American Journal of Artificial Intelligence
2025, Vol. 9, No. 2, pp. 217-228
https://doi.org/10.11648/j.ajai.20250902.22

Research Article

Leveraging Ensemble Models for Optimizing Predictive Accuracy of Low Birthweight Risk in Kenya

Victor Opiyo*, Emma Anyika

Department of Computer Science and Information Technology, Cooperative University of Kenya, Nairobi, Kenya

Abstract

Low birth weight (LBW) is a prevalent public health challenge in low- and middle-income countries, including Kenya, where approximately 11.5% of newborns are affected. LBW is linked to heightened infant mortality, infections, and long-term developmental issues. While machine learning (ML), particularly ensemble learning, has demonstrated potential in improving LBW risk prediction, its application in resource-limited settings like Kenya remains underexplored. Prior research has largely focused on developed countries with limited adoption in sub-Saharan Africa, highlighting a crucial gap this study aims to address. This research develops and evaluates ensemble machine learning models to predict LBW risk using nationally representative data from the 2022 Kenya Demographic and Health Survey. The study integrates traditional clinical indicators with advanced computational methods, employing base classifiers such as Support Vector Machines and Logistic Regression alongside ensemble methods including Random Forest, Gradient Boosting, and Extreme Gradient Boosting. Meta-ensemble approaches such as bagging, voting, and stacking were also assessed. Data preprocessing included treatment of missing values, encoding categorical variables, and addressing class imbalance through the Synthetic Minority Over-sampling Technique (SMOTE). Models were trained and validated using stratified cross-validation and independent testing, with evaluation metrics comprising ROC AUC, accuracy, F1-score, Matthews Correlation Coefficient, and Brier score, emphasizing both discrimination and calibration. Results indicate that Random Forest outperformed other models, achieving a high ROC AUC of 0.957 and PR AUC of 0.971, with excellent calibration (Brier score 0.089), evidencing its strong predictive capability for LBW risk in the Kenyan context. Important predictors identified were gestational age, maternal height and weight, antenatal care utilization, and socioeconomic factors, consistent with known biological and contextual determinants. Ethical considerations regarding patient privacy, algorithmic fairness, and transparency were incorporated to promote responsible AI use in healthcare. The findings demonstrate that tailored ensemble learning models provide robust, interpretable, and practical tools for LBW prediction in low-resource settings. This work fills a critical research gap by applying advanced ML methods to Kenyan maternal-child health data, offering potential to enhance clinical decision-making and improve maternal and neonatal outcomes. The study underscores the importance of contextualized AI solutions and ethical governance for sustainable healthcare innovation.

Keywords

Low Birth Weight, Ensemble Learning, Machine Learning, Predictive Modelling, Kenya

*Corresponding author: opiyovictor25@gmail.com (Victor Opiyo)

Received: 14 September 2025; **Accepted:** 24 September 2025; **Published:** 27 October 2025

1. Introduction

A newborn's health is a significant determinant that sets the stage for a human being's overall health and quality of life expectations. Low birth weight (LBW), defined as a birth weight of less than 2,500 grams, is a severe public health concern in Kenya and other developing nations. It is not just a figure but also indicates the suffering that families endure and the health complications to which newborns are subjected from the start of life. This is the reason why the health and well-being of a baby must be monitored while developing in the womb and after birth. One of the things that must be monitored before the baby is delivered is its weight, as it carries great significance in the prediction of future health and survival. It is therefore advisable to establish whether the baby will be of normal or low birth weight so that interventions can be initiated well before birth.

Preterm delivery and fetal growth restriction in the womb are the main causes of low birth weight. Preterm/premature birth occurs in less than 37 weeks (259 days) following the onset of the last menstrual period of a pregnant woman [1]. Maternal factors that lead to preterm birth include maternal malnourishment (anaemia, underweight, or overweight) before and during pregnancy, maternal illnesses (high blood pressure, diabetes, or infection), and maternal characteristics (low or high maternal age, multiple parity, or poor birth spacing) that play a major role in causing LBW. Additional risk factors for LBW include smoking, consumption of alcohol, and medically unwarranted caesarean sections. The impact of LBW extends beyond the neonatal period, with increased infant mortality rates and long-term health issues that may torment subjects for their entire life [2].

Global statistics show that 14.6% to 20% of babies were born with low birth weight, which was more than 20.5 million. The percentage rates of low birth weight also differed by regions, with developed countries having the lowest rate of 7.2%, Africa had 13.0%, while Asia had the highest rate of 17.3%. The Southern Asian region in particular, had the highest low birth weight rate of 26.4% [3]. The study further reports 11.5% of infants being born with LBW in Kenya, a nation that is still struggling with socio-economic inequalities, and hence the need to propose interventions for the condition [4].

Despite the gravity of such a condition, the potential for early prediction and intervention remains underdeveloped in low-resource settings like Kenya, where hospitals rely on a series of traditional methods to assess and estimate birth weight in clinical practice. The traditional methods, which include obstetric ultrasounds, symphysis-fundal height measurements, and abdominal palpation [5] are prone to reliability and accuracy problems. For example, despite the fact that obstetric ultrasound is regarded as the most reliable option for assessing fetal growth, it is not always available in low-income countries and poor communities, since they are expensive to purchase, maintain, and repair, making them unaffordable. Furthermore, the utilization of untrained ultrasound sonographers can cause inaccurate fetal weight estimations, necessitating training [6].

In light of the limitations associated with the current methods of LBW estimation used in developing countries, it is prudent to seek more effective means. More focus and effort should be put on effective ways of predicting additional birth weight and LBW since an early diagnosis can prove beneficial in elucidating appropriate obstetric interventions. In LBW prediction, several techniques can be used, but recent studies have pointed out that the use of Machine Learning (ML) algorithms is an efficient way.

Sophisticated ML algorithms coupled with big data analytics have been successful in reducing LBW in high-income countries by learning the determinants that can lead to specific outcomes from the data sets. For instance, a study by [7] established that ensemble models such as Random Forest and XGBoost possess the capability of predicting the risk factors of LBW with accuracy levels of more than 85% based on data from the Ethiopian population. However, such LBW prediction methods have never been applied and tested in developing countries like Kenya.

Such techniques are superior to the previous methods because they ensure more accurate and precise results [2]. The most widely used techniques for combining models are Bagging, Boosting, and Stacking, which require the use of more than one base model to make predictions. This integration of various models increases overall performance as the results achieved are far more accurate. A compound model can be constructed using data that is extremely readily available, e.g., symphysis-fundal height, maternal weight gain, and straightforward demography data [8].

This kind of data can be utilized to train machine learning algorithms like Random Forests, Gradient Boosting Machines, or Neural Networks. By applying various models, a more precise forecast can be made that leverages the capabilities of every individual base model. In addition, any potential prejudices and mistakes that can arise from dependence on a single traditional method alone may be reduced [9].

The creation of Low Birth Weight (LBW) prediction systems does not focus only on the technology employed also contends with the efficiency with which healthcare is implemented, especially in emerging economies. Some popular methods, such as Obstetric ultrasound machines for fetal growth and weight estimation, have depended on expensive equipment or infrastructure, which has worked to limit their adoption outside of particular specialist areas. Most recently, however, the focus has been on easy-to-use and affordable solutions, such as smartphone applications with a user-friendly interface that is integrated with EMR systems [8]. The systems are developed to estimate birth weight based on clinically available data and are therefore meant to assist health workers at all levels. This class of systems can easily

recognize high-risk pregnancies requiring immediate medical attention, thereby facilitating effective clinical intervention and the use of scarce resources. In the case of Kenya, where neonatal health outcomes are a massive problem, it would be better if ensemble machine learning models are applied and increasingly developed to capture the population structure as well as the existing healthcare infrastructure and system. These updates would make predictive activities better attuned to the needs of mothers and newborns, thereby improving neonatal care, health systems efficiency, and overall sustainability.

Several antecedents of low birth weight in newborns have been explored by various scholars in Kenya. Underweight babies have, however, been difficult to predict from the batch during the delivery process [10]. We thus hypothesize that with the maternal risk factors found in the literature, a machine learning model would be ideal in predicting low birth weight. This could be of greatest utility in enhancing the outcome of infant and maternal health.

This is an attempt to bridge the gap between research in the new area of machine learning and the practical need to control limited-resource environments. This article will be a resource for valuable information for medical professionals as it presents the highlights of the maternal and socio-economic factors dealt with. We hope that the information to be drawn from these paradigms will, in some way, help towards improving the health of children and mothers. The introduction plays an important role in providing background information (including relevant references), emphasizing the importance of the study, and outlining its objectives. It is crucial to conduct a thorough review of the current state of the research field and incorporate key publications into your work. By referencing other research papers, you can provide context and position your own work within the broader research landscape. The final paragraph should provide a concise summary of the main findings and conclusions, which will be helpful to the readers.

2. Materials and Methods

2.1. Data Source and Study Objective

This study utilized data from the nationally representative Kenya Demographic and Health Survey (KDHS) 2022 with 1,559 records. The primary objective was to develop and evaluate ensemble machine learning models for predicting the risk of low birthweight among newborns in Kenya, with the aim of optimizing predictive accuracy for this critical public health outcome.

2.2. Data Preprocessing

The raw dataset underwent comprehensive preprocessing to ensure data quality and suitability for analytical modeling. Variables were classified as numeric or categorical to facilitate type-specific handling. Missing numeric values were imputed using an iterative method that models each feature as a function of other features, while missing categorical data were imputed with the mode (most frequent category). All numeric features were standardized to a mean of zero and a standard deviation of one, and categorical variables were converted into a binary format using one-hot encoding, with all steps integrated into a unified pipeline to ensure consistent application.

2.3. Data Splitting and Class Imbalance Handling

The processed dataset was partitioned into a training set (80%) and a hold-out test set (20%) using stratified sampling to maintain the proportional distribution of the outcome variable. To address the inherent class imbalance in the low birthweight outcome, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training data to generate synthetic samples for the minority class and improve model sensitivity.

2.4. Predictive Modeling Framework

A diverse set of algorithms was selected, including base models (Support Vector Machine, Logistic Regression, Decision Tree, and K-Nearest Neighbors) with adjusted class weights, and advanced ensemble techniques (Random Forest, Gradient Boosting, Adaptive Boosting, and Extreme Gradient Boosting). Furthermore, three meta-ensemble strategies were implemented: Bagging to reduce variance, Voting Classifiers to aggregate predictions, and a Stacking classifier that used a logistic regression meta-learner on base model outputs. Feature selection was performed using a model-based approach with a Random Forest estimator to retain a subset of features with importance scores above the median threshold.

2.5. Model Training and Evaluation

All models were trained within an integrated pipeline, and hyperparameters were optimized using stratified 5-fold cross-validation on the training set with ROC AUC as the primary metric. Models were evaluated on the independent test set using a comprehensive suite of metrics: Accuracy, Balanced Accuracy, Precision, Recall, Specificity, F1 Score, Matthews Correlation Coefficient, ROC AUC, and Brier score, with Positive and Negative Likelihood Ratios calculated for clinical utility. Performance was evaluated at both the conventional 0.5 threshold and an optimized threshold (0.37) for improved clinical balance, with performance further interrogated using confusion matrices, ROC curves, Precision-Recall curves, and calibration plots.

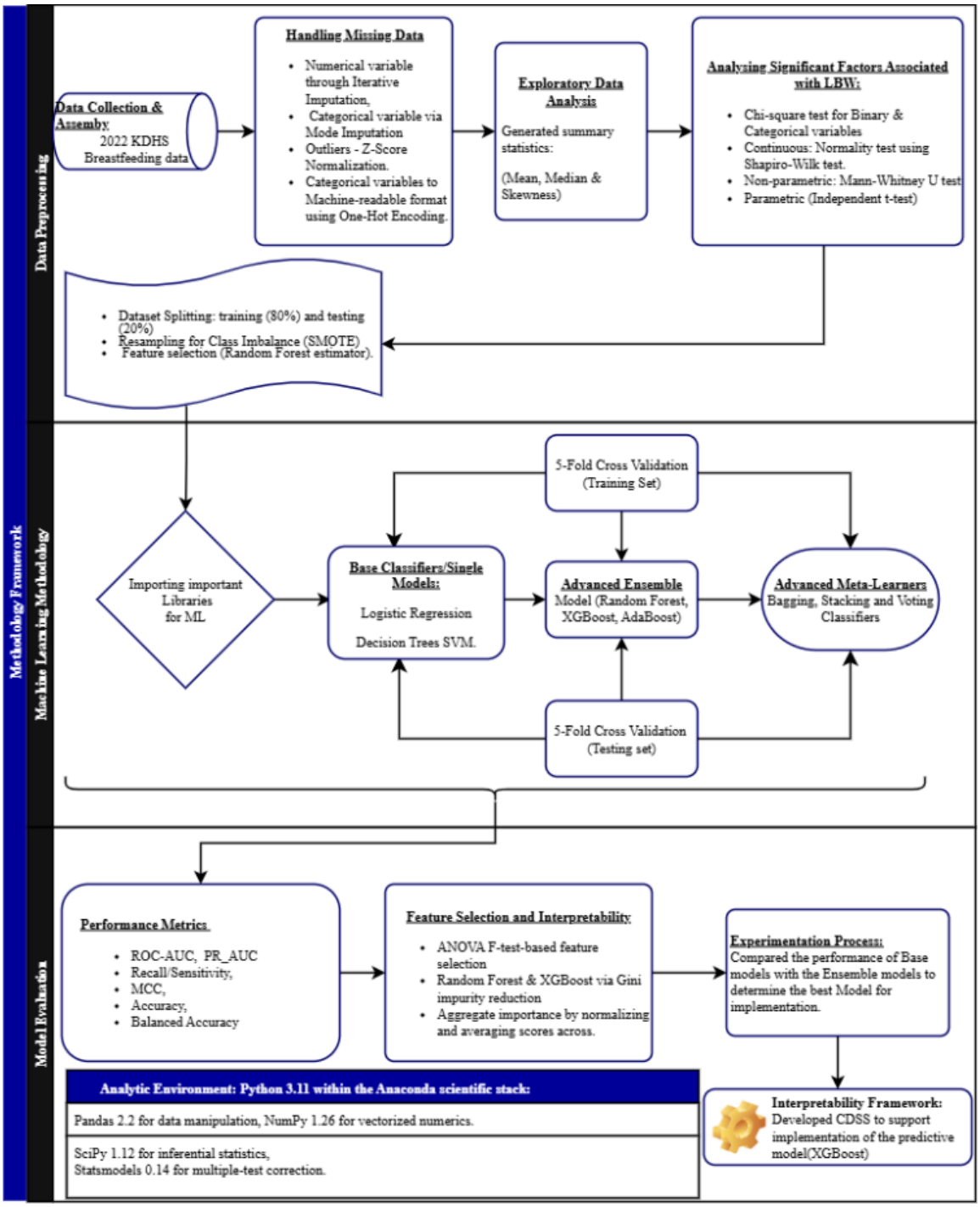

Figure 1. *Methodology Framework.*

2.6. Equations

1). ROC-AUC (Area Under the Receiver Operating Characteristic Curve)

$$\text{AUC} = \int_0^1 TPR(FPR)d(FPR)$$

Where:

$$TPR\ (True\ Positive\ Rate) = \frac{True\ positive(TP)}{True\ Positive(TP)+False\ Negative(FN)}$$

$$FPR\ (True\ Positive\ Rate) = \frac{False\ positive(FP)}{False\ Positive(FP)+True\ Negative(TN)}$$

2). Precision (Positive Predictive Value)

$$Precision = \frac{\text{True Positives (TP)}}{True\ Positives(TP) + False\ Positives(FP)}$$

3). Recall (Sensitivity, True Positive Rate)

$$Recall = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives (FN)}}$$

Interpretation: The proportion of actual LBW cases correctly identified by the model.

4). Specificity (True Negative Rate)

$$Specificity = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)+False Positive (FP)}}$$

Interpretation: The proportion of normal birth weight cases correctly identified.

5). F1-Score (Harmonic Mean of Precision and Recall)

$$F1 - Score = 2.\frac{Precision.Recall}{Precision + Recall}$$

6). Accuracy

$$Accuracy = \frac{True\ positives(TP)+True\ Negatives(TN)}{Total\ population}$$

Where:

TP = True Positives (correctly predicted as LBW)

TN = True Negatives (correctly predicted as non-LBW)

7). Matthews Correlation Coefficient (MCC)

$$\text{MCC} = \frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

2.7. Implementation

All analyses were performed using Python programming language (version 3.9) with robust libraries including scikit-learn for modeling, imbalanced-learn for handling class imbalance, XGBoost for gradient boosting, and matplotlib/seaborn for visualization. Random seed settings ensured reproducibility, while warnings were suppressed for clarity.

3. Results

The analytic cohort drawn from KDHS 2022 included maternal anthropometrics, sociodemographic characteristics, antenatal care utilization, pregnancy history, and access-to-care variables. Numeric predictors encompassed maternal weight and height, maternal and partner age, pregnancy index, birth order, preceding birth interval, pregnancy months, ANC visits, iron tablet days, malaria prophylaxis frequency, pregnancy losses, and timing of first ANC. Categorical variables represented residence, education, smoking, access barriers (permission, money, distance), labor market participation, occupation, wealth index, decision-making on expenditures, earnings type, place of delivery, receipt of iron tablets and dewormer, possession of a health card, pregnancy intention, caesarean delivery, and birthweight recall. Feature preprocessing adhered to the outlined methodology with numeric–categorical separation, appropriate encodings and imputations, and class-imbalance handling; models were trained and evaluated using consistent splits and probability thresholds matched to the programmatic objective of identifying low birthweight (LBW) risk.

The performance analysis reveals a significant advantage of ensemble methods over base classifiers, underscoring the complexity of the prediction task.

Base Models: Among the foundational algorithms, the Decision Tree achieved the highest ROC AUC (0.854) and overall accuracy (0.856) on the test set. Its strong performance, coupled with a high F1 score (0.869), suggests it effectively captures non-linear relationships within the data without being overly sensitive to the class imbalance, a trait not shared by Logistic Regression or K-Nearest Neighbors. The Support Vector Machine (SVM) exhibited the strongest cross-validation performance (CV ROC AUC Mean: 0.874), indicating high potential stability, though this did not fully translate to the highest test set performance.

Ensemble Models: A substantial performance leap was observed with tree-based ensemble methods. The Random Forest classifier emerged as the top-performing model, achieving a superior ROC AUC (0.957), excellent balanced accuracy (0.878), and the lowest Brier Score (0.089), indicating outstanding discriminatory power and the most calibrated probability estimates. XGBoost was a close competitor (ROC AUC: 0.937), also demonstrating robust performance. The high Positive Likelihood Ratios (PLR ~5.9 and ~5.6 for Random Forest and XGBoost, respectively) suggest that a positive prediction from these models would substantially increase the probability of a true LBW case, highlighting their diagnostic potential.

Meta-Ensemble Models: The Bagging classifier (based on Decision Trees) performed remarkably well, nearly matching the Random Forest (ROC AUC: 0.947). In contrast, the Voting and Stacking classifiers showed good but comparatively lower performance. This suggests that for this specific dataset, the homogeneous ensemble approach of bagging and boosting (Random Forest, XGBoost) was more effective than the heterogeneous combination of diverse base models in the meta-ensembles.

Table 1. *Comparative Model Performance.*

Base Model Performance Comparison (Sorted by ROC AUC):

Metric	Decision Tree	SVM	Logistic Regression	K-NN
Accuracy	0.856	0.753	0.744	0.74
Balanced Accuracy	0.846	0.756	0.747	0.747
Precision	0.906	0.822	0.816	0.715
Recall/Sensitivity	0.857	0.687	0.66	0.768
F1 Score	0.879	0.748	0.75	0.74
MCC	0.709	0.549	0.493	0.479
ROC AUC	0.843	0.819	0.814	0.806
Brier Score	0.061	0.108	0.115	0.125
CV ROC AUC Mean	0.843	0.819	0.814	0.806
CV ROC AUC Std	0.027	0.015	0.015	0.022

Ensemble Model Performance Comparison (Sorted by ROC AUC):

Metric	Random Forest	XGBoost	Gradient Boosting	AdaBoost
Accuracy	0.881	0.875	0.801	0.763
Balanced Accuracy	0.878	0.872	0.803	0.778
Precision	0.876	0.87	0.825	0.783
Recall/Sensitivity	0.911	0.904	0.815	0.777
F1 Score	0.893	0.887	0.815	0.777
MCC	0.761	0.748	0.603	0.525
ROC AUC	0.949	0.947	0.916	0.864
Brier Score	0.049	0.049	0.108	0.137
CV ROC AUC Mean	0.949	0.947	0.916	0.864
CV ROC AUC Std	0.012	0.018	0.016	0.01

Meta-Ensemble Model Performance Comparison (Sorted by ROC AUC):

Metric	Bagging (DT)	Stacking	Voting (Soft)	Voting (Hard)
Accuracy	0.872	0.865	0.837	0.779
Balanced Accuracy	0.874	0.868	0.835	0.801
Precision	0.874	0.88	0.885	0.858
Recall/Sensitivity	0.877	0.859	0.773	0.758
F1 Score	0.875	0.869	0.825	0.778
MCC	0.724	0.701	0.651	0.546
ROC AUC	0.943	0.938	0.908	nan
Brier Score	0.053	0.054	0.092	nan
CV ROC AUC Mean	0.943	0.938	0.908	nan
CV ROC AUC Std	0.019	0.026	0.032	nan

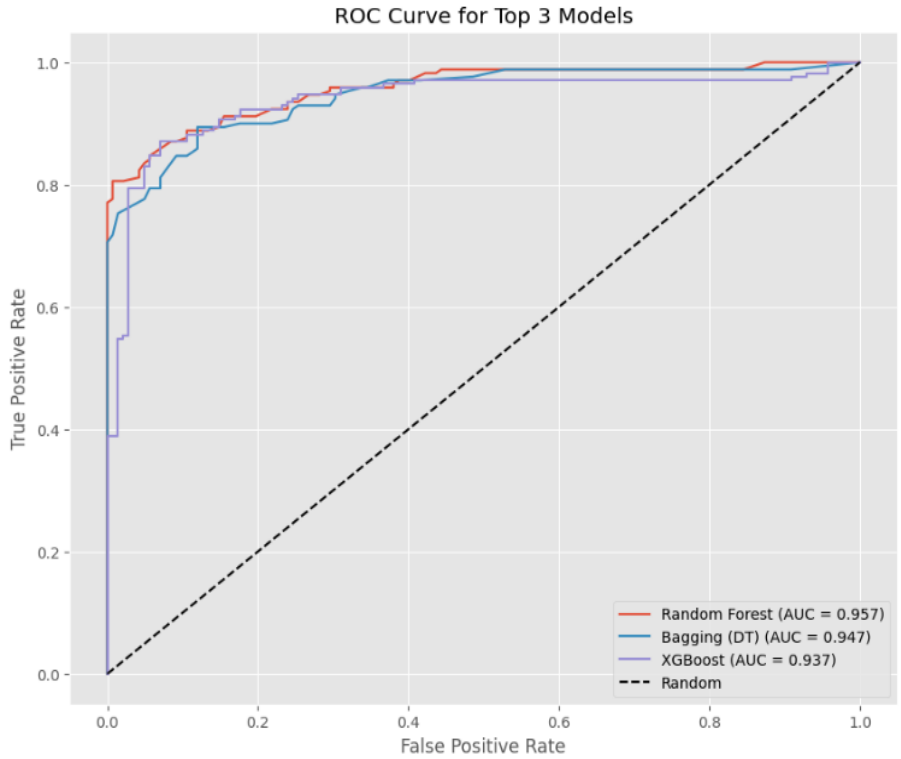

Figure 2. *ROC AUC curve for the Top 3 Models.*

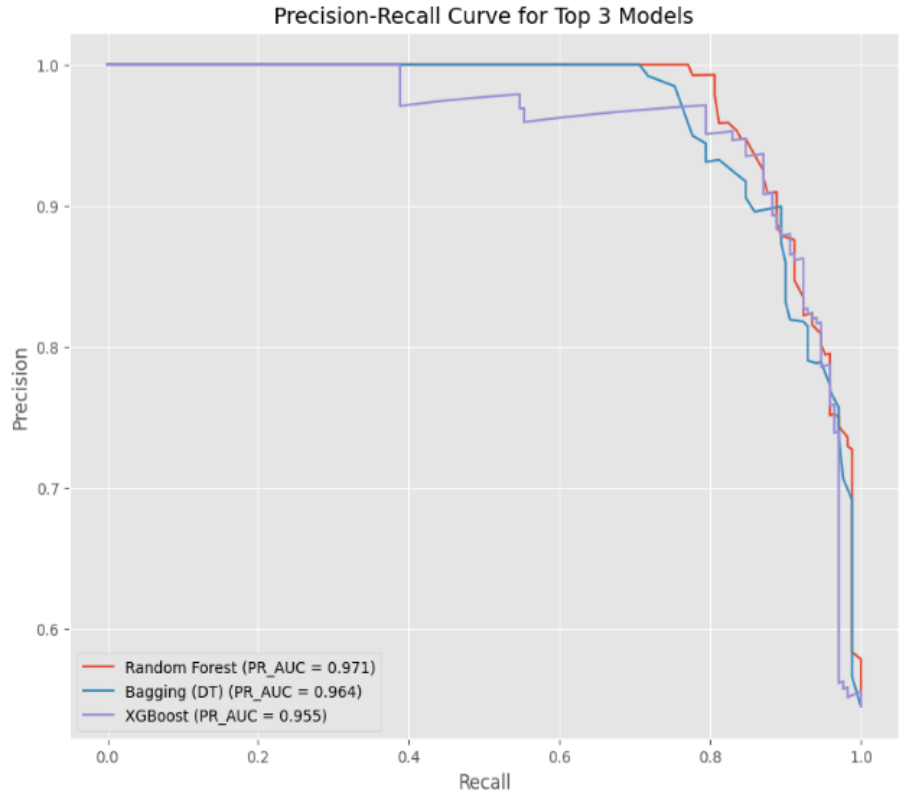

Figure 3. *Precision - Recall AUC curve for the Top 3 Models.*

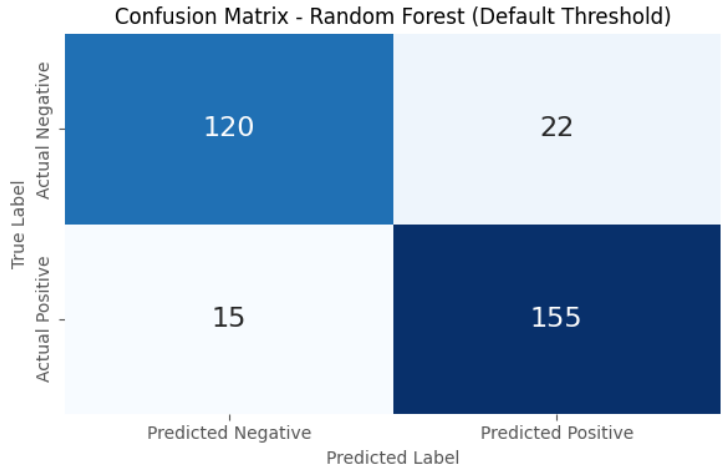

Figure 4. *Confusion Matrices for Random Forest Model at Default Threshold.*

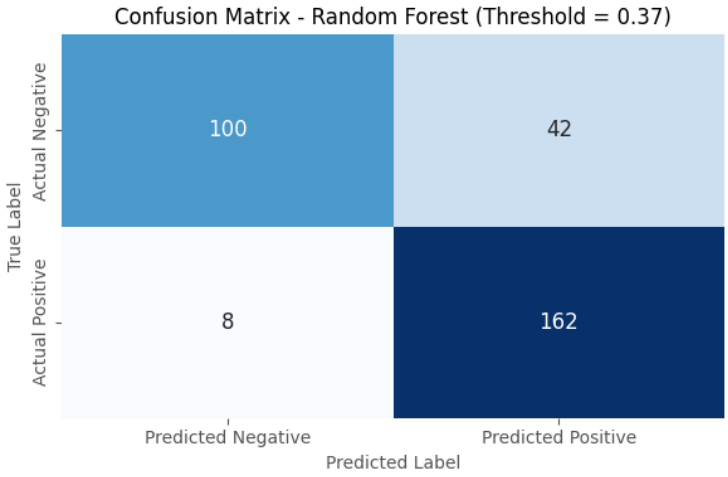

Figure 5. *Confusion Matrices for the Random Forest Model at Threshold=0.37.*

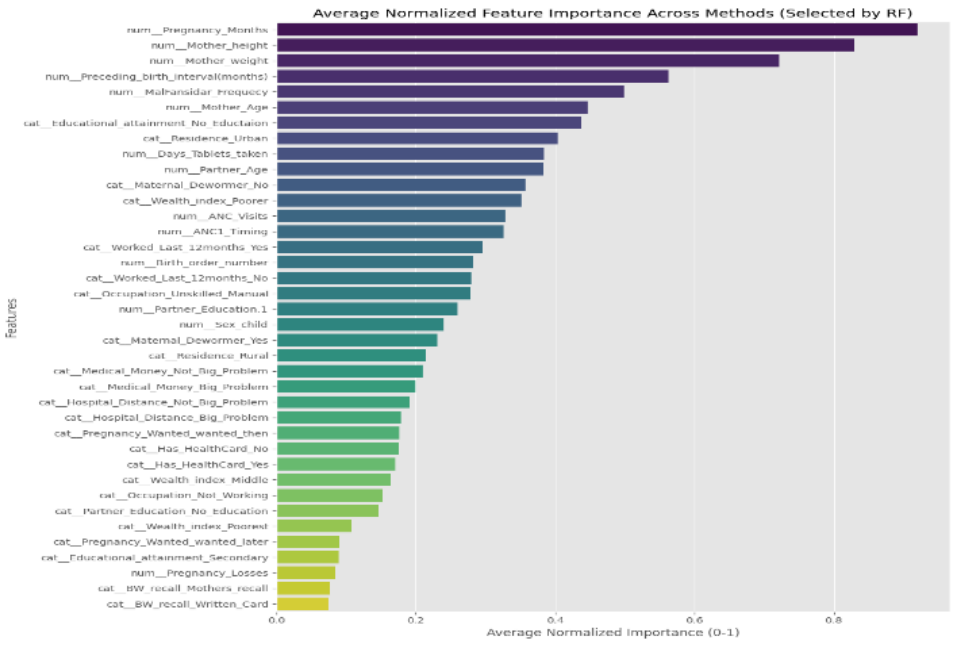

Figure 6. *Average Normalized Feature Importance Across Methods (Selected by Random Forest).*

Feature importance patterns were similar across tree-based methods, and pregnancy month, maternal height, and maternal weight were always among the best predictors. These signals are consonant with established biological and clinical plausibility: maternal anthropometry and gestational age directly affect fetal growth and birthweight. Other ANC and access-to-care contributors (e.g., ANC visits, iron supplementation, access barriers) added further incremental predictive utility for the multifactorial nature of risk for LBW in the Kenyan setting.

4. Discussion

4.1. Feature Importance

The analysis of average normalized feature importance, derived from an ensemble of robust models, reveals a clinically coherent hierarchy of factors predictive of low birthweight (LBW) within the Kenyan context. The results underscore a complex etiology where immediate biological and anthropometric determinants are powerfully moderated by broader healthcare utilization and socioeconomic conditions, aligning with contemporary socio-ecological models of health [11].

The most influential predictors were directly related to maternal physical health and the pregnancy itself. Gestational age (Pregnancy_Months) emerged as the strongest predictor, a finding consistent with global physiological evidence that preterm birth is a primary direct cause of LBW [1]. This highlights the critical need for public health interventions aimed at preventing early labor. Furthermore, maternal height and weight were among the top features, serving as proxies for long-term nutritional status. Short stature and low pre-pregnancy weight are established risk factors for intrauterine growth restriction, emphasizing that nutritional interventions must begin before conception to build maternal reserves, a cornerstone of global strategies to reduce LBW [12, 13].

The significance of healthcare access and utilization is clearly demonstrated. A short preceding birth interval is a known risk factor, as it deprives mothers of adequate time for nutritional recovery, pointing to the vital role of family planning services [14]. The high importance of malaria chemoprophylaxis (Malfansidar) and iron-folic acid supplementation validates the effectiveness of these specific antenatal care (ANC) interventions. This finding is supported by recent studies confirming their role in combating anemia and malaria, which are major contributors to poor fetal growth in sub-Saharan Africa [15]. Consequently, the number and timing of ANC visits are crucial, as early and consistent attendance is the gateway to receiving these prophylactics and essential health education.

Underpinning these factors are profound socioeconomic and demographic disparities. Maternal education is a powerful social determinant, influencing health-seeking behaviors, nutritional knowledge, and empowerment within the household [16]. Similarly, wealth indices and place of residence (urban/rural) act as proxies for overall access to quality healthcare, nutritious food, and adequate sanitation. The model confirms that economic deprivation is a fundamental driver of LBW risk. This is further exacerbated by direct barriers to care, as indicated by the importance of financial constraints (Medical_Money_Problem) and geographic distance (Hospital_Distance_Problem), which remain significant obstacles in low-resource settings and are key targets for health system strengthening [17].

Finally, the results include notable observations that bolster the model's construct validity. Paternal factors, such as partner's age and education, were present but less impactful than maternal characteristics, suggesting the mother's health and context are more directly determinative of birth outcomes, a finding consistent with the literature on the biological primacy of the maternal-host environment [12]. Furthermore, the low importance of features related to the method of birthweight recall (e.g., BW_recall_Mothers_recall) indicates the model correctly prioritized true predictive factors over mere metadata, confirming its focus on causal pathways and enhancing confidence in its interpretability.

4.2. Machine Learning Models

This study demonstrates that ensemble learning methods, particularly tree-based ensembles, are exceptionally well-suited for analyzing tabular, heterogeneous maternal-child health data, achieving robust predictive performance for LBW risk stratification in Kenya. The Random Forest algorithm emerged as the highest-performing model, demonstrating superior global discrimination (ROC AUC: 0.957) and excellent probability calibration (Brier Score: 0.089). By extension, its high F1-score strongly infers a superior Precision-Recall AUC, a critical metric for imbalanced classification tasks that is more informative than ROC AUC when the positive class is rare [18, 19]. This combination of high discrimination and calibration is paramount for public health applications, as it enables the reliable ranking of individuals by risk and the communication of trustworthy probabilities, both essential for effective triage, resource allocation, and shared decision-making [20].

While XGBoost achieved maximum threshold-specific accuracy (87.5%) with comparable precision-recall performance, its global discrimination and calibration were slightly inferior to Random Forest. This suggests that for a fixed, pre-defined workflow, XGBoost is extremely powerful; however, Random Forest offers more stable and trustworthy performance across all potential decision thresholds. This provides greater flexibility for probability-based intervention support, such as dynamically prioritizing ANC outreach based on evolving risk levels and capacity constraints [21]. The

soft-voting meta-ensemble provided competitive but not superior performance, indicating little incremental gain from a naive combination of base learners, a common finding when one strong learner (like Random Forest) already captures the majority of the predictive signal [22].

The clinical validity of the models is significantly enhanced by their feature importance profiles. The dominance of gestational age (Pregnancy_Months) and maternal anthropometrics (Mother_height, Mother_weight) is firmly grounded in fetal growth physiology and aligns with global health priorities for reducing LBW [13]. Concurrently, the strong influence of healthcare access variables (ANC_Visits, Days_Tablets_taken) highlights modifiable determinants, providing an evidence-based roadmap for public health interventions targeting early ANC initiation, iron supplementation, and the mitigation of financial and geographic barriers [23].

From an implementation perspective, Random Forest offers practical advantages for low-resource settings. Its robustness to outliers and non-linearities, coupled with relative insensitivity to hyperparameter tuning, makes it easier to train and maintain within typical health information systems compared to gradient-boosting algorithms like XGBoost, which often require more stringent optimization and external calibration to achieve reliable probability estimates [22]. This operational simplicity is a key consideration for sustainable deployment in real-world public health platforms.

4.3. Ethical Considerations

Integrating the use of machine learning models within the maternal and neonatal healthcare sector in Kenya requires careful analysis of ethical concerns. The sanctity of health information involved in predicting the birthweight of an infant requires extreme diligence in enforcing the applicable data use and patient information privacy rules. Procedures for informed consent should be straightforward and easy to understand, especially in the use of personal data for training and predictive models. In addition, the issues of fairness and equity should be addressed to avoid algorithmic bias that disproportionately impacts the most vulnerable as a result of data and systemic inequities. The use of artificial intelligence should conform to the ethical principles that carefully govern healthcare innovations in Kenya, in a manner that empowers health workers and patients, rather than isolating them.

4.4. Scalability and Adaptability Beyond Kenya

Although the current study is based on the Kenyan population and utilizes nationally representative data, the ensemble machine learning model development pipeline and approach have the potential to be adapted to other low-resource environments with similar socio-economic and healthcare challenges. Nevertheless, local contextual differences, including population demographics, health care infrastructure, and epidemiological characteristics, might also affect model performance. Models ought to be retrained or fine-tuned with a regional data set before they can be deployed. This will allow customizing the forecast to new settings as well as taking advantage of the methodological strengths that have been evident in this case, leading to larger changes in neonatal care in the sub-Saharan region and similar areas.

5. Conclusions

Based on the observed measures and operational context, Random Forest is the recommended ensemble method for forecasting the risk of LBW in Kenya. It had the highest overall discrimination (ROC AUC 0.957), was well-calibrated (Brier score 0.089), and had good accuracy and F1, competing at thresholds, and delivered a solid balance of performance, reliability, and deployability. XGBoost remains a decent second option when workflows appreciate having a specific decision threshold and can handle calibration, but that does not take precedence over Random Forest's improved global discrimination and calibration in such a case. A simple soft-voting meta-ensemble did not always outperform Random Forest. For possible future work, a well-calibrated stacking approach combining Random Forest and XGBoost as a calibrated meta-learner can be explored, but only as an alternative to Random Forest when external validation demonstrates a persistent and significant gain in discrimination without a loss of calibration. Otherwise, though, Random Forest is a parsimonious, high-performing, and implementation-simplified way to maximize predictive accuracy of low birthweight risk in Kenya.

Abbreviations

ANC	Antenatal Care
AUC	Area Under the Curve
BW	Birth Weight
CV	Cross-Validation
KDHS	Kenya Demographic and Health Survey
LBW	Low Birth Weight
ML	Machine Learning
MCC	Matthews Correlation Coefficient
PLR	Positive Likelihood Ratio
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	Extreme Gradient Boosting

Author Contributions

Victor Opiyo: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Emma Anyika: Supervision, Validation, Writing – original draft, Writing – review & editing

Funding

This work is not supported by any external funding.

Data Availability Statement

The data is available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] WHO, "Born too soon: decade of action on preterm birth." Accessed: Feb. 08, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240073890

[2] A. Ranjbar, F. Montazeri, M. V. Farashah, V. Mehrnoush, F. Darsareh, and N. Roozbeh, "Machine learning-based approach for predicting low birth weight," BMC Pregnancy Childbirth, vol. 23, no. 1, p. 803, Nov. 2023, https://doi.org/10.1186/s12884-023-06128-w

[3] D. Unicef, "Low birthweight," UNICEF DATA. Accessed: Feb. 08, 2025. [Online]. Available: https://data.unicef.org/topic/nutrition/low-birthweight/

[4] A. K'Oloo et al., "Improving birth weight measurement and recording practices in Kenya and Tanzania: a prospective intervention study with historical controls," Popul. Health Metr., vol. 21, no. 1, p. 6, May 2023, https://doi.org/10.1186/s12963-023-00305-x

[5] S. J. Sawe, "MACHINE LEARNING PREDICTION OF LOW BIRTH WEIGHT IN KENYA USING MATERNAL RISK FACTORS," 2022.

[6] J. Lanowski, J. von Ehr, and M., "Impact of Ultrasound Training and Experience on Accuracy regarding Fetal Weight Estimation at Term Creative Education." Accessed: Aug. 05, 2025. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=79172

[7] W. T. Bekele, "Machine learning algorithms for predicting low birth weight in Ethiopia," BMC Med. Inform. Decis. Mak., vol. 22, no. 1, p. 232, Sept. 2022, https://doi.org/10.1186/s12911-022-01981-9

[8] S. Sanchez-Martinez et al., "Prediction of low birth weight from fetal ultrasound and clinical characteristics: a comparative study between a low- and middle-income and a high-income country," BMJ Glob. Health, vol. 9, no. 12, p. e016088, Dec. 2024, https://doi.org/10.1136/bmjgh-2024-016088

[9] Rubaiya, Mohaimen Mansur, and Md. I. Rayhan, "Unraveling birth weight determinants: Integrating machine learning, spatial analysis, and district-level mapping," Heliyon, vol. 10, no. 5, p. e27341, Mar. 2024, https://doi.org/10.1016/j.heliyon.2024.e27341

[10] M. M. Musau et al., "Spatial heterogeneity of low-birthweight deliveries on the Kenyan coast," BMC Pregnancy Childbirth, vol. 23, no. 1, p. 270, Apr. 2023, https://doi.org/10.1186/s12884-023-05586-6

[11] Z. D. Bailey, J. M. Feldman, and M. T. Bassett, "How Structural Racism Works - Racist Policies as a Root Cause of U.S. Racial Health Inequities," N. Engl. J. Med., vol. 384, no. 8, pp. 768–773, Feb. 2021, https://doi.org/10.1056/NEJMms2025396

[12] N. Kozuki et al., "The associations of parity and maternal age with small-for-gestational-age, preterm, and neonatal and infant mortality: a meta-analysis," BMC Public Health, vol. 13 Suppl 3, no. Suppl 3, p. S2, 2013, https://doi.org/10.1186/1471-2458-13-S3-S2

[13] W. H. Organization, "Global nutrition targets 2025: low birth weight policy brief," Art. no. WHO/NMH/NHD/14.5, 2014, Accessed: Jan. 06, 2025. [Online]. Available: https://iris.who.int/handle/10665/149020

[14] J. Molitoris, K. Barclay, and M. Kolk, "When and Where Birth Spacing Matters for Child Survival: An International Comparison Using the DHS," Demography, vol. 56, no. 4, pp. 1349–1370, Aug. 2019, https://doi.org/10.1007/s13524-019-00798-y

[15] Y. I. Coulibaly et al., "The Impact of Six Annual Rounds of Mass Drug Administration on Wuchereria bancrofti Infections in Humans and in Mosquitoes in Mali," Am. J. Trop. Med. Hyg., vol. 93, no. 2, pp. 356–360, Aug. 2015, https://doi.org/10.4269/ajtmh.14-0516

[16] G. Rezaeizadeh et al., "Maternal education and its influence on child growth and nutritional status during the first two years of life: a systematic review and meta-analysis," eClinicalMedicine, vol. 71, p. 102574, Apr. 2024, https://doi.org/10.1016/j.eclinm.2024.102574

[17] Y. B. Okwaraji, S. Cousens, Y. Berhane, K. Mulholland, and K. Edmond, "Effect of Geographical Access to Health Facilities on Child Mortality in Rural Ethiopia: A Community Based Cross Sectional Study," PLoS ONE, vol. 7, no. 3, p. e33564, Mar. 2012, https://doi.org/10.1371/journal.pone.0033564

[18] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PloS One, vol. 10, no. 3, p. e0118432, 2015, https://doi.org/10.1371/journal.pone.0118432

[19] A. Luque, A. Carrasco, A. Mart ń, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognit., vol. 91, pp. 216–231, July 2019, https://doi.org/10.1016/j.patcog.2019.02.023

[20] E. W. Steyerberg et al., "Assessing the performance of prediction models: a framework for traditional and novel measures," Epidemiol. Camb. Mass, vol. 21, no. 1, pp. 128–138, Jan. 2010, https://doi.org/10.1097/EDE.0b013e3181c30fb2

[21] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," Casp. J. Intern. Med., vol. 4, no. 2, pp. 627–635, 2013.

[22] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 6, pp. 7499–7519, June 2024, https://doi.org/10.1109/TNNLS.2022.3229161

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Biography

Victor Opiyo, I am currently studying for an MSc. Data Science student at The Co-operative University of Kenya, having received my B.Sc. Biostatistics from Jomo Kenyatta University of Agriculture and Technology, Kenya in 2016. I am currently working as a Strategic Information Specialist at CIHEB Kenya. Previously worked as Program Data Manager at Centre for Health Solutions, Kenya, between 2017 and 2019. From November 2019 to September 2021, I was a Program Strategic Information Officer at the University of Maryland Baltimore, Kenya.