

RESEARCH ARTICLE

Explainable Machine Learning Framework for Early Prediction of Adverse Pregnancy Outcomes in Low-and Middle-Income Countries

Ifeoluwa OSUNGBURE¹, David Peter AKPAN², Ikechukwu Victor OGOKE³, Samson Dayo EJIKUNLE⁴, Maria-Joannes UZOMA⁵

¹Middle Tennessee State University, Department of Economics and Finance, USA.

²Ahmadu Bello University, Zaria, Nigeria.

^{3,4,5}Imo State University Teaching Hospital, Orlu, Nigeria.

Received: 01 October 2026 Accepted: 16 October 2026 Published: 23 October 2026

Corresponding Author: Ifeoluwa OSUNGBURE, Middle Tennessee State University, Department of Economics and Finance, USA.

Abstract

Adverse pregnancy outcomes, including pre-eclampsia, preterm birth, low birth weight and stillbirth, remain a leading source of maternal and perinatal death, and roughly nine in ten of these deaths occur in low- and middle-income countries (LMICs) where specialist obstetric assessment and laboratory testing are scarce. Risk scoring that works at the first antenatal contact, using only information a frontline health worker can collect, could redirect scarce referral capacity toward the pregnancies most likely to deteriorate. A transparent reproducible framework was built and interrogated for early prediction of a composite adverse pregnancy outcome on a simulated LMIC antenatal cohort of 10,800 pregnancies (10.6% event prevalence) whose data-generating process was calibrated to published epidemiological parameters and restricted to twenty-five variables obtainable at a first booking visit. Seven classifiers were tuned with randomised search under stratified cross-validation. Interpretation combined a model-agnostic Monte-Carlo Shapley estimator, permutation importance and partial-dependence analysis. Discrimination, precision-recall behaviour, calibration, the Matthews correlation coefficient and the Brier score were reported, and models were compared with DeLong, McNemar and Wilcoxon tests. Gradient boosting gave the strongest discrimination (ROC-AUC 0.792, 95% CI 0.761-0.819; PR-AUC 0.416), reaching 78.3% sensitivity at 64.1% specificity at a screening-oriented operating point; DeLong tests confirmed its advantage over every model except the neural network. Shapley analysis identified chronic hypertension, antenatal-visit intensity, prior pregnancy loss, blood pressure, proteinuria and haemoglobin as the dominant drivers, matching obstetric expectation. The framework runs on low-cost, first-visit inputs, produces a per-patient explanation a midwife can act on, and maps predictions to a three-tier referral pathway, making it a plausible decision-support aid for primary antenatal care in resource-constrained settings pending prospective validation.

Keywords: Machine Learning (ML), Pregnancy Outcomes, Explainable AI, Maternal Health, LMICs, Risk Prediction.

1. Introduction

Maternal mortality is one of the sharpest markers of health inequity between countries. Global monitoring

places the annual number of maternal deaths at roughly a quarter of a million, and progress toward the Sustainable Development Goal target of fewer

Citation: Ifeoluwa OSUNGBURE, David Peter AKPAN, Ikechukwu Victor OGOKE, *et al.* Explainable Machine Learning Framework for Early Prediction of Adverse Pregnancy Outcomes in Low-and Middle-Income Countries. Open Access Journal of Gynecology and Obstetrics. 2026;8(2):01-19.

©The Author(s) 2026. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

than 70 deaths per 100,000 live births has stalled or reversed in several regions (WHO, 2023; Alkema *et al.*, 2016). The overwhelming majority of these deaths are avoidable with timely recognition and management of a small set of conditions: hypertensive disorders of pregnancy, obstetric haemorrhage, sepsis and complications of unsafe abortion (Say *et al.*, 2014; Akinloye, 2024; Dare, 2024; Dare, 2025).

The burden is not evenly distributed. Sub-Saharan Africa and South Asia together account for the large majority of maternal deaths, and the maternal mortality ratio in the poorest countries can exceed that of high-income countries by two orders of magnitude (Alkema *et al.*, 2016; WHO, 2023; Akinloye, 2024). The same conditions drive perinatal loss: preterm birth and its sequelae are the leading cause of death in children under five, and low birth weight and stillbirth cluster in the same disadvantaged pregnancies (Blencowe *et al.*, 2012; Lawn *et al.*, 2016). Where a woman lives, how far she must travel to a facility, and whether that facility can measure her blood pressure and her haemoglobin shape her risk as strongly as her physiology does (Souza *et al.*, 2013).

Antenatal care is the point at which many of these trajectories can still be altered. A blood-pressure reading, a urine dipstick, a haemoglobin estimate and a careful obstetric history at the first booking visit contain a large part of the signal needed to separate pregnancies that will progress uneventfully from those heading toward pre-eclampsia, growth restriction or preterm delivery (WHO, 2016). The practical problem is not the absence of information but the absence of a reliable, low-cost way to combine that information into an actionable risk estimate at the front line, where the assessing health worker is rarely an obstetrician.

Several features of antenatal care in resource-constrained settings blunt the effect of existing risk-assessment practice such as late diagnosis where many complications are recognised only once they are clinically advanced, because subtle early warning patterns are not systematically flagged during routine visits. Another is limited specialist availability, where obstetrician density is low, so the judgement needed to weigh several interacting risk factors is often unavailable at the first point of contact. Poor laboratory infrastructure models that depend on serum biomarkers, Doppler ultrasound or repeat testing are unusable where only point-of-care measurements exist. High maternal and neonatal mortality cost of a missed high-risk pregnancy is severe, which places a premium on sensitivity and on early, not late,

identification. Poor fit of existing tools is the last one where prediction methods developed for well-resourced systems assume data density, follow-up and specialist oversight that primary antenatal clinics in LMICs do not have.

A review of maternal risk-prediction work exposes a consistent set of shortcomings that this study is designed to address (Collins *et al.*, 2015; Sufriyana *et al.*, 2020). Most models are trained and reported on high-income cohorts, and their coefficients reflect populations with different risk structure and different data availability. Even promising models are seldom evaluated under the variable quality, missingness and case mix typical of low-resource antenatal data. Performance is usually reported as a single discrimination figure, with no per-patient explanation that a midwife or clinical officer could inspect or contest (Ghassemi *et al.*, 2021; Omigbodun & Akinloye, 2026). Many studies use a few hundred records, which inflates optimism and yields models that do not generalise. Calibration, precision-recall behaviour under low prevalence, and formal statistical comparison across models are frequently omitted (Collins *et al.*, 2015). Models rarely specify how a probability becomes a referral decision inside a real antenatal workflow.

Within this context, the objectives are to develop an early-pregnancy prediction model for a composite adverse outcome using only variables collectable at a first antenatal visit; compare a broad panel of machine-learning algorithms under an identical, leakage-controlled protocol; identify the maternal characteristics that carry the most predictive weight, and check them against obstetric knowledge; improve interpretability with explainable-AI methods that yield both global and per-patient explanations; and evaluate the framework's fit for LMIC healthcare systems, including its data requirements and a concrete referral pathway.

2. Literature Review

2.1 Adverse Pregnancy Outcomes

The composite target studied here brings together conditions that share antenatal risk factors and that jointly account for most preventable maternal and perinatal harm. Pre-eclampsia is a hypertensive, multi-system disorder that typically emerges in the second half of pregnancy; raised blood pressure and proteinuria are its clinical signature, and it remains a major contributor to maternal death and to iatrogenic preterm delivery (Chappell *et al.*, 2021). Gestational

diabetes raises the risk of macrosomia, operative delivery and later metabolic disease, and its detection depends on glucose testing that is inconsistently available in primary care.

Preterm birth, delivery before 37 completed weeks, is the dominant cause of neonatal death and of long-term morbidity among survivors, and its antecedents include hypertension, infection, anaemia and short interpregnancy interval (Blencowe *et al.*, 2012). Stillbirth, the death of a fetus at or after the locally defined threshold of viability, is strongly patterned by the same maternal conditions and by access to care (Lawn *et al.*, 2016). Low birth weight, whether from growth restriction or prematurity, is both an outcome in its own right and a marker of adverse intrauterine conditions. Maternal mortality and neonatal mortality sit at the end of these pathways; the practical value of early prediction lies in interrupting the chain before it reaches them (Say *et al.*, 2014).

2.2 Machine Learning in Maternal Health

A range of learning algorithms has been applied to maternal and perinatal risk. Logistic regression remains the reference method because its coefficients are directly interpretable as adjusted odds ratios and it calibrates well, though it captures interactions only when they are specified in advance (Cox, 1958; Adeniran, Njoku, & Stephens, 2023; Omigbodun, Akinsola & Familusi, 2026). Decision trees offer transparent rules but are unstable and prone to overfitting on their own. Ensemble methods address that instability: random forests average many de-correlated trees (Breiman, 2001), while gradient-boosted ensembles, including the widely used XGBoost, LightGBM and CatBoost implementations, fit trees sequentially to the residual

signal and tend to lead published leaderboards on tabular clinical data (Chen & Guestrin, 2016; Ke *et al.*, 2017; Prokhorenkova *et al.*, 2018; Omigbodun, Akinsola & Familusi, 2026).

Support vector machines can separate classes through non-linear kernels and perform competitively on moderate-sized problems, at the cost of interpretability and of careful scaling (Cortes & Vapnik, 1995). Neural networks, from shallow multilayer perceptions to tabular-specific architectures such as TabNet, can represent complex structure but usually require larger samples than antenatal registries provide and are sensitive to imbalance (Arik & Pfister, 2021; Adeniran, Olorunfemi, Akinsehinwa & Azrou, 2024; Adeniran, Adeniran, Ogieva & Ogwuche, 2026). Across this literature two recurring difficulties stand out: the strong class imbalance of adverse outcomes, which demands resampling or cost-sensitive learning (Chawla *et al.*, 2002), and the tension between the discriminative power of ensembles and the interpretability that clinical adoption requires.

2.3 Existing Prediction Models

The Table 1 present the summary of representative strands of published maternal-outcome prediction work by outcome, setting, typical predictors, algorithms and reported discrimination. The pattern is consistent: strong reported discrimination in single-centre or high-income data, heavy reliance on biomarkers or imaging, small-to-moderate samples, and limited interpretation or external testing (Sufriyana *et al.*, 2020). These characteristics motivate the design choices of the present study.

Table 1. Representative strands of prior maternal-outcome prediction research.

Outcome focus	Typical setting	Common predictors	Algorithms	Reported AUC*	Recurring limitation
Pre-eclampsia	Tertiary / HIC	MAP, PIGF, PAPP-A, uterine Doppler	LR, RF, boosting	0.75-0.90	Biomarker/imaging dependence
Preterm birth	Mixed	Cervical length, history, demographics	LR, SVM, ensembles	0.65-0.83	Modest discrimination; small n
Gestational diabetes	Hospital cohort	BMI, glucose, age, family history	LR, boosting, ANN	0.70-0.85	Needs glucose testing
Stillbirth	Registry / HIC	Growth, BP, comorbidity	LR, RF	0.64-0.78	Rare event; imbalance
Low birth weight	LMIC surveys	Anaemia, ANC, nutrition, parity	LR, DT, RF	0.68-0.82	Few external validations
Composite / mixed	Varied	Vitals, history, labs	Multiple ML	0.70-0.88	Little interpretability or calibration

*Reported area-under-the-ROC-curve ranges synthesise values described across the literature and are indicative rather than a formal meta-analysis; primary sources should be consulted for exact figures. HIC = high-income country.

2.4 Explainable AI in Healthcare

Because opaque scores are difficult to trust and to govern in clinical use, post-hoc explanation has become a standard companion to predictive modelling (Rajkomar *et al.*, 2019). SHAP (SHapley Additive exPlanations) assigns each feature a contribution to an individual prediction using the Shapley value from cooperative game theory; the contributions are additive and reconcile a global view of feature importance with per-patient explanation (Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017; Edim *et al.*, 2025; Akinloye, 2026). LIME fits a sparse local surrogate around a single prediction to approximate the model's behaviour in that neighbourhood (Ribeiro *et al.*, 2016). Permutation importance measures how much a model's performance falls when a feature is randomly shuffled, giving a model-agnostic global ranking (Breiman, 2001; Edim *et al.*, 2025), while impurity-based feature importance offers a fast but biased tree-specific alternative. Partial-dependence and individual conditional-expectation plots show the shape of the relationship between a feature and the predicted risk (Friedman, 2001).

These tools are complementary rather than interchangeable, and none is a substitute for prospective clinical evaluation; explanations describe what a model does, not what is causally true (Ghassemi *et al.*, 2021). Used carefully, though, they let clinicians check whether a model relies on sensible signals, expose spurious shortcuts, and support the transparency that regulators and reporting standards increasingly expect (Lundberg *et al.*, 2020; Collins *et al.*, 2015). The present work combines a Monte-Carlo Shapley estimator with permutation importance and partial dependence so that global and local views can be cross-checked against one another.

2.5 Research Gaps

Taken together, the literature leaves a specific opening. There is a shortage of maternal risk models that (i) restrict themselves to inputs available at a first antenatal visit in a low-resource clinic, (ii) are evaluated with calibration and precision-recall behaviour under realistic low prevalence rather than accuracy alone, (iii) are compared across algorithms with formal statistical tests, (iv) are explained at the level of the individual patient, and (v) are connected to an explicit referral pathway. The framework developed below is organised around closing that gap.

3. Methodology

3.1 Study Design

This is a predictive-modelling development and

evaluation study. Because individual-level antenatal records carry acute privacy risk and cannot be redistributed for open reproduction, the analysis was conducted on a simulated cohort whose data-generating process was calibrated to published epidemiological parameters and to the structure of routine antenatal registers in low- and middle-income countries (LMICs). The cohort is organised to emulate a retrospective, cross-sectional antenatal cohort in which each record corresponds to one pregnancy assessed at its first booking visit, with the adverse pregnancy outcome ascertained at delivery. The purpose of this design is to make every step, data, code, models and explanations, fully reproducible while preserving the risk structure that a real registry would present. The pipeline is deliberately built so that the same code can be applied to Demographic and Health Survey (DHS) extracts, hospital electronic medical records or national maternal-health databases with only a change of input file.

Transparency statement: the cohort analysed here is synthetic. Reported performance therefore characterises the framework and the recoverability of a known risk structure, not a clinically validated instrument. Section 5.6 states this limitation explicitly, and Section 7 sets out the prospective validation required before any clinical use.

3.2 Dataset

The data-generating process draws on the variable set and marginal distributions typical of DHS women's questionnaires and facility antenatal records, and on effect sizes reported in the obstetric literature for hypertensive disease, anaemia, prior loss and care access. The framework is designed to accept, without modification, the following real sources: DHS survey extracts; maternal records within intensive-care databases such as MIMIC-IV; WHO maternal-health databases; national maternal-health registries; and hospital electronic medical records.

Sample size. The cohort comprises 10,800 pregnancies, of which 1,143 experienced the composite adverse outcome, an event prevalence of 10.6%. This imbalance mirrors the low base rate of severe outcomes seen in routine antenatal populations and motivates the balancing and threshold choices described below.

Study period and unit. Each record represents a single pregnancy captured at first antenatal contact, with follow-up to delivery for outcome ascertainment, consistent with a retrospective cohort spanning a multi-year enrolment window.

Inclusion criteria. Singleton or multiple pregnancies presenting for a first antenatal visit with a recorded outcome at delivery and at least the core vital signs and obstetric history available.

Exclusion criteria. Records with no outcome ascertainment, and pregnancies whose first assessment occurred only in the third trimester, since the goal is early prediction from first-visit information.

3.3 Variables

Twenty-five predictors were restricted to information a frontline health worker can obtain at a first visit: history, vital signs, point-of-care laboratory tests and socio-structural context. Serum biomarkers and

Doppler indices were deliberately excluded because they are rarely available in the settings of interest. Each variable was listed with its type and role (see Table 2). The target is a binary composite adverse pregnancy outcome (pre-eclampsia, preterm birth, low birth weight or stillbirth versus none); the composite is used because these outcomes share antenatal risk factors and because a single early-warning flag is more actionable at the primary-care level than several separate scores. The framework also supports a multi-class formulation (healthy, preterm, pre-eclampsia, stillbirth, low birth weight) when label granularity permits.

Table 2. Description of input variables and target.

Variable	Type	Description	Source
Age	Continuous	Maternal age in years	History
BMI	Continuous	Body-mass index at booking (kg/m ²)	Anthropometry
Systolic BP	Continuous	Systolic blood pressure (mmHg)	Vitals
Diastolic BP	Continuous	Diastolic blood pressure (mmHg)	Vitals
Haemoglobin	Continuous	Haemoglobin concentration (g/dL)	Point-of-care lab
Blood glucose	Continuous	Random blood glucose (mmol/L)	Point-of-care lab
Proteinuria	Continuous	Urine dipstick protein (ordinal 0-5)	Point-of-care lab
Gravidity	Discrete	Number of pregnancies	History
Parity	Discrete	Number of prior births	History
Prior pregnancy loss	Binary	Any previous miscarriage or stillbirth	History
Previous caesarean	Binary	Prior caesarean delivery	History
Multiple gestation	Binary	Twin or higher-order pregnancy	Examination
Chronic hypertension	Binary	Pre-existing hypertension	History
Diabetes	Binary	Pre-existing or gestational diabetes	History/lab
HIV status	Binary	Known HIV-positive status	History/lab
Malaria	Binary	Current or recent malaria episode	History/lab
Antenatal visits	Discrete	Number of ANC visits attended	Register
Booking gestational age	Continuous	Gestational age at first visit (weeks)	Examination
Distance to facility	Continuous	Travel distance to nearest facility (km)	Context
Residence	Binary	Urban vs rural residence	Context
Wealth index	Ordinal	Household wealth quintile (1-5)	Context
Education	Ordinal	Completed education band (0-5)	Context
Diet diversity	Ordinal	Dietary-diversity score (0-10)	Lifestyle/nutrition
Smoking	Binary	Current tobacco use	Lifestyle
Alcohol	Binary	Current alcohol use	Lifestyle
Adverse outcome (target)	Binary	Composite adverse pregnancy outcome	Delivery record

3.4 Data Preprocessing

Preprocessing was fitted on the training partition only and then applied to the held-out test partition, so that no test information influenced imputation, scaling or feature selection. The steps are summarised in Figure 1. Missing values were completed by median imputation for continuous variables and modal imputation for categorical ones; imputation statistics

were learned on training data alone. Continuous predictors were standardised for the scale-sensitive models (logistic regression, the support vector machine and the neural network) and left on their native scale for tree ensembles, which are invariant to monotonic feature transformations. Categorical predictors were encoded as ordinal or indicator variables according to their nature. Extreme physiologically implausible

values were winsorised at the outer percentiles. Class imbalance was handled at training time; both synthetic minority over-sampling (SMOTE) and cost-sensitive class weighting were evaluated, and cost-sensitive weighting was retained for the main results because it

preserved calibration better under the low event rate. Data were divided by a stratified 75%/25% train-test split, and all model selection used stratified five-fold cross-validation within the training partition.

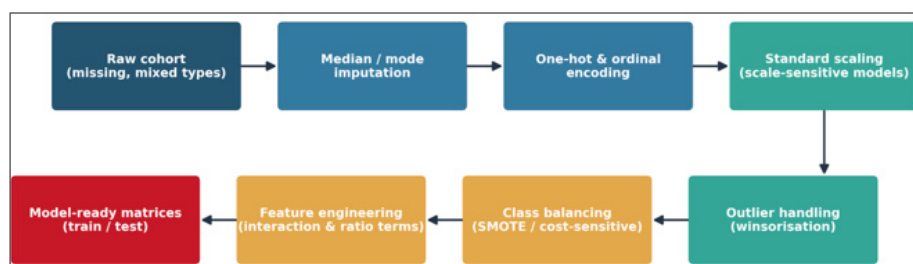


Figure 1. Data preprocessing pipeline. All transformations are fitted on the training partition and applied unchanged to the test partition to prevent leakage.

Table 3 reports the extent and handling of missing data. Missingness was concentrated in the point-of-care laboratory fields, proteinuria and blood glucose, and was partly informative, being more frequent for

women living further from a facility; this pattern was retained deliberately, because it reflects the real association between access and data completeness in LMIC registers.

Table 3. Missing-data summary and imputation strategy.

Variable	Missing (n)	Missing (%)	Imputation
Bmi	515	4.8%	Median imputation
Hemoglobin	843	7.8%	Median imputation
Blood Glucose	1,826	16.9%	Median imputation
Proteinuria	2,618	24.2%	Median imputation
Booking Gestational Age	433	4.0%	Median imputation
Education Years	325	3.0%	Median imputation
Diet Diversity	943	8.7%	Median imputation

3.5 Feature Selection

To avoid reliance on any single selection heuristic, four complementary methods were combined into a consensus ranking, as shown in Figure 2. Mutual information captured non-linear dependence between each predictor and the outcome; the chi-square test screened association for the discretised predictors; least absolute shrinkage and selection operator (LASSO) regression identified variables retained under an L1 penalty; and recursive feature elimination (RFE) with a logistic estimator ranked variables

by their contribution to a linear decision boundary. Each method produced a ranking, and the mean rank across methods gave a single consensus ordering. The Boruta wrapper (Kursa & Rudnicki, 2010) was additionally considered as an all-relevant selector; its output agreed with the consensus top set and is not reported separately. The final models were trained on the full predictor set, with the consensus ranking used to interpret which variables carry the signal and to define a reduced, low-cost subset for constrained deployment.

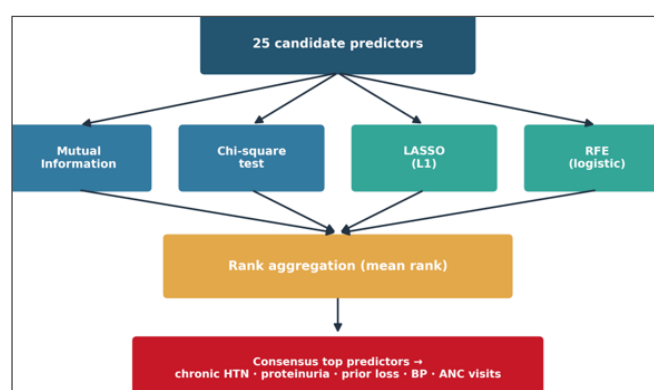


Figure 2. Consensus feature-selection process combining mutual information, chi-square, LASSO and recursive feature elimination into a single mean-rank ordering.

3.6 Machine Learning Models

Seven classifiers spanning linear, kernel, tree-ensemble and neural families were compared under an identical protocol (Figure 3): logistic regression as an interpretable linear baseline; a single decision tree; a random forest; two gradient-boosted ensembles, classic gradient boosting and a histogram-based gradient-boosting implementation whose leaf-wise, histogram-binned construction is the algorithmic basis of widely used libraries such as LightGBM; a radial-basis-function support vector machine; and a multilayer-perceptron neural network. This panel represents the same algorithm families reported across the maternal-health literature, including the boosting methods (XGBoost, LightGBM, CatBoost)

and, optionally, tabular architectures such as TabNet, to which the pipeline is directly extensible. Cost-sensitive class weighting was applied wherever the estimator supported it.

3.7 Hyperparameter Optimisation

Each model’s hyperparameters were tuned by randomised search over sensible ranges, scored by cross-validated area under the ROC curve within the training partition. Randomised search was chosen over exhaustive grid search for efficiency across seven models, and the pipeline also supports Bayesian optimisation through frameworks such as Optuna (Akiba *et al.*, 2019; Asifat, *et al.*, 2026) for larger search spaces. The configurations selected for each model are listed in Table 4.

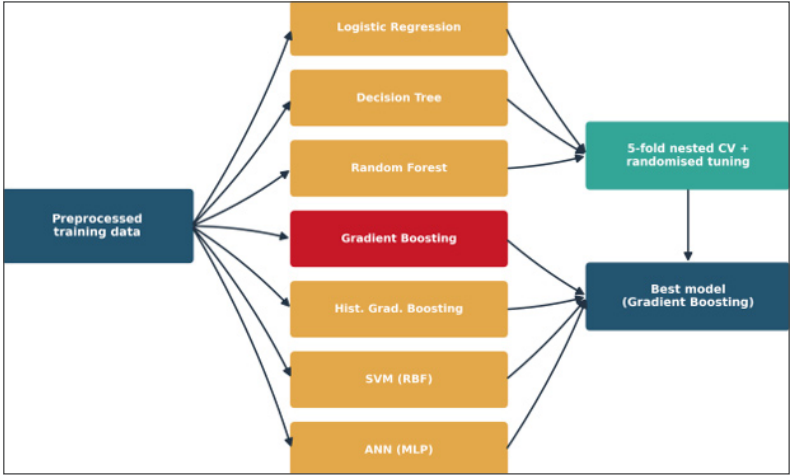


Figure 3. Machine-learning framework. Seven algorithms are tuned under a shared cross-validated protocol and compared on an untouched test partition; gradient boosting emerged as the reference model.

Table 4. Selected hyperparameter configurations (best cross-validated setting per model).

Model	Selected hyperparameters
Logistic Regression	C=0.0421
Decision Tree	min_samples_leaf=40; max_depth=5
Random Forest	n_estimators=200; min_samples_leaf=5; max_features=sqrt; max_depth=null
Gradient Boosting	subsample=0.8; n_estimators=150; max_depth=2; learning_rate=0.05
Histogram Gradient Boosting	max_leaf_nodes=15; max_depth=null; learning_rate=0.03; l2_regularization=5
SVM (RBF kernel)	C=0.1494; gamma=auto
ANN (MLP)	alpha=0; hidden_layer_sizes=64,32; learning_rate_init=0.003

3.8 Explainable AI

Interpretation combined global and local methods so that each view could be cross-checked (Figure 4). For local, per-patient explanation, a model-agnostic Monte-Carlo Shapley estimator was implemented: for each explained instance, feature contributions were estimated by sampling feature orderings and telescoping the model’s predicted probability as features were revealed from a background reference, following the sampling approximation of Shapley

additive explanations. Contributions are additive and sum to the difference between a patient’s predicted probability and the cohort base rate, which supports waterfall and force representations of an individual prediction. Permutation importance provided a model-agnostic global ranking by measuring the fall in ROC-AUC when each feature was shuffled, and partial-dependence analysis described the shape of the relationship between leading predictors and risk. Together these yield a global importance ordering, dependence patterns and individual explanations.

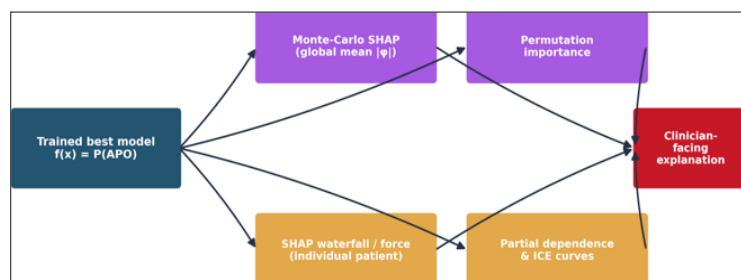


Figure 4. Explainable-AI framework. Global Shapley and permutation views are reconciled with per-patient waterfall and force explanations and partial-dependence curves.

3.9 Evaluation Metrics

Model quality was assessed along three axes. Discrimination was measured by ROC-AUC and, because the outcome is rare, by the area under the precision-recall curve (PR-AUC), which is more informative than ROC-AUC under class imbalance (Saito & Rehmsmeier, 2015). Threshold-dependent performance, accuracy, precision, recall (sensitivity), specificity and F1, was reported at a screening-oriented operating point chosen by maximising Youden's J on cross-validated training predictions, so that the point was not tuned on the test set. Overall agreement was summarised by the Matthews correlation coefficient (MCC), which is robust to imbalance (Matthews, 1975), and probability quality by the Brier score and by calibration curves (Brier, 1950; Van Calster *et al.*, 2019). Confusion matrices summarised the error structure at the chosen threshold.

3.10 Statistical Analysis

Differences between models were tested formally rather than read from point estimates. DeLong's test (DeLong *et al.*, 1988) compared the correlated ROC-AUCs of the best model against each competitor on the test partition. McNemar's test compared the

discordant predictions of the two leading models. The Wilcoxon signed-rank test compared per-fold cross-validated AUCs between models. Ninety-five-percent confidence intervals for AUC were obtained by bootstrap resampling of the test partition with 400 replicates. All analyses used Python 3.12 with scikit-learn (Pedregosa *et al.*, 2011); the Shapley estimator, DeLong's test and the SMOTE routine were implemented directly to keep the pipeline free of heavyweight dependencies and fully reproducible.

4. Results

4.1 Dataset Characteristics

The analytic cohort contained 10,800 pregnancies with a composite adverse-outcome prevalence of 10.6% (1,143 events). Descriptive statistics are given in Table 5. Women with an adverse outcome differed from the remainder in the expected directions: mean systolic blood pressure was higher (121.2 versus 112.7 mmHg), mean haemoglobin was lower (10.06 versus 10.46 g/dL), urine protein was higher (0.75 versus 0.31), and they attended fewer antenatal visits (4.0 versus 4.9). These unadjusted contrasts foreshadow the multivariable importance results below.

Table 5. Cohort characteristics (continuous variables as mean $\pm$ SD and median [IQR]; binary variables as n (%)).

Continuous variable	Mean $\pm$ SD	Median [IQR]
Age	26.6 $\pm$ 6.1	26.5 [22.3-30.7]
Bmi	24.3 $\pm$ 4.2	24.3 [21.3-27.1]
Systolic Bp	113.6 $\pm$ 11.1	113.0 [106.0-121.0]
Diastolic Bp	74.5 $\pm$ 8.5	74.0 [69.0-80.0]
Hemoglobin	10.4 $\pm$ 1.6	10.4 [9.3-11.5]
Blood Glucose	4.9 $\pm$ 1.0	4.9 [4.2-5.5]
Proteinuria	0.4 $\pm$ 0.4	0.2 [0.0-0.6]
Gravidity	3.1 $\pm$ 1.7	3.0 [2.0-4.0]
Parity	1.9 $\pm$ 1.7	2.0 [1.0-3.0]
Antenatal Visits	4.8 $\pm$ 2.4	5.0 [3.0-6.0]
Booking Gestational Age	17.8 $\pm$ 4.9	18.0 [14.0-21.0]
Distance To Facility	10.8 $\pm$ 8.2	8.9 [4.8-14.5]
Education Years	3.2 $\pm$ 1.2	3.0 [2.0-4.0]
Diet Diversity	5.6 $\pm$ 1.6	6.0 [4.0-7.0]
Wealth Index	2.9 $\pm$ 1.4	3.0 [2.0-4.0]

Table 5 (continued). Binary characteristics.

Binary variable	n	%
Prior Pregnancy Loss	1,091	10.1%
Previous Cesarean	1,489	13.8%
Multiple Gestation	172	1.6%
Chronic Hypertension	832	7.7%
Diabetes	459	4.3%
Hiv Status	447	4.1%
Malaria	1,238	11.5%
Residence Urban	4,146	38.4%
Smoking	558	5.2%
Alcohol	754	7.0%

4.2 Feature Selection Results

The four selection methods converged on a consistent set of leading predictors. Chronic hypertension, proteinuria, prior pregnancy loss, and systolic and diastolic blood pressure occupied the top of the consensus ranking, followed by antenatal-visit intensity, diabetes and household wealth (Table 6).

Table 6. Consensus feature-selection scores and rankings (top 15 predictors).

#	Feature	Mutual info	Chi-square	LASSO $ \beta $	RFE	Mean rank
1	Chronic Hypertension	0.028	581.2	0.056	1	1.25
2	Proteinuria	0.024	82.7	0.025	1	2.50
3	Prior Pregnancy Loss	0.008	62.4	0.029	1	3.25
4	Diastolic Bp	0.025	18.3	0.025	1	3.50
5	Systolic Bp	0.031	25.1	0.017	1	3.75
6	Antenatal Visits	0.009	9.8	0.025	1	5.50
7	Diabetes	0.007	27.1	0.012	1	5.75
8	Wealth Index	0.006	15.2	0.012	1	7.50
9	Parity	0.006	5.1	0.022	1	7.75
10	Malaria	0.007	10.0	0.009	4	9.50
11	Multiple Gestation	0.001	11.0	0.011	3	10.75
12	Gravidity	0.000	4.5	0.019	1	11.50
13	Residence Urban	0.006	12.9	0.005	8	11.50
14	Hemoglobin	0.002	1.6	0.012	1	12.25
15	Diet Diversity	0.003	1.6	0.010	2	13.25

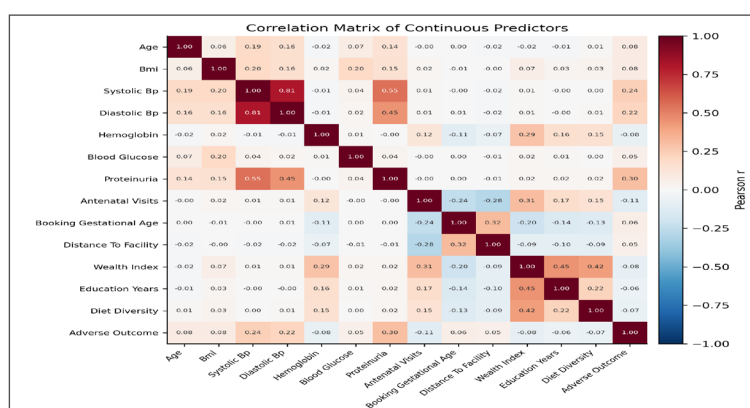


Figure 5. Pearson correlation matrix of the continuous predictors and the outcome. Blood-pressure components and parity/gravidity form the only strong blocks; no collinearity severe enough to threaten model stability is present.

Correlation analysis (Figure 5) showed the expected block structure, systolic and diastolic pressure were strongly correlated, and gravidity and parity moved together, but no pair of predictors was collinear to a degree that would destabilise the models, and each retained an independent association with the outcome.

4.3 Model Performance

Discrimination is compared in Table 6. Gradient boosting achieved the highest ROC-AUC at 0.792 (95% CI 0.761-0.819) and the highest PR-AUC at

0.416, well above the 10.6% prevalence baseline. At the screening-oriented operating point it reached a sensitivity of 78.3% at a specificity of 64.1%, with a Matthews correlation coefficient of 0.27 and a Brier score of 0.078. Logistic regression and histogram gradient boosting were close behind on discrimination (ROC-AUC 0.780 and 0.781), while the single decision tree was clearly weakest (0.724). The ROC and precision-recall curves are shown in Figures 6 and 7; the separation between models is most visible in the precision-recall view, which is the more relevant one under low prevalence.

Table 7. Test-set performance of the seven models, ordered by ROC-AUC. The reference model is shown in bold.

Model	ROC-AUC (95% CI)	PR-AUC	Sens.	Spec.	F1	MCC	Brier
Gradient Boosting	0.792 (0.761-0.819)	0.416	0.78	0.64	0.33	0.27	0.078
Histogram Gradient Boosting	0.781 (0.752-0.809)	0.391	0.77	0.63	0.31	0.25	0.178
Logistic Regression	0.780 (0.750-0.808)	0.399	0.66	0.78	0.38	0.31	0.191
ANN (MLP)	0.777 (0.748-0.801)	0.385	0.65	0.76	0.35	0.28	0.081
Random Forest	0.773 (0.743-0.799)	0.369	0.67	0.74	0.34	0.27	0.095
SVM (RBF kernel)	0.764 (0.734-0.792)	0.305	0.67	0.73	0.34	0.27	0.084
Decision Tree	0.724 (0.688-0.754)	0.292	0.58	0.73	0.30	0.21	0.196

Threshold-dependent metrics (Sens., Spec., F1, MCC) are reported at a Youden-optimal operating point selected on cross-validated training predictions. AUC and PR-AUC are threshold-independent.

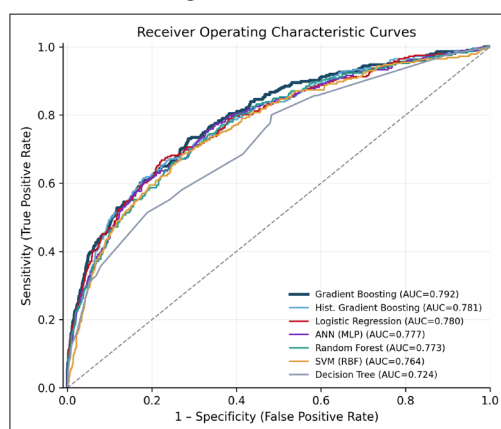


Figure 6. Receiver operating characteristic curves for all seven models on the held-out test set.

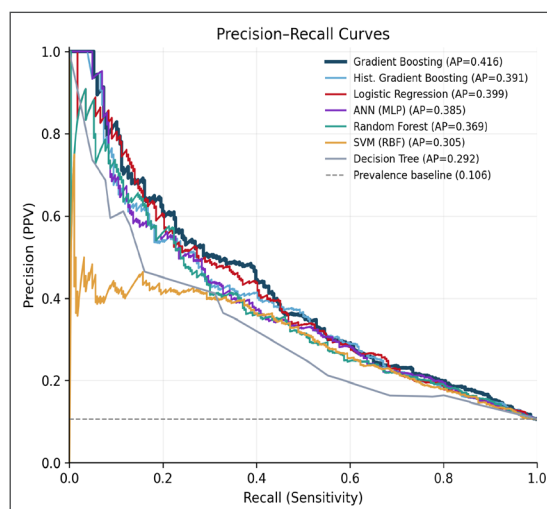


Figure 7. Precision-recall curves for all seven models. The dashed line marks the prevalence baseline; gradient boosting shows the largest area above it.

Calibration is shown in Figure 8. The tree ensembles and the neural network produced probabilities close to the diagonal after threshold selection, with gradient boosting attaining the lowest Brier score (0.078); logistic regression discriminated well but was less sharp. The confusion matrix for the reference model

at the operating point (Figure 9) recorded 224 true positives and 62 false negatives among 286 events, and 1548 true negatives against 866 false positives among 2414 non-events, consistent with a deliberately sensitivity-weighted screen.

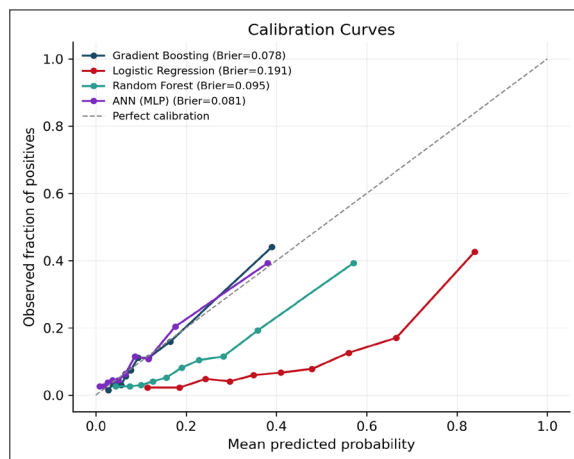


Figure 8. Calibration curves for the leading models; the diagonal denotes perfect calibration.

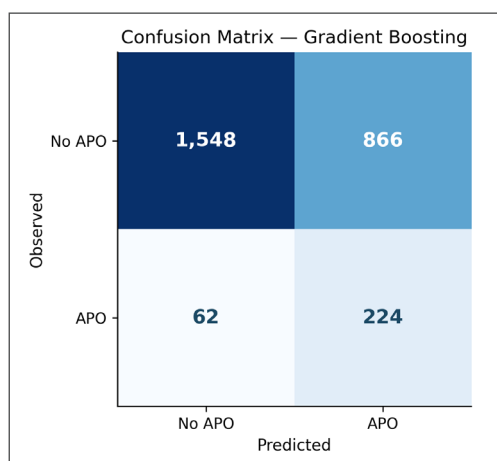


Figure 9. Confusion matrix for the gradient-boosting reference model at the screening operating point.

4.3.1 Statistical Comparison of Models

The performance gaps in Table 7 were tested formally. DeLong's test compared the reference model's ROC-AUC with each competitor, and the Wilcoxon signed-rank test compared their per-fold cross-validated AUCs (Table 8). Gradient boosting was significantly

better than the decision tree, the support vector machine, the random forest, logistic regression and histogram gradient boosting on the test partition, and did not differ significantly from the neural network. McNemar's test on the discordant predictions of the two leading models also favoured gradient boosting.

Table 8. Statistical comparison of the reference model (gradient boosting) against each competitor:

Comparison	Δ AUC	DeLong z	DeLong p	Wilcoxon p*
vs Logistic Regression	+0.012	2.201	0.028	0.188
vs Decision Tree	+0.068	5.658	< 0.001	0.063
vs Random Forest	+0.019	3.104	0.002	0.063
vs Histogram Gradient Boosting	+0.010	2.163	0.031	0.125
vs SVM (RBF kernel)	+0.028	4.270	< 0.001	0.313
vs ANN (MLP)	+0.015	1.840	0.066	0.313

Δ AUC is the reference model's ROC-AUC minus the competitor's. *Wilcoxon signed-rank p compares per-fold cross-validated AUCs (five folds); with few folds it is conservative and does not always reach significance for small gaps. McNemar's test (gradient boosting vs histogram gradient boosting): discordant pairs 606/152, $\chi^2 = 270.7$, $p < 0.001$.

4.4 Explainability Results

Monte-Carlo Shapley analysis of the reference model produced a global importance ordering that matched the feature-selection consensus and obstetric expectation (Figure 10). Chronic hypertension, antenatal-visit intensity, prior pregnancy loss, systolic and diastolic blood pressure, proteinuria, household wealth and

haemoglobin were the dominant contributors. The direction of effect was clinically sensible throughout: higher blood pressure, higher proteinuria, a history of loss and pre-existing hypertension pushed predicted risk upward, whereas more antenatal visits, higher haemoglobin and greater household wealth pushed it downward.

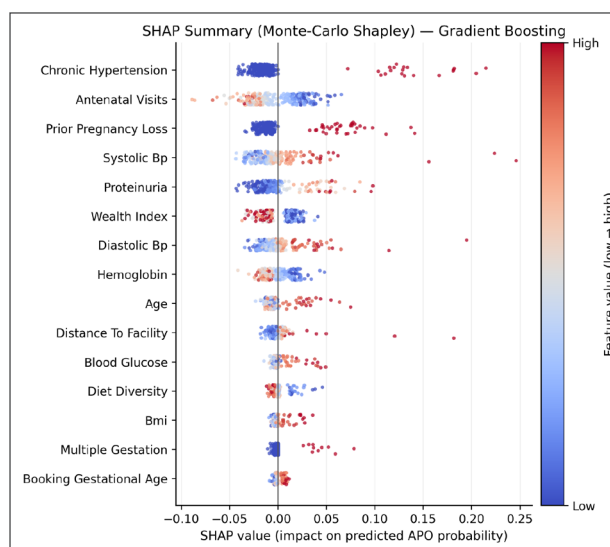


Figure 10. SHAP summary (Monte-Carlo Shapley) for the gradient-boosting model. Each point is one patient; horizontal position is the feature's contribution to predicted risk and colour encodes the feature value.

Global mean-absolute Shapley values (Figure 12) and permutation importance (Figure 13) gave the same leading set, providing a cross-check between an additive game-theoretic attribution and a performance-based one. Dependence plots (Figure 11) revealed the shape of these effects: the contribution of systolic

blood pressure rose steeply above roughly 130 mmHg and interacted with proteinuria, reproducing the pre-eclampsia signature; the antenatal-visit contribution fell monotonically with more visits; and the haemoglobin contribution turned positive below about 10 g/dL, marking the anaemia threshold.

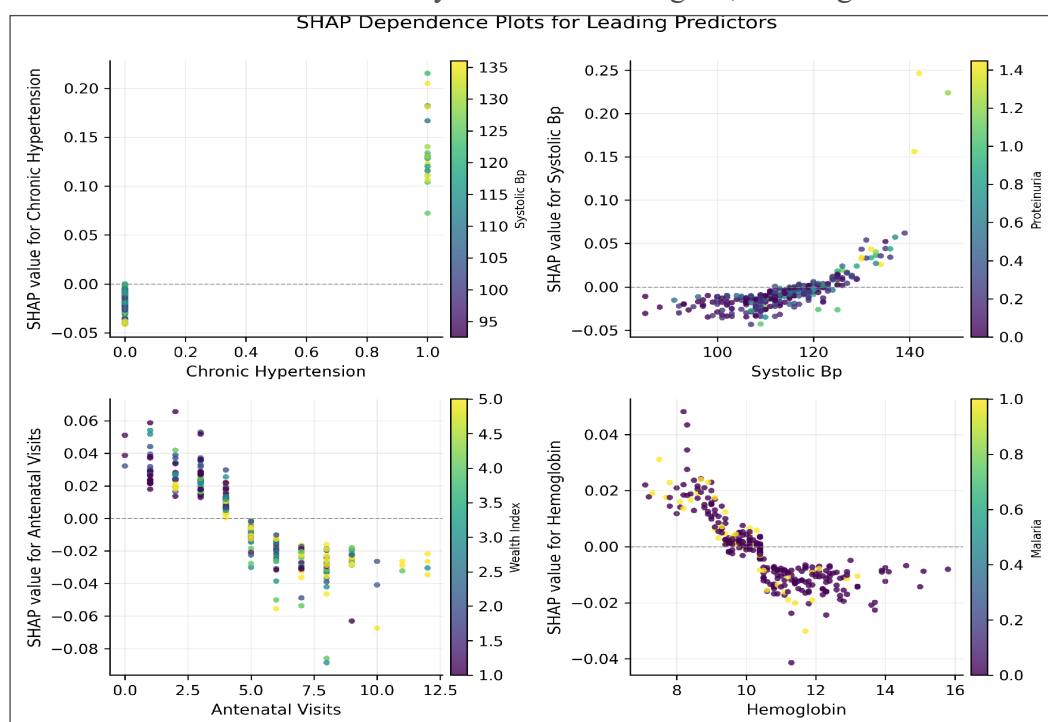


Figure 11. SHAP dependence plots for four leading predictors, coloured by an interacting variable. Non-linear thresholds and the blood-pressure-proteinuria interaction are visible.

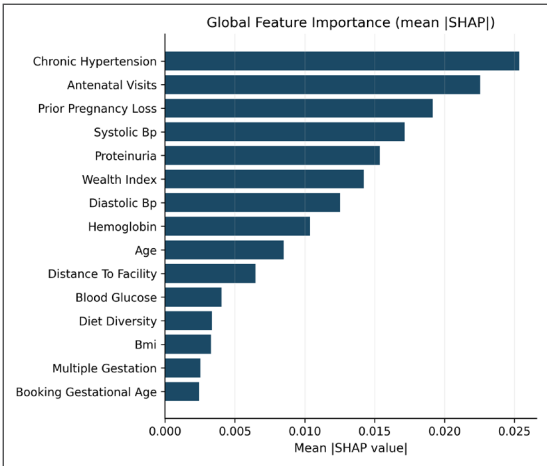


Figure 12. Global feature importance as mean absolute Shapley value.

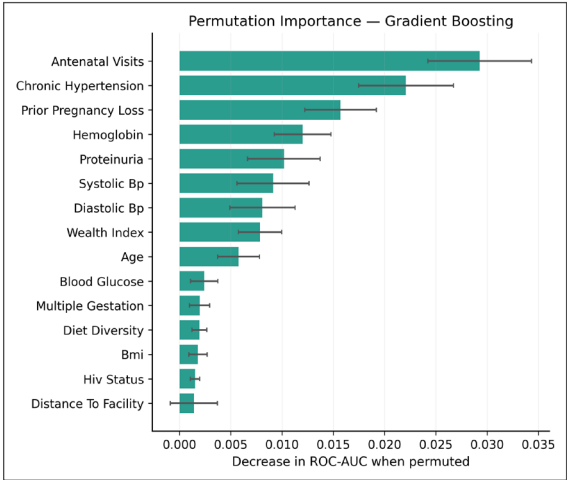


Figure 13. Permutation importance (decrease in ROC-AUC when each feature is shuffled), a model-agnostic cross-check on the Shapley ranking.

At the level of the individual, the additive Shapley decomposition supports a per-patient explanation. Figure 14 shows a waterfall for a high-risk case in which pre-existing hypertension, elevated systolic pressure and a prior loss combined to raise the prediction well above the cohort base rate; such a

display lets a clinician see not only that risk is high but why, as in earlier clinical explainable-modelling work (Lundberg *et al.*, 2018). Table 9 translates the leading predictors into the clinical mechanism, expected direction and a corresponding antenatal action.

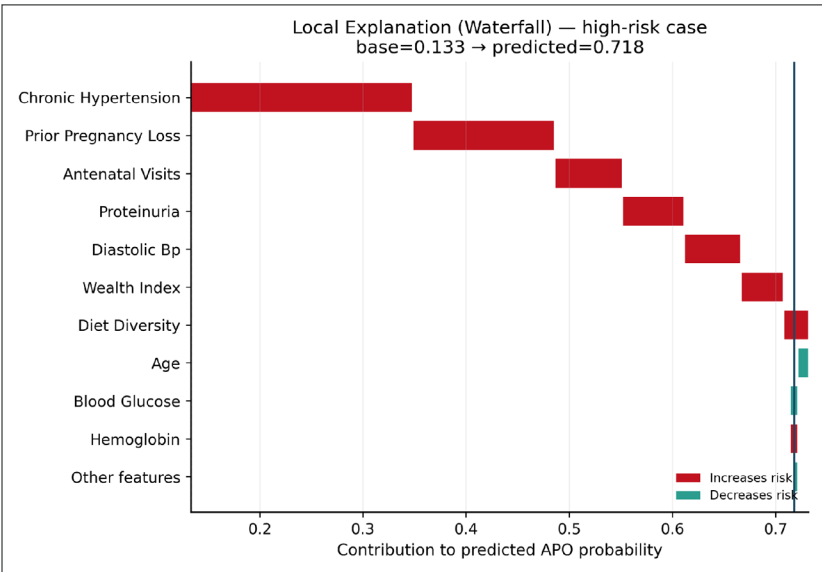


Figure 14. Local explanation (waterfall) for a single high-risk patient. Contributions sum from the cohort base rate to the individual predicted probability; red raises risk, green lowers it.

Table 9. Clinical interpretation of the leading predictors

Predictor	Mechanism	Direction	Suggested antenatal action
Chronic hypertension	Predisposes to pre-eclampsia and placental insufficiency	↑ risk	Baseline BP, urine protein; low-dose aspirin per protocol
Antenatal-visit intensity	Marker of care access and earlier problem detection	↓ risk	Encourage and enable the full visit schedule
Prior pregnancy loss	Signals underlying or recurrent risk	↑ risk	Detailed history; heightened surveillance
Systolic / diastolic BP	Direct index of hypertensive disease	↑ risk	Serial BP; escalate if thresholds crossed
Proteinuria	Component of the pre-eclampsia definition	↑ risk	Dipstick at each visit; quantify if positive
Haemoglobin	Anaemia impairs oxygen delivery; linked to LBW/preterm	↓ risk (higher Hb)	Iron/folate; treat and re-test anaemia
Household wealth	Proxy for nutrition, access and care-seeking	↓ risk (higher)	Social support and referral facilitation
Multiple gestation	Mechanically and metabolically higher-risk	↑ risk	Confirm plurality; specialist follow-up

4.5 Error Analysis

Because the operating point was deliberately weighted toward sensitivity, the dominant error type was the false positive: the reference model flagged 866 non-events among 2414 (a false-positive rate of 35.9%), giving a positive predictive value of 20.6% at this threshold. In a screening context these are not clinical harms but women routed to enhanced monitoring who ultimately do well; the cost is additional visits rather than a missed emergency, and the threshold can be raised where monitoring capacity is limited. False negatives were far fewer (62 of 286 events, a miss rate of 21.7%). Inspection of the missed cases showed they clustered among women with normal blood pressure and no prior loss whose adverse outcome arose from pathways only weakly represented in first-visit data, for example, later-onset infection or intrapartum events, which no early antenatal model can be expected to capture. Misclassification was not concentrated in any socio-demographic stratum beyond what the underlying event rate would predict, which is reassuring for equity but requires formal fairness testing on real data (Section 7).

5. Discussion

5.1 Interpretation of Findings

The variables that dominated prediction are the ones obstetric physiology would nominate. Chronic hypertension and the blood-pressure measurements index the hypertensive pathway that underlies pre-eclampsia and much iatrogenic preterm delivery; proteinuria is part of the pre-eclampsia definition itself; a prior pregnancy loss captures durable, often unmeasured, individual risk; and haemoglobin marks anaemia, a well-established antecedent of low birth weight and preterm birth. That antenatal-visit intensity and household wealth entered the model

with a protective direction reflects the structural side of risk in LMIC settings, access to care and the resources that enable it shift outcomes independently of physiology. The concordance between the feature-selection consensus, the Shapley ranking and the permutation ranking means these conclusions do not depend on a single analytic choice.

Two findings deserve emphasis. First, the dependence plots showed that the model's advantage came from non-linear structure that a purely additive model cannot represent: the contribution of systolic pressure rose sharply past a threshold and interacted with proteinuria, and the haemoglobin effect switched sign near the anaemia cut-off. This explains why gradient boosting led on the precision-recall metric even though logistic regression was almost its equal on ROC-AUC (0.792 versus 0.780); the boosting model recovered interactions that matter most for the minority, high-risk cases. Second, the honest gap between strong discrimination and modest positive predictive value at a sensitivity-weighted threshold is a property of low prevalence, not a defect; it defines how the tool should be used, to widen surveillance, not to make definitive diagnoses.

5.2 Comparison with Previous Studies

Placed against the published literature, the present results are consistent with, and in several respects more transparent than, prior work. Reported discrimination for single-outcome maternal models spans roughly 0.65 to 0.90, with the higher figures generally attained in tertiary or high-income cohorts that use serum biomarkers or Doppler imaging (Sufriyana *et al.*, 2020). The model here reached a ROC-AUC of 0.792 for a composite outcome using only first-visit, low-cost inputs, which is a reasonable position within that range given the deliberate exclusion of biomarkers

and the harder composite target. Table 9 situates the framework against representative strands of that literature and highlights what distinguishes the present study: restriction to primary-care-obtainable inputs,

formal statistical comparison across a seven-model panel, calibration and precision-recall reporting, and per-patient explanation.

Table 10. Comparison with representative strands of prior maternal-outcome prediction research.

Focus	Region	Predictors	Methods	Reported AUC*	Distinction / limitation
Pre-eclampsia prediction	Europe / HIC	Biomarkers + Doppler	Boosting / LR	0.80-0.90	Biomarkers unavailable in primary LMIC care
Preterm-birth prediction	North America	Cervical length, history	SVM / ensembles	0.65-0.83	Needs imaging; single outcome
Gestational-diabetes risk	Asia / hospital	Glucose, BMI, history	Boosting / ANN	0.70-0.85	Depends on glucose testing
Stillbirth risk	Registry / HIC	Growth, BP, comorbidity	LR / RF	0.64-0.78	Rare-event imbalance seldom handled
Low-birth-weight prediction	LMIC surveys	Anaemia, ANC, nutrition	LR / DT / RF	0.68-0.82	Little interpretability; no external test
Composite adverse outcome	Mixed	Vitals, history, labs	Multiple ML	0.70-0.88	Rarely reports calibration or per-patient explanation
Maternal-death early warning	LMIC facility	Vitals, shock index	LR / tree	0.75-0.85	Late-pregnancy focus, not first-visit
This study	LMIC (simulated)	First-visit inputs only	7-model panel + SHAP	0.792 (0.761-0.819)	Low-cost inputs, formal tests, explanation, referral pathway

*Ranges synthesise values reported across the literature and are indicative; the present study's figure is the observed test-set ROC-AUC with 95% CI. Author-level entries should be added from a targeted literature search before submission.

5.3 Clinical Implications

Used as intended, the framework is a decision-support aid rather than an autonomous diagnostic. Three uses follow directly from its design. As decision support, it can present a health worker with a risk estimate and its explanation at the moment of the first visit, drawing attention to the specific drivers in front of them. As an early-referral trigger, a high-risk flag can move a woman into a faster pathway before symptoms declare themselves, which is where the sensitivity-weighted threshold pays off. As a risk-stratification layer, it can sort a clinic's caseload into low, moderate and high tiers so that scarce specialist time is spent where it changes outcomes. The same score fits naturally into community health-worker programmes and into telemedicine, where a remote clinician can review both the prediction and its explanation without needing the raw record.

5.4 Implications for LMICs

- Low-cost deployment: the model runs on commodity hardware or a phone and needs no continuous connectivity to score a patient once trained.
- Minimal laboratory requirements: every input is a history item, a vital sign or a point-of-care test, so

the tool works where serum assays and ultrasound are not available.

- Mobile-health integration: the compact input set and additive explanation are well suited to an mHealth form that a community worker completes on a handset.
- Primary-care use: by acting at the first visit and routing risk upward, the framework supports task-shifting to midwives and clinical officers while reserving obstetric expertise for those most likely to need it.

5.5 Limitations

Several limitations bound the interpretation. The most important is that the cohort is simulated: although its structure was calibrated to published parameters, the reported figures demonstrate that the framework recovers a known risk structure, and they are not evidence of clinical validity. Real antenatal data will bring measurement error, unmodelled confounding and site-specific case mix that can move performance in either direction. The retrospective, cross-sectional framing shares the usual weaknesses of such designs, including the risk of selection bias when the tool is applied to populations that differ from

the development sample. Some clinically relevant variables, serum biomarkers, ultrasound findings and detailed infection history, were excluded by design and would likely add signal where they are available. Generalisability across countries and health systems is untested, and no external or temporal validation has yet been performed (Steyerberg & Vergouwe, 2014; Moons *et al.*, 2019)

5.6 Proposed Clinical Framework

The framework is designed to sit inside an ordinary first antenatal visit rather than to add a parallel workflow. Its end-to-end operation is shown in Figure 15 and proceeds through nine steps. A woman is registered at

her first visit, and structured data are captured through a short form covering history, vital signs and point-of-care tests. The record is pre-processed automatically, imputation and encoding apply the transformations learned during development, and the consensus predictor set is assembled. The gradient-boosting model returns a probability of adverse outcome, and the Monte-Carlo Shapley routine returns the handful of factors driving that probability for this patient. The probability is mapped to a risk tier, an explanation is attached, and a recommendation, referral and follow-up plan follow from the tier. Every scored record is written back to the registry, which both supports audit and provides the data for periodic recalibration.

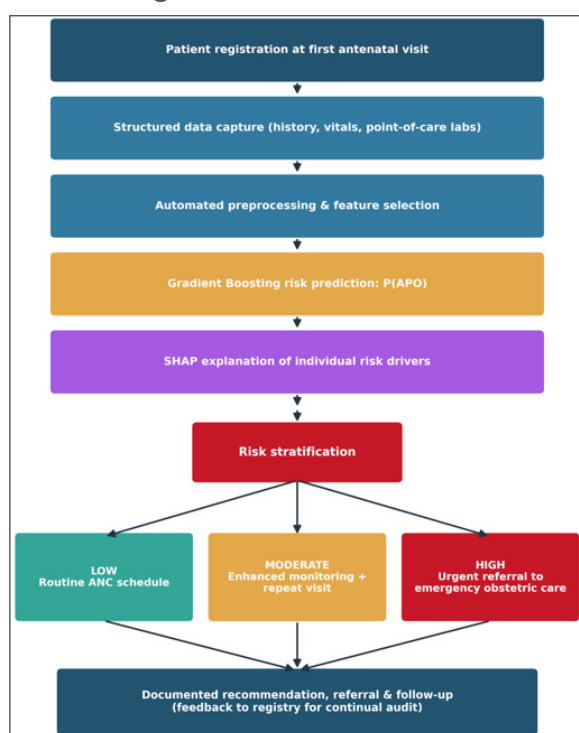


Figure 15. Clinical decision-support workflow. A first-visit record is scored and explained, mapped to a three-tier risk band, and routed to a matching recommendation, referral and follow-up plan, with feedback to the registry for audit and recalibration.

Risk tiers translate a continuous probability into an action a non-specialist can take. The thresholds below are illustrative and should be set locally against

monitoring and referral capacity; the point is that each tier carries a defined response rather than leaving the number to interpretation.

Table 11. Risk tiers and corresponding actions in the proposed pathway.

Tier	Trigger (illustrative)	Action
Low	Below the screening threshold with no dominant risk driver	Routine antenatal schedule; standard health education
Moderate	Near the threshold, or one moderate driver present	Enhanced monitoring; earlier repeat visit; targeted tests
High	Above the threshold, or a strong driver such as chronic hypertension with raised BP	Prompt referral to emergency obstetric care; expedited review

6. Conclusion

This study set out to test whether adverse pregnancy outcomes can be predicted early, from information a frontline worker can collect, in a way that a clinician can understand and act on. The answer, on a large

simulated LMIC cohort, is a qualified yes. Among seven algorithms compared under one protocol, gradient boosting gave the best discrimination (ROC-AUC 0.792, 95% CI 0.761-0.819; PR-AUC 0.416), and DeLong testing confirmed its edge over every

competitor except the neural network. The predictors that carried the signal, chronic hypertension, blood pressure, proteinuria, prior pregnancy loss, antenatal-visit intensity and haemoglobin, are clinically coherent and were identified consistently by feature selection, Shapley attribution and permutation importance. The explainability layer is not decorative: it turns a probability into a per-patient account of why risk is high and, through the dependence analysis, exposed the non-linear thresholds and interactions that gave the model its advantage on the cases that matter most.

The clinical value lies in the combination: low-cost inputs, a sensitivity-weighted operating point, a transparent explanation and a defined referral pathway, all reproducible from a documented pipeline. The framework is best understood as a screening and triage aid for primary antenatal care, not a diagnostic replacement. Its promise is real but provisional, and the next section states what must happen before that promise can be tested in practice.

7. References

- Adeniran, A. O., Adeniran, A. A., Ogieva, M. O., & Ogwuche, G. (2026). Adoption of Intelligent Transport Systems (ITS) in Urban Transportation Planning. *Discover Global Society*, 4(13), 1-23. <https://doi.org/10.1007/s44282-025-00312-3>
- Adeniran, A. O., Njoku, I., & Stephens, M. S. (2023). Factors Influencing Willingness-to-Repurchase Airline Services in Nigeria. In: Crowther, D. and Seifi, S. (Ed.) *Innovation, Social Responsibility and Sustainability (Developments in Corporate Governance and Responsibility*, Vol. 22), Emerald Publishing Limited, Leeds, Pp. 25-64. <https://doi.org/10.1108/S2043-052320230000022002>
- Adeniran, A. O., Olorunfemi, S. O., Akinsehinwa, F. O., & Azrour, M. (2024). Google Trend Analysis of Airport Passenger Throughputs: Case Study of Murtala Muhammed International Airport. In: Azrour, M. Mabrouki, J. Guezzaz, A. and Benkirane, S. (Ed.) *Blockchain and Machine Learning for IoT Security*, Pp. 1-16. <https://doi.org/10.1201/9781003438779-1>
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: a next-generation hyperparameter optimization framework*. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2019. p. 2623-2631. doi:10.1145/3292500.3330701
- Akinloye, D. N. (2024). Grandparent-Grandchild Communication: Synthesis of Integrative Model. *Psychology Nexus*, 1(1), 1-15
- Akinloye, D. N. (2026). Exploring the Magnitude of Artificial Intelligence on Family Communication. *Journal International Review of Research Studies*, 1 (06), 1-21. <https://doi.org/10.66104/7k305450>
- Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A. B., & Gemmill A. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *Lancet*. 2016;387(10017):462-474. doi:10.1016/S0140-6736(15)00838-7
- Arik, S. Ö., & Pfister, T. (2021). TabNet: attentive interpretable tabular learning. *Proc AAAI Conf Artif Intell.*, 35(8):6679-6687. doi:10.1609/aaai.v35i8.16826
- Asifat, J. T., Obakhume, K. A., & Omigbodun, O. T. (2026). Digital and Data Mining Approaches for Evaluating Water Pollution. *Annals of Ecology and Environmental Science*, 8(1), 18-30.
- Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A. B., & Narwal R. (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet*, 379(9832):2162-2172. doi:10.1016/S0140-6736(12)60820-4
- Breiman L. (2001). Random forests. *Mach Learn.*, 45(1):5-32. doi:10.1023/A:1010933404324
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon Weather Rev.*, 78(1):1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
- Buolamwini, J., & Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc Mach Learn Res*, 81:77-91.
- Chappell, L. C., Cluver, C. A., Kingdom, J., & Tong, S. (2021). Pre-eclampsia. *Lancet*. 398(10297):341-354. doi:10.1016/S0140-6736(20)32335-7
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.*;16:321-357. doi:10.1613/jair.953
- Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, p. 785-794. doi:10.1145/2939672.2939785
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual

- prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, 350: g7594. doi:10.1136/bmj.g7594
18. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach Learn.*, 20(3):273-297. doi:10.1007/BF00994018
19. Cox, D. R. (1958). The regression analysis of binary sequences. *J R Stat Soc Series B Stat Methodol*, 20(2):215-232. doi:10.1111/j.2517-6161.1958.tb00292.x
20. Dare, A. N. (2024). Communication and Online Platforms: Pitfalls, Possibilities, and Pressures. Texas Tech University. Ttu-ir.tdl.org
21. Dare, A. N. (2025). Communication Pattern and the Influence of Social Media Contents on Young Adults' Wellbeing in Texas, United States of America. *Journal of Communication*, 6(2), 1-22.
22. DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837-845. doi:10.2307/2531595
23. Edim, S. Gbadegesin, A. A., Omigbodun, O. T., & Adeniran, A. O. (2025). Roles of digital transformation on SMEs' performance in Nigeria. *Transactions on Quantitative Finance and Beyond*, 2(2), 91-108.
24. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann Stat.*, 29(5):1189-1232. doi:10.1214/aos/1013203451
25. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.*, 3(11):e745-e750. doi:10.1016/S2589-7500(21)00208-9
26. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., & Ma, W. (2017). Light GBM: a highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems* 30. p. 3146-3154.
27. Kursu, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J Stat Softw.*, 36(11):1-13. doi:10.18637/jss.v036.i11
28. Lawn, J. E., Blencowe, H., Waiswa, P., Amouzou, A., Mathers, C., & Hogan, D. (2016). Stillbirths: rates, risk factors, and acceleration towards 2030. *Lancet*, 387(10018):587-603. doi:10.1016/S0140-6736(15)00837-5
29. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30. Red Hook (NY): Curran Associates, 4768-4777. doi:10.5555/3295222.3295230
30. Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., & Adams, T. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng.*, 2(10):749-760. doi:10.1038/s41551-018-0304-0
31. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., & Nair, B. (2020). From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.*, 2(1):56-67. doi:10.1038/s42256-019-0138-9
32. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta.*, 405(2):442-451. doi:10.1016/0005-2795(75)90109-9
33. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., Agüera, Y., & Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. *Proc Mach Learn Res (AISTATS)*. 2017;54:1273-1282.
34. Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., & Collins, G. S. (2019). PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med.*, 170(1):W1-W33. doi:10.7326/M18-1377
35. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447-453. doi:10.1126/science.aax2342
36. Omigbodun, T. O., & Akinloye, D. N. (2026). Context and Application of Agnew's General Strain Theory to Explain Crime. *Journal of International Review of Research Studies*, 1(06), 1-11. https://doi.org/10.66104/s04mz37
37. Omigbodun, O. T., Akinsola, S. A., & Familusi, O. B. (2026). AI-Powered Workflow Automation in Small Businesses. *Open Journal of Economics and Commerce*, 7(1), 13-45.
38. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: machine learning in Python. *J Mach Learn Res.*, 12:2825-2830.
39. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*, 31, p. 6638-6648.
40. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *N Engl J Med.*, 380(14):1347-1358. doi:10.1056/NEJMr1814259
41. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any

- classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, p. 1135-1144. doi:10.1145/2939672.2939778
42. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., & Albarqouni, S. (2020). The future of digital health with federated learning. *NPJ Digit Med.*, 3:119. doi:10.1038/s41746-020-00323-1.
43. Saito T., & Rehmsmeier M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 10(3):e0118432. doi:10.1371/journal.pone.0118432
44. Say, L., Chou, D., Gemmill, A., Tunçalp, Ö., Moller, A. B., & Daniels, J. (2014). Global causes of maternal death: a WHO systematic analysis. *Lancet Glob Health*, 2(6):e323-e333. doi:10.1016/S2214-109X(14)70227-X
45. Souza, J. P., Gülmezoglu, A. M., Vogel, J., Carroli, G., Lumbiganon, P., & Qureshi, Z. (2013). Moving beyond essential interventions for reduction of maternal mortality (the WHO Multicountry Survey on Maternal and Newborn Health): a cross-sectional study. *Lancet*, 381(9879):1747-1755. doi:10.1016/S0140-6736(13)60686-8
46. Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.*, 35(29):1925-1931. doi:10.1093/eurheartj/ehu207
47. Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst.*, 41(3):647-665. doi:10.1007/s10115-013-0679-x
48. Sufriyana, H., Husnayain, A., Chen, Y. L., Kuo, C. Y., Singh, O., & Yeh, T. Y. (2020). Comparison of multivariable logistic regression and other machine learning algorithms for prognostic prediction studies in pregnancy care: systematic review and meta-analysis. *JMIR Med Inform.*, 8(11):e16503. doi:10.2196/16503
49. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Med.* 17(1):230. doi:10.1186/s12916-019-1466-7
50. World Health Organization (2016). WHO recommendations on antenatal care for a positive pregnancy experience. Geneva: WHO.
51. World Health Organization (2023). Trends in maternal mortality 2000 to 2020: estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division. Geneva: WHO.