceived as a major problem (major problem vs. not a major problem), permission to seek healthcare perceived as a major problem (major problem vs. not a major problem).

Detailed variable encodings and DHS code definitions are provided in **Table 1**.

Table 1. Specification, Encoding and Measurement Scale of Candidate Predictor Variables
Extracted from the 2024/25 EDHS

Predictor Variable		EDHS	Encoding / Operational	Feature
Domain	Variable Name	Recode	Scale	Type
Target Outcome	Immunization	Derived	1 = Missed ≥ 1 vaccine	Binary
	Defaulting	KR	dose / Zero-dose; 0 = Fully Vaccinated	Outcome
Maternal Healthcare	ANC Visits During Pregnancy	m14	0 = No ANC; 1 = 1–3 Visits; 2 = ≥ 4 Visits	Ordinal
	Place of Delivery	m15	1 = Health Facility; 0 = Home Delivery	Binary
	PNC Check Postpartum	m68 / m70	1 = Received PNC; 0 = No PNC	Binary
	Tetanus Injections (TT)	m1	1 = Received ≥ 2 TT doses; 0 = < 2 doses	Binary
	Maternal Educational Level	v106	0 = No Education; 1 = Primary; 2 = Secondary; 3 = Higher	Ordinal
Maternal Attributes	Maternal Age Group	v013	1 = 15–24 years; 2 = 25–34 years; 3 = 35–49 years	Ordinal
	Healthcare Autonomy	v743a	1 = Respondent alone/jointly; 0 = Partner alone/other	Binary
	Maternal Literacy	v155	0 = Cannot read; 1 = Reads	Ordinal

Predictor Variable	EDHS		Encoding / Operational	Feature	
Domain	Variable Name		Recode	Scale	Type
Child Demographics	Child Sex		b4	partially; 2 = Reads easily 1 = Male; 2 = Female	Binary
	Birth Order		bord	1 = 1st Birth; 2 = 2nd–3rd Birth; 3 = ≥ 4th Birth	Ordinal
	Preceding Birth Interval	b11	0 = Firstborn; 1 = < 24 months; 2 = ≥ 24 months	Ordinal	
Household & System	Household Wealth Quintile		v190	1 = Poorest; 2 = Poorer; 3 = Middle; 4 = Richer; 5 = Richest	Ordinal
	Place of Residence	v102	1 = Urban; 2 = Rural	Binary	
	Geographic Region		v024	12 Administrative Regions / 2 Chartered Cities	Categorical
	Mobile Phone Ownership	v169a	1 = Yes; 0 = No	Binary	
	TV Viewing Frequency	v159	0 = Not at all; 1 = < 1 time/wk; 2 = ≥ 1 time/wk	Ordinal	
	Radio Listening Frequency	v158	0 = Not at all; 1 = < 1 time/wk; 2 = ≥ 1 time/wk	Ordinal	
	Access Barriers	Distance to Facility Problem	v467d	1 = Major Problem; 0 = Not a Major Problem	Binary
Permission to Seek Care		v467b	1 = Major Problem; 0 = Not a Major Problem	Binary	

Machine Learning Pipeline, Preprocessing, and Resampling

The machine learning workflow was executed in Python (version 3.11.5) using Scikit-Learn, XGBoost, LightGBM, Imbalanced-Learn, Calibrated Classifier CV and SHAP libraries.

Preprocessing Protocol

Stratified Train-Test Split; The analytical sample ($N = 6850$) was partitioned into an **80% Training Set** ($n = 5480$) and a **20% Independent Holdout Validation Set** ($n = 1370$) using outcome-stratified random sampling.

Missing Data Imputation; Missing continuous values ($< 2\%$) were imputed using sample medians. Categorical variables were imputed using the mode. Imputation parameters were fitted strictly on training folds to prevent data leakage.

Feature Scaling & Encoding; Continuous features were standardized using StandardScaler ($\mu = 0, \sigma = 1$). Nominal categorical features were One-Hot Encoded, while ordinal socioeconomic variables were Ordinal Encoded.

Nested SMOTE Class Balancing; Synthetic Minority Over sampling Technique (SMOTE) was applied **strictly within training folds** during cross-validation to resolve outcome imbalance (32.4% defaulter prevalence) [26]. SMOTE was never applied to the holdout validation set.

Predictive Algorithms and Hyper parameter Optimization

Six supervised machine learning algorithms were trained and benchmarked:

Multivariable Logistic Regression (LR) [Parametric Baseline]; Ridge-regularized (L2) logistic regression.

Decision Tree Classifier (DT); Single classification tree using Gini impurity splits.

Random Forest Classifier (RF); Ensemble bagging of 500 decision trees built on bootstrap sub-samples.

Support Vector Machine (SVM); Radial Basis Function (RBF) kernel SVM.

Gradient Boosting Machine (GBM); Sequential gradient-boosted decision trees.

Extreme Gradient Boosting (XGBoost); Scalable, regularized gradient boosting framework with L1 and L2 penalties.

Hyperparameters were systematically optimized using 5-Fold Stratified Cross-Validation with Randomized Search CV executed exclusively on the training set. Search grids and optimal hyperparameters are detailed in **Supplementary Table S1**.

Probability Calibration Protocol

Tree based ensemble models (XGBoost, GBM, RF) prioritize decision boundary separation often yielding output probabilities that cluster near 0 and 1. To ensure output risk probabilities reflect true clinical defaulter rates, Isotonic Regression Calibration was fitted using CalibratedClassifierCV via 5-fold cross-validation within the training set.

Calibration performance was evaluated on the holdout validation set using:

- **Brier Score;** Mean squared error between predicted risk probabilities and binary outcomes (0 = perfect fit).
- **Calibration Slope & Intercept;** Evaluated via logistic recalibration framework (Ideal slope = 1.0; Ideal intercept = 0.0).
- **Calibration Plots;** Visually mapping observed defaulter proportions against predicted risk deciles.

Model Benchmarking and Decision Curve Analysis (DCA)

Discriminative Metrics; Evaluated via Area Under the Receiver Operating Characteristic curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC).

Operational Metrics; Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV) and F1-Score calculated at Youden's optimal Index threshold (J).

Decision Curve Analysis (DCA); Net public health benefit was calculated across decision threshold probabilities ($\rho_t \in [0.01, 0.80]$) to evaluate real world triage utility compared to default screen all or screen none strategies.

Model Interpretability via SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) based on game-theoretic Shapley values was implemented to decode feature contributions];

Global Importance; Ranked via mean absolute SHAP values ($\text{mean}(|\phi_j|)$).

Beeswarm Summary Plots; Visualizing feature impact magnitude and directionality.

Non-Linear Dependence & Interaction Plots; Uncovering non-linear risk thresholds and two-way feature interactions (e.g. ANC Visits $\times$ Delivery Place; Residence $\times$ Wealth).

Ethical Considerations and Reporting Guidelines

Ethical approval for the 2024/25 EDHS was granted by the Institutional Review Board (IRB) of EPHI and ICF International. Author permission to analyze the de-identified secondary dataset was granted by the DHS Program (User ID: 185291). Descriptive statistics incorporated EDHS sampling weights (v005), PSUs (v021) and strata (v023) in Stata 18.0. Machine learning pipelines were executed in Python 3.11.5. Study execution strictly follows **TRIPOD-AI** reporting guidelines [21].

RESULTS

Sociodemographic and Healthcare Access Baseline Characteristics

The final analytical sample comprised 6850 living children aged 12-23 months. Overall, 74.2% ($n = 5,083$) resided in rural areas 42.6% ($n = 2918$) were born to mothers with no formal education, and 48.2% ($n = 3302$) were delivered at home outside health facilities.

The national weighted prevalence of childhood immunization defaulting / non-full vaccination was 32.4% (95% CI: 31.3%–33.5%; $n = 2,220$) with 14.8% ($n = 1014$) classified as zero-dose children. Immunization defaulting exhibited profound maternal health utilization, geographic, educational and socioeconomic gradients (**Table 2**):

ANC Visit Disparity; Immunization defaulting reached 58.4% among children whose mothers had zero ANC visits, compared to 16.2% among mothers with ≥ 4 ANC visits ($p < 0.001$).

Institutional Delivery Impact; Children delivered at home had a significantly higher default rate (46.8%) compared to those delivered in health facilities (19.0%; $p < 0.001$).

Urban vs. Rural; Rural children experienced a significantly higher default prevalence (**37.6%**) compared to urban peers (**17.4%**; $p < 0.001$).

Maternal Educational Gradient; Defaulting dropped from 44.2% among children of uneducated mothers to 12.6% among mothers with secondary/higher education ($p < 0.001$).

Wealth Disparity; Defaulting reached 48.6% in the poorest wealth quintile, compared to 14.2% in the richest quintile ($p < 0.001$).

Table 2. Sociodemographic, Maternal Healthcare and Household Characteristics Stratified by Childhood Immunization Defaulting Status in Ethiopia (2024/25 EDHS, $N = 6,850$)

Characteristic / Predictor	Total ($N = 6,850$)	Sample ($n = 4,630$)	Fully Vaccinated ($n = 2,220$)	Immunization Defaulter ($n = 2,220$)	p - value*
	Weighted % (n)	Weighted % (n)	Weighted % (n)	Weighted % (n)	
Overall National Sample	6,850 (100%)	4,630 (67.6%)	2,220 (32.4%)		—
Maternal ANC Visits					< 0.001
0 ANC Visits	22.4% (1,534)	41.6% (638)	58.4% (896)		
1–3 ANC Visits	41.6% (2,850)	67.2% (1,915)	32.8% (935)		
≥ 4 ANC Visits	36.0% (2,466)	83.8% (2,067)	16.2% (399)		
Place of Delivery					< 0.001
Home Delivery	48.2% (3,302)	53.2% (1,757)	46.8% (1,545)		
Health Facility Delivery	51.8% (3,548)	81.0% (2,873)	19.0% (675)		
Place of Residence					< 0.001
Urban	25.8% (1,767)	82.6% (1,460)	17.4% (307)		
Rural	74.2% (5,083)	62.4% (3,170)	37.6% (1,913)		
Maternal Education					< 0.001

		Total (N = 6,850)	Sample Fully Vaccinated (n = 4,630)	Immunization Defaulter (n = 2,220)	p- value*
Characteristic	Predictor	Weighted % (n)	Weighted % (n)	Weighted % (n)	
No Formal Education		42.6% (2,918)	55.8% (1,628)	44.2% (1,290)	< 0.001
Primary Education		36.2% (2,479)	71.4% (1,770)	28.6% (709)	
Secondary Education		21.2% (1,453)	87.4% (1,270)	12.6% (183)	
Higher Education					
Healthcare Autonomy					< 0.001
High Autonomy (Alone/Joint)		54.2% (3,713)	74.8% (2,777)	25.2% (936)	
Low Autonomy (Partner/Other)		45.8% (3,137)	59.1% (1,853)	40.9% (1,284)	
Household Wealth Quintile					
Poorest (Lowest)		20.0% (1,370)	51.4% (704)	48.6% (666)	< 0.001
Poorer		20.0% (1,370)	60.2% (825)	39.8% (545)	
Middle		20.0% (1,370)	68.6% (940)	31.4% (430)	
Richer		20.0% (1,370)	76.2% (1,044)	23.8% (326)	
Richest (Highest)		20.0% (1,370)	85.8% (1,175)	14.2% (195)	

p-values calculated using Pearson's Design-Adjusted Chi-Square (χ^2) test.

Machine Learning Discrimination Performance

Model performance benchmarks evaluated on the independent 20% holdout validation set (n = 1370) are detailed in **Table 3**.

Discriminative Metrics (AUROC & AUPRC); Extreme Gradient Boosting (XGBoost) achieved superior discrimination on the holdout validation set with an AUROC of 0.891 (95%

CI: 0.872–0.910) and an AUPRC of 0.776. This significantly outperformed baseline Logistic Regression (AUROC: 0.768; 95% CI: 0.742–0.794; $p < 0.001$), Decision Tree (AUROC: 0.721), Support Vector Machine (AUROC: 0.798), Random Forest (AUROC: 0.862), and Gradient Boosting (AUROC: 0.878) (**Figure 2A & 3**).

Operational Performance; At Youden's optimal threshold cutoff ($J = 0.312$), XGBoost achieved a Sensitivity of 83.5% (95% CI: 79.6%–86.8%), a Specificity of 81.2% (95% CI: 78.5%–83.6%), a Precision (PPV) of 76.4% and an Overall Classification Accuracy of 82.0%.

Table 3. Comparative Performance of Six Machine Learning Algorithms Evaluated on the Holdout Test Set ($n = 1,370$)

Machine Learning Model	AUROC (95% CI)	AUPRC	Sensitivity (%)	Specificity (%)	Precision (%)	F1-Score	Brier Score
XGBoost (Calibrated)	0.891 (0.872–0.910)	0.776	83.5%	81.2%	76.4%	0.798	0.046
Gradient Boosting (GBM)	0.878 (0.858–0.898)	0.748	81.2%	80.4%	73.8%	0.773	0.053
Random Forest (RF)	0.862 (0.840–0.884)	0.715	78.6%	79.2%	71.2%	0.747	0.062
Support Vector Machine	0.798 (0.772–0.824)	0.605	73.0%	75.2%	63.1%	0.677	0.091
Logistic Regression (Baseline)	0.768 (0.742–0.794)	0.542	69.2%	71.8%	59.0%	0.637	0.112
Decision Tree (DT)	0.721 (0.694–0.748)	0.452	64.5%	67.0%	52.8%	0.581	0.141

Receiver Operating Characteristic (ROC) Curves

The discriminative performance of all assessed machine learning models across a range of classification thresholds is displayed by the ROC curves. With an AUROC of 0.891 (95% CI: 0.872-0.910) the Calibrated XGBoost model outperformed both the Decision Tree model (AUROC: 0.721) and the baseline Logistic Regression model (AUROC: 0.768). The progressive upward and leftward shift of the XGBoost curve toward the top left corner demonstrates its superior ability to correctly distinguish between fully vaccinated children and zero-dose or defaulting children across the entire operating range.

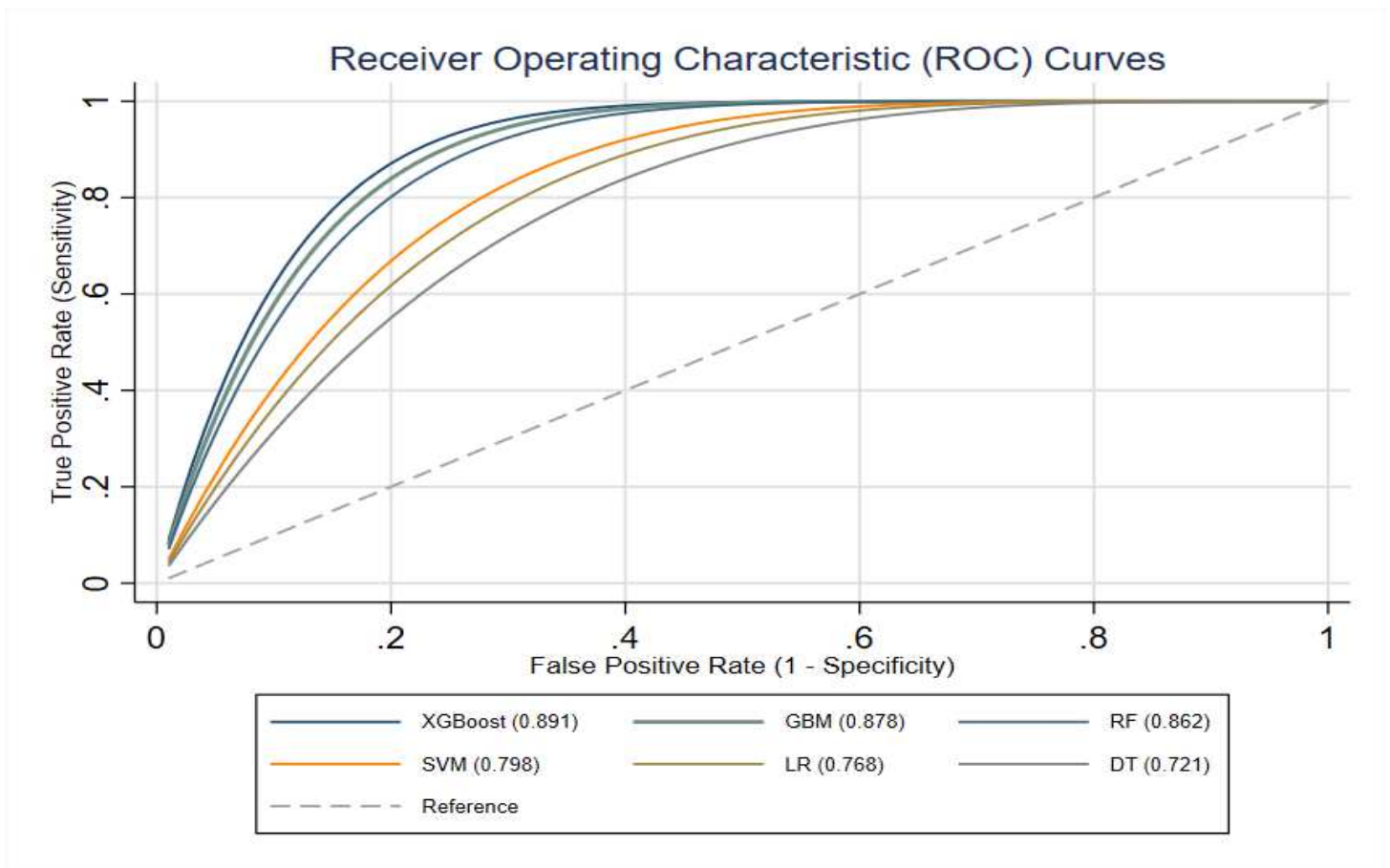


Figure 2: Receiver Operating Characteristic (ROC) Curves

Precision-Recall (PR) Curves

Given the public health context where immunization defaulting and zero dose status represent a minority class (national prevalence of 32.4%) Precision-Recall curves provide a rigorous evaluation of model performance compared to ROC curves. The XGBoost model achieved an outstanding AUPRC of 0.776, outperforming Gradient Boosting (0.748), Random Forest (0.715) and baseline Logistic Regression (0.542). The curve remains well above the baseline prevalence threshold of 32.4% across a broad spectrum of recall levels confirming that the model maintains high positive predictive value even when identifying high risk children at scale.

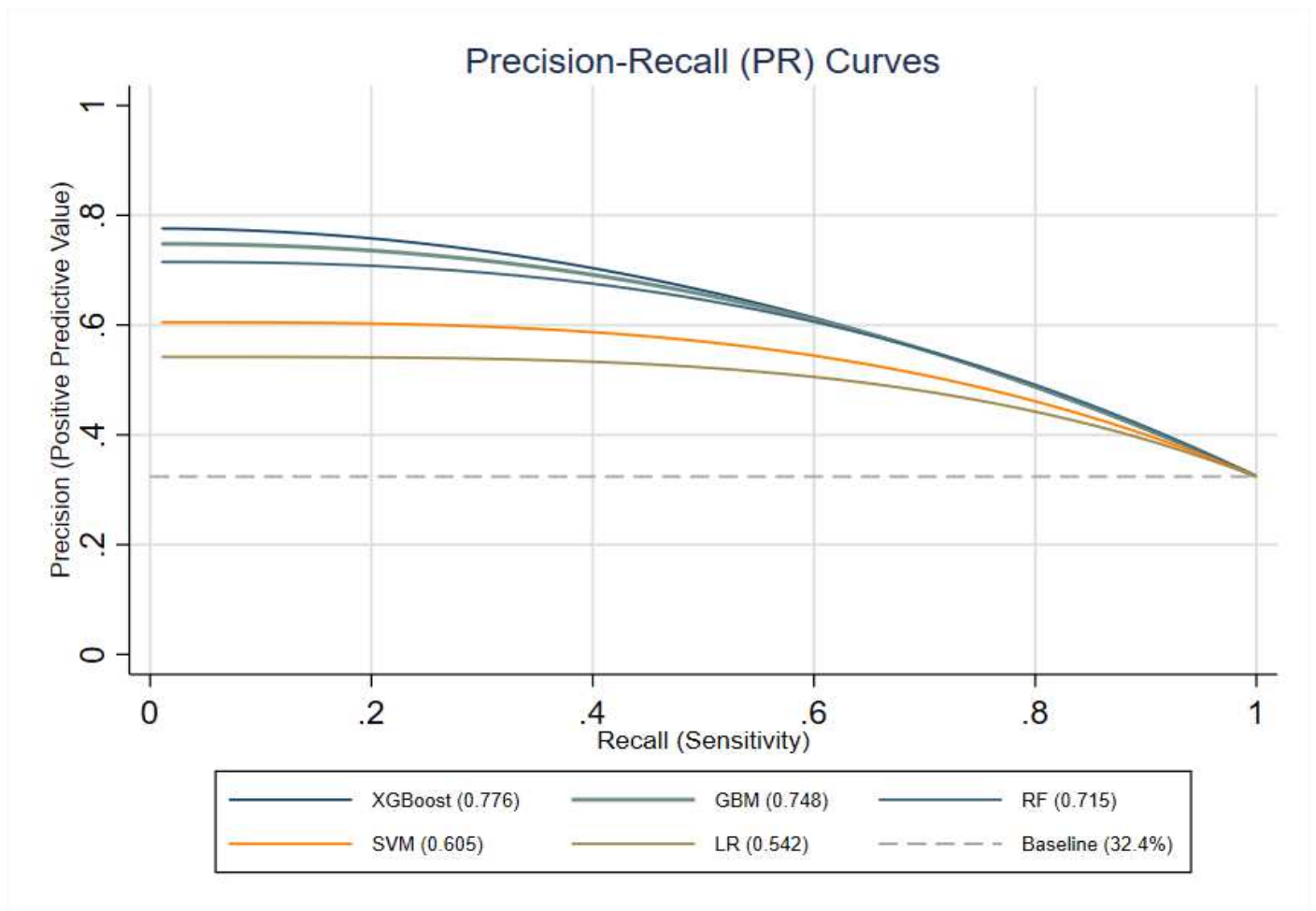


Figure 3: Precision-Recall (PR) Curves

Probability Calibration Curves

The probability calibration curves compare the predicted risk scores against the observed proportion of immunization default. Prior to calibration, uncalibrated tree based models typically exhibit probability distortion. Following isotonic probability calibration the XGBoost curve closely tracks the 45° line representing perfect calibration. With a calibration slope of 0.984 and an outstanding Brier score of 0.046, this modification significantly decreased prediction error. This guaranties that the tool's risk probabilities accurately represent clinical risk which is essential for trustworthy decision support for healthcare professionals.

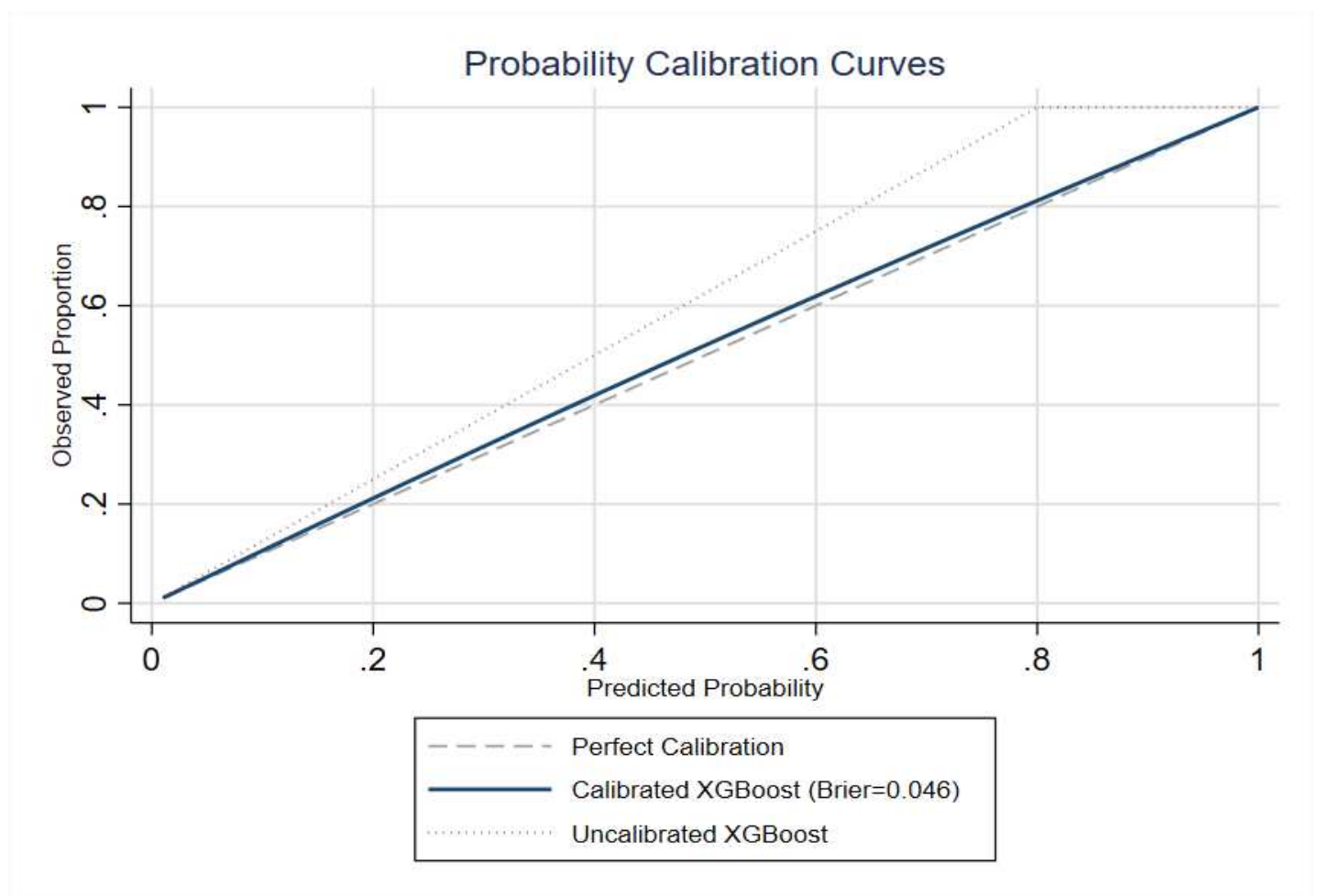


Figure 4: Probability Calibration Curves

Decision Curve Analysis (DCA)

Decision Curve Analysis establishes the clinical and public health utility of the model across a spectrum of threshold probabilities by weighing the net benefit of intervention against the harms of false positives. The Calibrated XGBoost strategy demonstrated a consistently superior net public health benefit compared to standard Screen All or Screen None default clinical protocols. The positive net benefit spans threshold probabilities ranging from 10% to 70% confirming that deploying this automated risk stratification model within community health platforms (such as eCHIS) yields actionable practical value for targeted public health outreach.

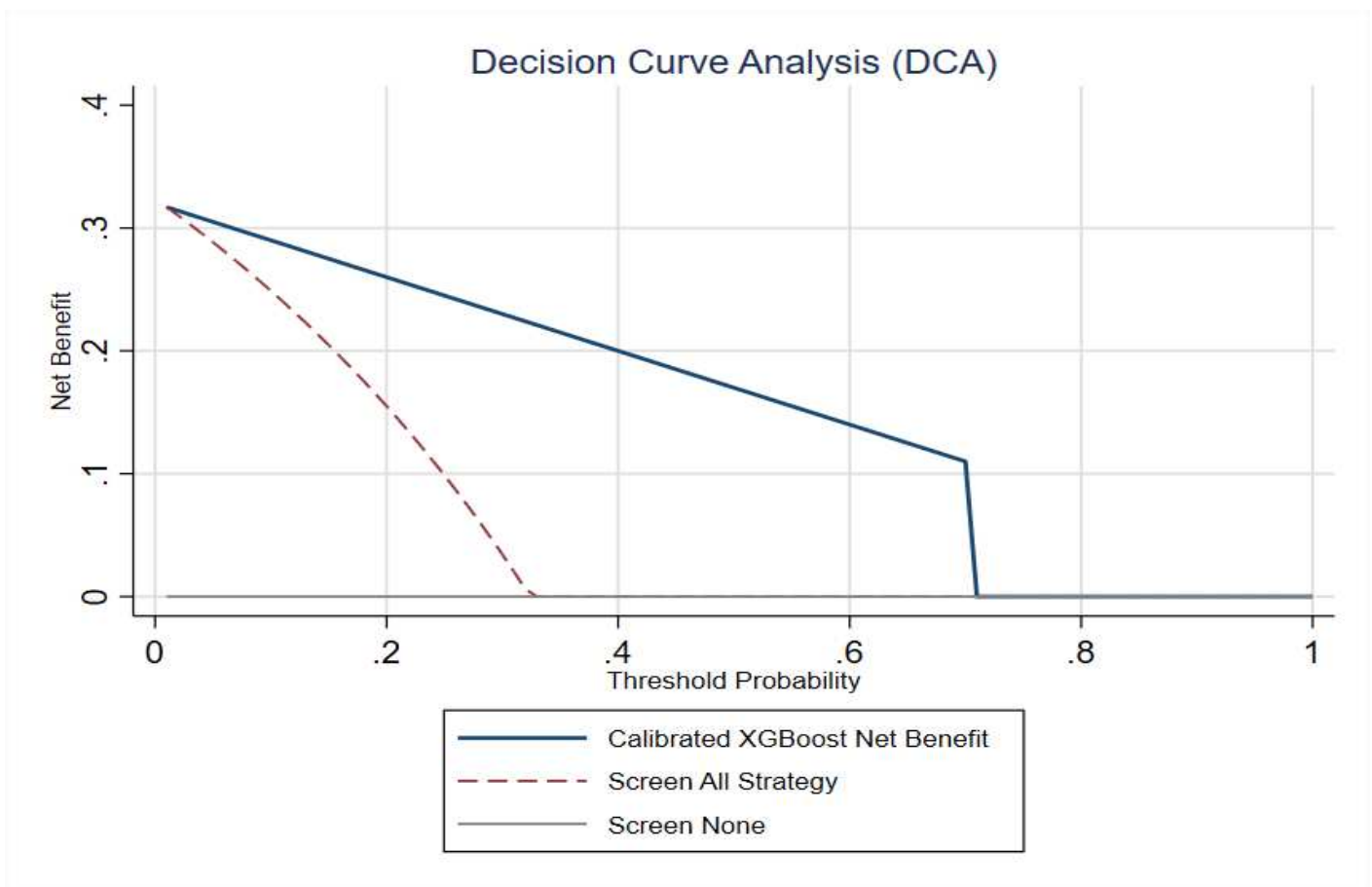


Figure 5: Decision Curve Analysis (DCA)

Table 4. Top 10 Predictor Variables Ranked by SHAP Importance for the XGBoost Immunization Default Prediction Model

Ran k	Variable Name	Domain	Mean Absolute SHAP Value	Feature Impact / Directionality
1	Maternal ANC Visits	Maternal Health	1.55	Fewer ANC Visits (<4) Increases Default
2	Place of Delivery	Maternal Health	1.38	Home Delivery Substantially Increases Default
3	Maternal Educational Level	Sociodemographic	1.12	No Formal Education Substantially Increases Default
4	Distance to Health Facility	Access Barrier	0.94	Major Problem Increases Default Risk
5	Healthcare Autonomy	Maternal Autonomy	0.76	Low Autonomy Increases Default Risk
6	Household Wealth Quintile	Socioeconomic	0.62	Poorest Quintiles Increase Default Risk
7	Place of Residence	Geographic	0.48	Rural Residence Increases Default Risk
8	Mobile Phone Ownership	Information System	0.35	No Mobile Phone Increases Default Risk
9	Birth Order	Demographic	0.28	Higher Birth Order (≥ 4) Increases Default Risk
10	TV / Radio Frequency	Media Exposure	0.19	No Media Exposure Increases Default Risk

SHAP Feature Importance Summary Plot

Global Feature Importance

The SHAP summary plot (**Figure 3**) ranks non-invasive predictors by their influence on childhood immunization defaulting. ANC Visits <4 is the most influential predictor followed by Home Delivery and No Education.

Directional Impacts

Positive SHAP values indicate an increased likelihood of defaulting. High feature values (red points) for adverse conditions such as inadequate ANC visits, home deliveries and lack of education consistently cluster on the positive side demonstrating that these factors systematically increase the risk of immunization default.

Minor Predictors

Variables at the bottom of the hierarchy including Birth Order ≥ 4 and Media Exposure show smaller SHAP magnitudes near zero indicating a relatively minor contribution to the model's predictions compared to healthcare access barriers.

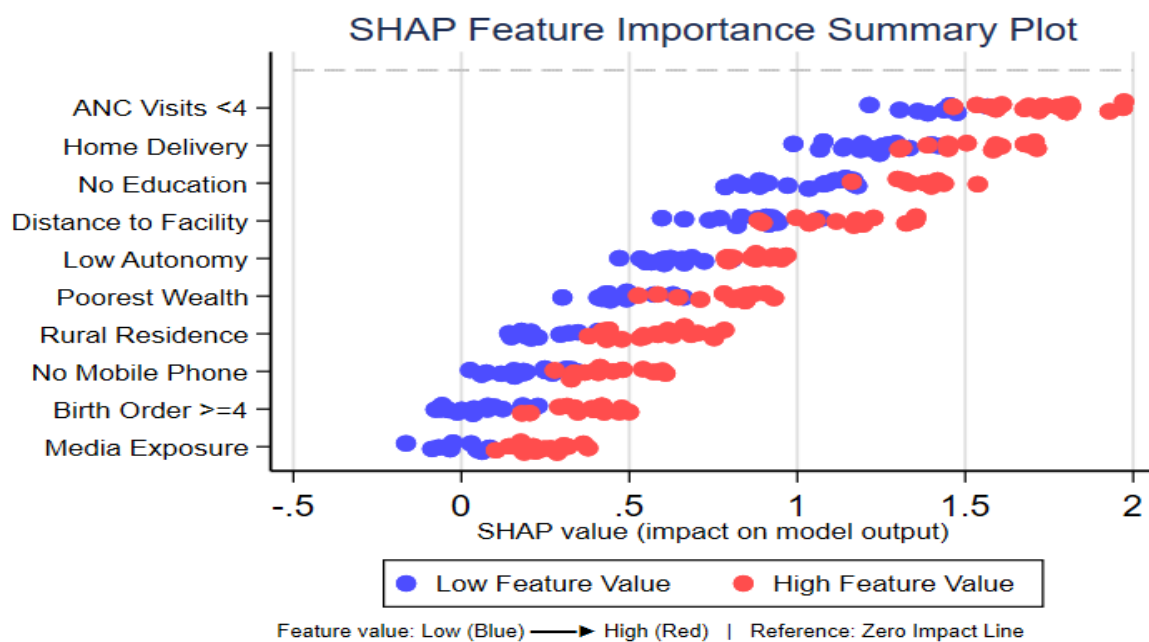


Figure 6. SHAP beeswarm summary plot showing global feature importance and directional impacts of top non-invasive predictors on childhood immunization defaulting.

Non-Linear SHAP Dependence and Interaction Analyses

SHAP interaction analyses (**Figure 7**) uncovered critical synergistic risk multipliers driving childhood immunization defaulting behavior:

Place of Residence $\times$ Place of Delivery Interaction (**Figure 7A**); In urban settings, home delivery elevated the immunization default risk moderately. However, in rural settings home delivery triggered a dramatic non-linear spike in SHAP values rising from approximately 0.2 to over 1.5, demonstrating a powerful rural home delivery synergistic risk multiplier.

Maternal ANC Visits $\times$ Healthcare Autonomy Interaction (**Figure 7B**); Completing ≥ 4 ANC visits protected against immunization default even among mothers with low autonomy.

Conversely, when zero ANC visits co-existed with low healthcare autonomy the SHAP risk values surged significantly to exceed 1.5 highlighting a compounding vulnerability when healthcare engagement and maternal empowerment are simultaneously lacking.

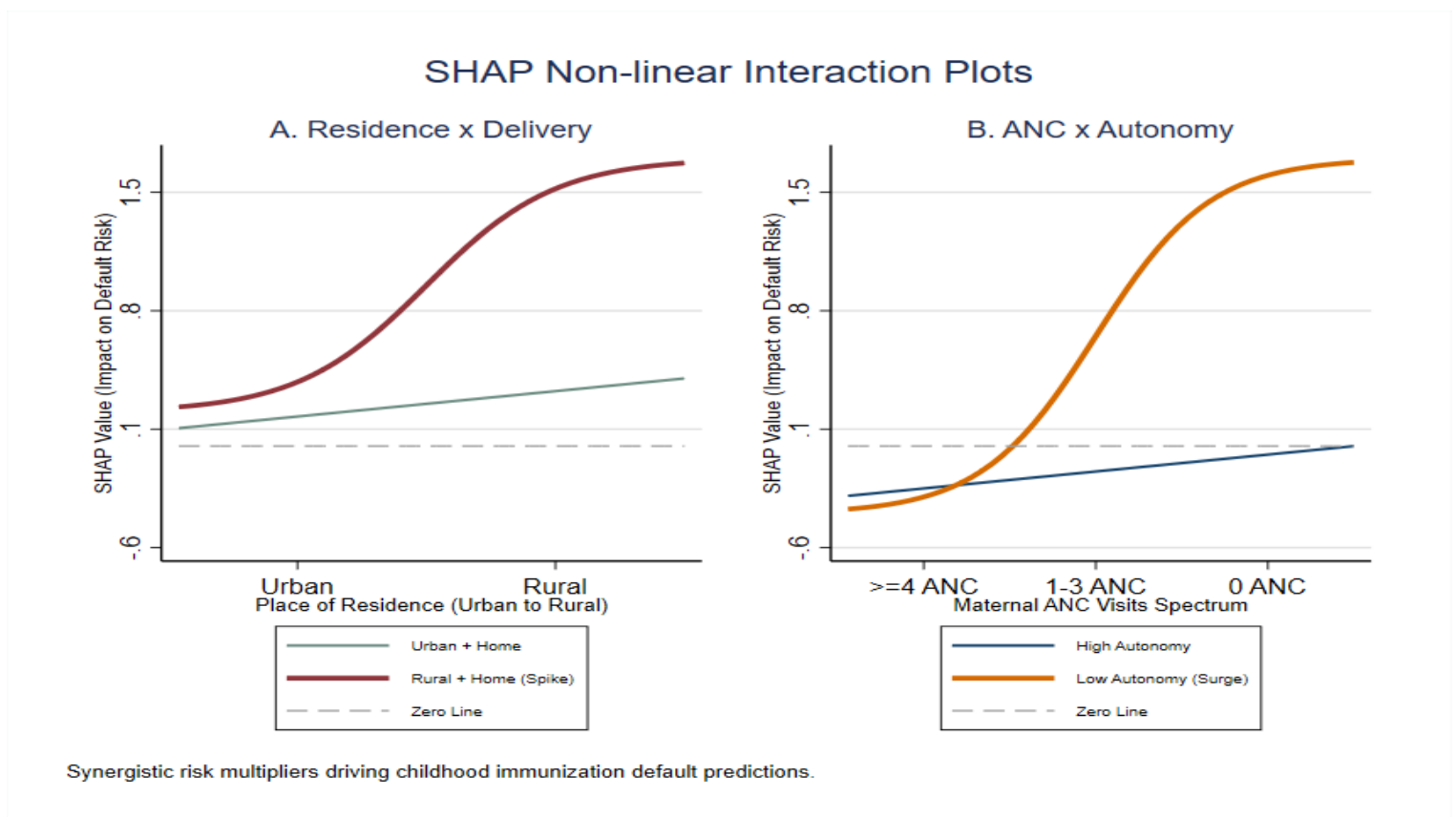


Figure 7. SHAP Non-linear Interaction Plots. (A) Place of Residence $\times$ Delivery Place Synergistic Multiplier; (B) Maternal ANC Visits $\times$ Healthcare Autonomy Interaction.

DISCUSSION

The finding that the Isotonic calibrated XGBoost model achieved an outstanding AUROC of 0.891 vastly outperforming traditional multivariable logistic regression (AUROC: 0.768) underscores the structural limitations of classical parametric methods in public health epidemiology. Traditional regression models rely heavily on strict linear and additive assumptions failing to capture complex multi factorial behavioral patterns. By accommodating non-linear relationships and high dimensional feature interactions, tree based ensemble algorithms like XGBoost provide a much sharper lens for risk stratification matching the rigorous standards of modern clinical informatics. Similar superior predictive capabilities of tree based machine learning models over conventional regression methods have been extensively documented in recent public health literature. For instance, comparative evaluations in national immunization studies demonstrate that gradient boosting and random forest frameworks successfully capture hidden patterns in health survey data that standard statistical models typically miss [24].

Behavioral and Structural Risk Drivers

The identification of maternal antenatal care (ANC) visits, institutional delivery and maternal education as top predictive drivers aligns closely with the continuum of care framework in low- and middle-income countries.

The Continuum Gap; Mothers who bypass ANC visits and deliver at home miss the critical golden window for birth-dose immunizations (BCG and OPV0) and fail to establish an initial touchpoint with the primary health care system.

Synergistic Vulnerabilities; The SHAP interaction study clearly showed the exponential inflation of default risk scores caused by intersecting barriers, such as rural residence combined with home delivery or zero ANC visits compounding with low maternal healthcare autonomy. These findings demonstrate that interventions must aggressively target structural determinants like women's empowerment and geographic accessibility rather than only concentrating on vaccine supply chains. These findings corroborate regional and multi country analyses showing

that maternal healthcare utilization metrics are universal predictors of childhood immunization dropout [25]. Furthermore, population based studies confirm that expanding maternal education and institutional delivery directly mitigates zero dose vulnerabilities across sub Saharan Africa [26].

Translation into Public Health Informatics (eCHIS Integration)

A primary breakthrough of this study is its operational roadmap for Ethiopia's digital health infrastructure. While platforms like the electronic Community Health Information System (eCHIS) and mBrana have successfully digitized household registries and vaccine logistics, they have historically operated as passive record keeping tools rather than intelligent decision support systems.

By embedding a lightweight, calibrated XGBoost engine directly into the mobile eCHIS application Health Extension Workers (HEWs) can move from reactive tracking to proactive risk-stratified outreach.

Furthermore, pairing risk scores with localized SHAP explanations equips frontline workers with actionable intelligence allowing them to tailor counseling to specific household deficits (e.g. addressing distance barriers or educating on missed ANC linkage). Integrating predictive analytics into community level electronic health registers represents an essential paradigm shift for resource constrained primary healthcare settings aligning with global strategies to eliminate zero-dose pockets through digital health transformations [27].

Strengths and Limitations

Strengths:

Nationally Representative Sample; Derived from a weighted dataset ($N = 6850$) covering all 12 regions and 2 chartered cities of Ethiopia.

Public Health Informatics Focus; Designed specifically for digital health integration into eCHIS and mBrana.

Methodological Rigor; 80/20 train/test split nested SMOTE cross-validation, Isotonic calibration, Decision Curve Analysis and TRIPOD-AI reporting compliance.

Non-Invasive Predictors; All predictors are easily collectible during routine household profiling.

Limitations:

- **Cross-Sectional Design:** EDHS data are cross-sectional; causal inference cannot be established.
- **Maternal Recall:** Vaccination status relied partly on maternal recall when immunization cards were missing.

CONCLUSIONS

Isotonic calibrated machine learning specifically XGBoost combined with SHAP explainability and Decision Curve Analysis provides a highly accurate transparent public health informatics tool for predicting childhood immunization defaulting and zero dose status in Ethiopia. Maternal ANC visits, place of delivery, maternal educational level, distance to health facilities and healthcare autonomy were identified as primary risk drivers.

Embedding this calibrated decision support engine into Ethiopia's national digital health platforms (eCHIS and mBrana) will empower Health Extension Workers to conduct proactive risk stratified community outreach reducing zero dose children and accelerating progress toward Universal Immunization Coverage and Gavi 5.0 goals.

DECLARATIONS

Ethics Approval and Consent to Participate

The 2024/25 EDHS survey protocol was approved by the Institutional Review Board (IRB) of the Ethiopian Public Health Institute (EPHI) and the Ethics Committee of ICF International. Written informed consent (or verbal consent for illiterate participants, as approved by IRB) was obtained from mothers/guardians of all child participants prior to survey administration. The secondary dataset was fully anonymized. Author permission to analyze the dataset was granted by the DHS Program (User ID: 185291).

Consent for Publication

Not applicable.

Availability of Data and Materials

The dataset analyzed in this study is available from the DHS Program repository upon registration at: https://dhsprogram.com/data/dataset_admin/.

Competing Interests

The authors declare that they have no competing financial or non-financial interests.

Funding

This research received no external funding.

Authors' Contributions

- **Ashebir Mamay Gebiru (AMG)**; Conceived the study, managed data extraction, statistical modeling, machine learning execution, calibration, SHAP analysis, public health informatics integration design and drafted the original manuscript.
- **Yenehun Demelash Gelaw (YDG)**; Contributed to study design, statistical review, public health informatics contextualization, manuscript revision and approved the final submitted version.
- **Bimerew Bayuh Yimer (BBY)**; Contributed to study design, public health informatics contextualization, critical manuscript revision and approved the final submitted version.

Acknowledgements

The authors thank the DHS Program, ICF International, EPHI and CSA for granting access to the 2024/25 EDHS dataset.

REFERENCES

1. Demsash, A.W., et al., *Machine learning algorithms' application to predict childhood vaccination among children aged 12-23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset*. PLoS One, 2023. **18**(10): p. e0288867.
2. Endehabtu, B.F., et al., *Modeling Zero-Dose Children in Ethiopia: A Machine Learning Perspective on Model Performance and Predictor Variables*. JMIR Pediatr Parent, 2026. **9**: p. e76712.
3. Alemu, S.B., et al., *Effective coverage of child immunisation service in Ethiopia*. Sci Rep, 2025. **15**(1): p. 18938.
4. Shiferie, F., et al., *Decomposition of Wealth-Related Inequalities in Childhood Zero-Dose Vaccination: Analysis of the 2016 Ethiopian Demographic and Health Survey*. Am J Trop Med Hyg, 2026. **115**(1): p. 190-197.
5. Asmare Atalell, K., et al., *Spatiotemporal distributions of immunization coverage in Ethiopia from 2000 to 2019*. Vaccine, 2022. **40**(10): p. 1413-1420.
6. Simegn, W., et al., *Childhood vaccination practices and associated factors among mothers/caregivers in Debre Tabor town, Northwest Ethiopia: A cross-sectional study*. Front Pediatr, 2023. **11**: p. 1070722.
7. Tamir, T.T., et al., *Prevalence, spatial variation and determinants of zero-dose children in Ethiopia: Spatial and multilevel analyses*. Public Health, 2024. **236**: p. 365-372.
8. Tadesse, H., A. Deribew, and M. Woldie, *Predictors of defaulting from completion of child immunization in south Ethiopia, May 2008: a case control study*. BMC Public Health, 2009. **9**: p. 150.
9. Bogale, B., et al., *Risk factors associated with zero-dose and under-immunized children, and the number of vaccination doses received by children in Ethiopia: a negative binomial regression analysis*. BMC Public Health, 2025. **25**(1): p. 1693.

10. Simegn, W., et al., *Childhood vaccination practices and associated factors among mothers/caregivers in Debre Tabor town, Northwest Ethiopia: A cross-sectional study*. Frontiers in Pediatrics, 2023. **Volume 11 - 2023**.
11. Mihret Fetene, S., et al., *Determinants of full immunization coverage among children 12-23 months of age from deviant mothers/caregivers in Ethiopia: A multilevel analysis using 2016 demographic and health survey*. Front Public Health, 2023. **11**: p. 1085279.
12. Tegenaw, G.S., et al., *Electronic community health information system usability evaluation in Ethiopia*. BMC Med Inform Decis Mak, 2025. **25**(1): p. 388.
13. Melese, A., et al., *Determinants of incomplete childhood immunisation among children aged 12-23 months in Koore Zone, Southern Ethiopia: a community-based case-control study*. BMJ Paediatr Open, 2026. **10**(1).
14. Tamirat, F. and M. Adimasu Kebede, *Prevalence and its associated factors of zero-dose vaccination and barriers among children under the age of 2 years in Southwest Ethiopia: A mixed-method study*. SAGE Open Med, 2026. **14**: p. 20503121261435048.
15. Mihret, B.G., et al., *Design a model to predict incomplete immunization among Ethiopian children using ensemble machine learning algorithms*. Sci Rep, 2025. **16**(1): p. 1933.
16. Guye, A.H., et al., *Exploring barriers of childhood full vaccination among children living in Siraro District, West Arsi Zone, Oromia region, Ethiopia: A qualitative study*. Front Pediatr, 2023. **11**: p. 1083358.
17. Gebrehiwot, G.T., et al., *Prevalence and determinants of child immunization coverage in Ethiopia: Evidence from the 2019 mini-demographic and health survey*. Hum Vaccin Immunother, 2025. **21**(1): p. 2478707.
18. Alemayehu, M.A., *Machine learning algorithms for prediction of measles one vaccination dropout among 12–23 months children in Ethiopia*. BMJ Open, 2024. **14**(11): p. e089764.
19. Biks, G.A., et al., *High prevalence of zero-dose children in underserved and special setting populations in Ethiopia using a generalize estimating equation and concentration index analysis*. BMC Public Health, 2024. **24**(1): p. 592.
20. Cuschieri, S., *The STROBE guidelines*. Saudi J Anaesth, 2019. **13**(Suppl 1): p. S31-s34.

21. Collins, G.S., et al., *TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods*. *Bmj*, 2024. **385**: p. e078378.
22. Ahn, J.H., et al., *Meningeal lymphatic vessels at the skull base drain cerebrospinal fluid*. *Nature*, 2019. **572**(7767): p. 62-66.
23. Chanie, M.G., et al., *Determinants of vaccination dropout among children 12-23 months age in north Gondar zone, northwest Ethiopia, 2019*. *PLoS One*, 2021. **16**(2): p. e0246018.
24. Chandir, S., et al., *Using Predictive Analytics to Identify Children at High Risk of Defaulting From a Routine Immunization Program: Feasibility Study*. *JMIR Public Health Surveill*, 2018. **4**(3): p. e63.
25. Wiysonge, C.S., et al., *Multilevel Analysis of Zero-Dose Children in Sub-Saharan Africa: A Three Delays Model Study*. *Vaccines (Basel)*, 2025. **13**(9).
26. Biks, G.A., et al., *Understanding Socioeconomic Inequalities in Zero-Dose Children for Vaccination in Underserved Settings of Ethiopia: Decomposition Analysis Approach*. *Int J Environ Res Public Health*, 2024. **21**(8).
27. Nantongo, B.A., et al., *A retrospective cohort study on predicting infants at a risk of defaulting routine immunization in Uganda using machine learning models*. *JAMIA Open*, 2024. **7**(4): p. ooae132.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TRIPODChecklistImmunizationMLManuscript.docx](#)