Discover Artificial Intelligence

RESEARCH Open Access

Early risk prediction of gestational diabetes using routine antenatal care clinical data using machine learning

Emmanuel Ahishakiye[1*], Justine Nakirijja[1] and Shallon Ahimbisibwe[1]

*Correspondence:
Emmanuel Ahishakiye
ahishema@gmail.com
[1]School of Computing and Information Science, Kyambogo University, Kampala, Uganda

Abstract

Gestational diabetes mellitus (GDM) is a major contributor to adverse maternal and neonatal outcomes, particularly in low-resource settings where timely diagnostic testing is not always feasible. Early identification of high-risk pregnancies using routinely collected antenatal information could support targeted screening and preventive care. This study evaluated the feasibility of early GDM risk prediction using machine learning models trained on routine antenatal care data from 3525 pregnancies in Uganda (2153 non-GDM and 1372 GDM cases). Logistic regression, Random Forest, XGBoost, and a soft-voting ensemble were developed and assessed using a held-out test set. Model performance was evaluated using receiver operating characteristic area under the curve (ROC AUC), precision–recall AUC (PR AUC), accuracy, precision, recall, and F1-score, and interpretability was examined using SHapley Additive exPlanations (SHAP). Tree-based models demonstrated strong discrimination, with Random Forest and XGBoost achieving ROC AUC values of 0.997 and PR AUC values above 0.995, while logistic regression achieved ROC AUC of 0.982. Random Forest achieved high sensitivity (recall = 0.994), missing only two GDM cases in the test set. SHAP analysis identified body mass index, high-density lipoprotein cholesterol, blood pressure, and prior metabolic history as influential predictors, with feature effects consistent with established clinical knowledge. The proposed approach is intended as an early risk-stratification and triage support tool rather than a diagnostic system. All predictors were available prior to confirmatory glucose testing, and model outputs are designed to prioritise women for further evaluation and follow-up. Although performance estimates reflect internal validation within a referral-hospital cohort, the findings demonstrate that routinely collected antenatal variables contain clinically meaningful predictive information. With external and prospective validation, interpretable machine learning models may support more efficient screening and monitoring of gestational diabetes in resource-constrained antenatal care settings.

Keywords Gestational diabetes, Antenatal care data, Early risk prediction, Machine learning, Clinical decision support

1 Introduction

Gestational diabetes mellitus (GDM) is a common metabolic disorder of pregnancy characterized by glucose intolerance first recognized during gestation. The condition is associated with serious short- and long-term consequences for both mother and child, including pre-eclampsia, obstructed labour, macrosomia, neonatal hypoglycaemia, and increased lifetime risk of type 2 diabetes mellitus and cardiovascular disease [1, 2]. Because these complications are partly preventable with timely intervention, early identification of women at elevated risk has become an important objective of antenatal care.

Globally, the prevalence of GDM varies widely depending on population characteristics and screening criteria, with estimates typically ranging from 5 to 18% [3]. However, growing evidence suggests that the burden is rising in low- and middle-income countries due to rapid urbanization, nutritional transition, and increasing maternal age and obesity [4, 5]. In Uganda, antenatal clinics, particularly those at national referral hospitals, serve heterogeneous populations that include a substantial proportion of high-risk pregnancies. Despite this, systematic early screening remains inconsistent, and many women are evaluated only after symptoms or complications appear [6]. Consequently, opportunities for early preventive counselling and monitoring are frequently missed.

Standard diagnosis of GDM relies on the oral glucose tolerance test (OGTT), which requires fasting, multiple blood samples, laboratory reagents, trained personnel, and repeat clinic attendance [7]. These requirements create practical challenges in many resource-constrained antenatal care (ANC) settings, including limited laboratory infrastructure, delayed test scheduling, long patient waiting times, and loss to follow-up between visits [8]. As a result, routine universal screening is difficult to sustain, and selective screening based solely on clinician judgement may fail to identify women at risk early in pregnancy. There is therefore a clinical need for low-cost pre-screening approaches that can operate using information already collected during routine ANC visits [8].

Early risk prediction refers, in this study, to prediction using information available at or shortly after the first antenatal visit, prior to routine diagnostic glucose testing. Such prediction is intended to function as a clinical decision-support triage tool to prioritize women for confirmatory testing and closer follow-up, rather than to replace diagnostic procedures. An effective early risk model could support timely counselling on diet, physical activity, and monitoring, thereby reducing preventable complications.

Recent advances in machine learning (ML) have demonstrated strong potential for clinical risk prediction because ML algorithms can capture complex, nonlinear interactions among multiple patient characteristics [9, 10]. Several studies have applied ML to GDM prediction and reported improved discrimination compared with traditional statistical risk scores [11]. However, many existing models depend on specialised biochemical measurements, longitudinal electronic health records, or variables collected at or near the diagnostic glucose test, limiting their applicability in low-resource clinical environments. Furthermore, many published models emphasize predictive performance without sufficient attention to interpretability, transparency, and clinician trust, factors that are particularly important in maternal healthcare decision-making [12–14].

In addition, evidence from sub-Saharan Africa remains limited. Models developed in high-income populations may not generalize well due to differences in genetics, nutrition, healthcare access, and demographic structure [15]. The lack of locally validated

predictive tools therefore represents a critical gap in applying artificial intelligence to maternal health in African settings [16]. A model developed and evaluated using routinely collected Ugandan ANC data could provide more contextually appropriate risk stratification and inform future multi-site validation.

Another important consideration is explainability. Clinical adoption of artificial intelligence systems depends not only on predictive accuracy but also on whether healthcare providers can understand and trust model recommendations [17]. In maternal healthcare, this requirement is particularly important because risk assessments may directly influence monitoring intensity, counselling, referral decisions, and patient anxiety during pregnancy. Unlike administrative prediction tasks, clinical decisions involve ethical accountability and shared decision-making between providers and patients. Opaque "black-box" predictions may therefore be difficult to justify in practice, especially when they affect clinical advice given to pregnant women. Interpretability methods, such as feature importance analysis and SHapley Additive exPlanations (SHAP), allow clinicians to examine how maternal characteristics influence predicted risk and to verify whether predictions align with established clinical understanding of gestational metabolic risk. Transparent explanations enable clinicians to assess whether the model is reasoning plausibly rather than relying on spurious correlations, thereby improving trust, accountability, and safe integration into routine care workflows.

Accordingly, this study addresses three key research questions:

RQ1 Can routinely collected antenatal care data be used to predict gestational diabetes risk with clinically meaningful discrimination?

RQ2 Do non-linear machine learning models outperform traditional logistic regression for GDM risk prediction?

RQ3 Are model explanations clinically plausible and consistent with known maternal risk factors?

To answer these questions, we develop and evaluate supervised learning models using de-identified routine antenatal records from Kawempe National Referral Hospital in Uganda. We compare logistic regression with two non-linear algorithms, Random Forest and Extreme Gradient Boosting (XGBoost), and also examine a soft-voting ensemble. Model interpretability is assessed using feature importance and SHAP analysis to identify clinically relevant predictors.

This paper contributes to AI-enabled maternal risk prediction in resource-constrained settings in three ways. First, it demonstrates the feasibility of predicting GDM risk using routinely available ANC variables that can be obtained without specialised diagnostic glucose testing procedures such as the oral glucose tolerance test. Second, it integrates explainable machine learning to enhance transparency and clinician trust. Third, it provides locally grounded empirical evidence from a Ugandan referral-hospital cohort, supporting future multi-site validation and deployment studies in sub-Saharan Africa. Importantly, the study does not exclude routine laboratory measurements entirely; rather, it avoids reliance on diagnostic glucose-based testing and focuses on measurements commonly obtained during standard antenatal booking investigations.

1.1 Contributions and organization of the paper

This paper contributes to the growing field of AI-assisted maternal risk prediction with a focus on practical deployment in resource-constrained antenatal care settings.

- First, we develop and compare multiple supervised learning models using routinely collected antenatal care (ANC) variables available at or near the first clinical visit. By evaluating logistic regression alongside non-linear methods (Random Forest and Extreme Gradient Boosting), the study examines whether machine learning can improve early risk stratification for gestational diabetes using data that can realistically be obtained in typical clinical workflows.
- Second, we incorporate an explainable artificial intelligence framework to enhance clinical transparency. Model interpretability is examined using feature importance analysis, logistic regression coefficient inspection, and SHapley Additive exPlanations (SHAP). These analyses allow assessment of whether model predictions align with known maternal risk factors and support clinician trust in decision-support systems.
- Third, the study provides locally grounded evidence from a Ugandan referral-hospital cohort, a setting in which published machine-learning applications for maternal metabolic risk prediction remain limited. The findings therefore contribute to contextual model evaluation and inform future multi-site validation and implementation studies in sub-Saharan Africa.

1.2 Organization of the paper

The remainder of the paper is organized as follows. Section 2 reviews related work on gestational diabetes prediction and clinical machine learning. Section 3 describes the dataset, preprocessing, and modelling methodology. Section 4 presents the experimental results and model performance evaluation. Section 5 discusses clinical interpretation, limitations, and deployment considerations. Section 6 concludes the paper and outlines directions for future research.

2 Related literature review

2.1 Gestational diabetes mellitus burden and risk landscape

Gestational diabetes mellitus (GDM) represents a significant pregnancy-related metabolic disorder with implications for both maternal and neonatal health outcomes worldwide. Reported prevalence varies widely, generally ranging between 5 and 18%, depending on population characteristics and diagnostic criteria [18, 19]. Women diagnosed with GDM are more likely to experience complications such as hypertensive disorders, operative delivery, fetal overgrowth, and an increased future risk of developing type 2 diabetes mellitus [20]. In sub-Saharan Africa, including Uganda, the frequency of GDM is rising alongside demographic and lifestyle changes such as urbanisation, increasing obesity rates, and altered dietary patterns [2]. Despite this increasing burden, effective screening remains inconsistent. Many health facilities face limited laboratory capacity, delayed detection, poor adherence to screening protocols, and weak referral systems, all of which hinder appropriate management [21]. For this reason, improving early identification through routinely collected antenatal information has been proposed as an important strategy for enhancing maternal care in resource-constrained environments.

2.2 Traditional risk scoring and clinical screening limitations

Historically, GDM risk assessment has depended on clinician assessment or checklist-based scoring tools incorporating factors such as maternal age, parity, family history, and anthropometric indicators [22]. Although straightforward to apply, these approaches frequently fail to detect high-risk individuals, particularly in diverse populations where metabolic variation does not follow simple threshold patterns. Statistical prediction techniques, including logistic regression, have also been employed; however, these models typically assume linear relationships between predictors and outcomes. Such assumptions limit their ability to represent complex biological interactions among metabolic and cardiovascular factors, a weakness commonly reported in maternal risk modelling studies [23]. Consequently, researchers have increasingly explored computational approaches capable of learning more complex patterns within clinical data.

2.3 Application of machine learning and artificial intelligence in GDM prediction

Machine learning methods have become increasingly prominent in medical prediction tasks because they can capture hidden relationships and multidimensional interactions within large datasets [24]. Investigations conducted across different regions, including China, India, and the Middle East, have demonstrated that supervised learning algorithms such as Random Forest, gradient boosting, and support vector machines often surpass conventional statistical models in identifying pregnancies at risk of GDM [11, 25]. In settings where detailed biochemical and clinical data are available, these models frequently achieve area-under-curve values exceeding 0.90 [26]. However, most existing studies rely on comprehensive laboratory tests and structured clinical datasets collected in high-income health systems. Such dependence limits applicability to low-resource environments. A recent systematic review reported a shortage of predictive models built exclusively on routine clinical observations without biochemical measurements, despite the widespread availability of such variables in many African facilities [27]. This highlights the need for affordable, explainable screening tools based on routinely captured antenatal indicators.

Recent comparative machine learning research in clinical prediction has emphasized the importance of benchmarking multiple algorithms rather than relying on a single modelling approach [28–30]. Multi-algorithm healthcare studies have shown that ensemble tree-based models, including Random Forest and gradient boosting methods, often achieve higher discrimination than traditional statistical techniques such as logistic regression, particularly when modelling non-linear physiological relationships [31, 32]. However, these studies also demonstrate that predictive performance depends strongly on dataset characteristics, feature availability, and class distribution rather than algorithm selection alone [32, 33]. In addition, recent medical artificial intelligence research highlights that strong internal performance does not necessarily translate into generalisable clinical performance without external validation across different patient populations and healthcare settings [34–38]. These findings support the use of comparative modelling strategies and contextual interpretation of performance metrics when developing predictive models for gestational diabetes risk stratification.

2.4 Feature explainability and the demand for transparent AI in maternal care

While predictive accuracy is important, clinical implementation of artificial intelligence depends strongly on interpretability. Opaque "black-box" models can reduce clinician confidence and complicate integration into routine workflows [39]. To overcome this challenge, recent research has promoted interpretable machine learning approaches, including decision trees, feature importance analysis, and Shapley Additive Explanations (SHAP) [40]. These methods allow practitioners to understand how individual patient characteristics influence predictions and to verify consistency with biological reasoning. In maternal health applications, transparency is especially important because automated risk assessments may affect clinical decisions and patient counselling. Ethical considerations related to accountability and potential misclassification therefore make explainable models preferable in healthcare environments [41].

2.5 Digital maternal health systems in low-resource settings

Digital health technologies and predictive analytics are increasingly recommended to strengthen maternal healthcare delivery in low-income regions. Such systems can support earlier referral, improved triage, and better surveillance of high-risk pregnancies [42]. Uganda and several other African countries have begun implementing electronic medical record systems, yet advanced clinical decision-support tools based on artificial intelligence remain limited [43]. Evidence from implementations in countries such as South Africa and India suggests that machine learning-based maternal triage platforms can improve detection and prioritisation of high-risk cases when incorporated into routine antenatal workflows [44]. Nevertheless, few solutions utilise interpretable models built solely from routinely collected clinical data, reinforcing the need for further research in this area.

2.6 Research gaps

Existing studies demonstrate the potential of artificial intelligence in predicting GDM, but important limitations remain. Many models depend on laboratory biomarkers rather than measurements available during routine antenatal visits. Additionally, relatively few investigations incorporate interpretability techniques to support clinical acceptance. Finally, empirical evidence from African healthcare settings is still limited. In particular, a number of previously proposed prediction models rely on specialised diagnostic glucose testing (e.g., fasting plasma glucose or oral glucose tolerance testing) or expanded biochemical panels that may not be available during early antenatal visits in low-resource settings. In contrast, routine antenatal booking assessments typically include basic demographic information, vital signs, anthropometric measurements, and a limited set of standard laboratory investigations. Distinguishing between specialised diagnostic testing and routinely collected booking measurements is important when evaluating the feasibility of early screening tools in resource-constrained clinical environments. The present study addresses these challenges by developing predictive models using routine antenatal care data, rigorously evaluating their performance, and examining decision behaviour through SHAP analysis, feature attribution, and tree-based interpretation. The findings aim to support scalable, interpretable, and cost-effective screening strategies for maternal health programmes in resource-constrained settings.

3 Materials and methods

3.1 Dataset description

This study utilised a retrospective, de-identified clinical dataset comprising antenatal and basic laboratory records collected from pregnant women attending routine antenatal care services in Uganda. The dataset contains 3525 patient-level observations, with each record representing a unique pregnancy episode. All personal identifiers were removed prior to analysis to ensure confidentiality and compliance with ethical data-governance requirements. Each record includes a set of maternal demographic, anthropometric, vital sign, and basic laboratory variables routinely collected during antenatal visits in public health facilities. Demographic attributes include maternal age and parity. Anthropometric and physiological measurements include body mass index (BMI), systolic and diastolic blood pressure, and haemoglobin concentration. The dataset further contains lipid profile indicators such as high-density lipoprotein (HDL) cholesterol, alongside other routinely recorded clinical variables relevant to metabolic and cardiovascular risk assessment during pregnancy. A unique case identifier was present in the raw dataset but was excluded from modelling as it carries no predictive information. The outcome variable is a binary class label indicating gestational diabetes mellitus (GDM) status, denoted as GDM or Non-GDM. Importantly, this label reflects clinically determined gestational diabetes status as recorded in patient charts, based on standard diagnostic practice within antenatal care services (including laboratory assessment and clinical evaluation). The outcome label was not constructed using deterministic rules derived from predictor variables, ensuring that the prediction task represents a valid supervised learning problem rather than rule reconstruction. Within the dataset, 1372 records (38.9%) correspond to women diagnosed with GDM, while 2153 records (61.1%) represent non-GDM cases. This class distribution reflects a moderate imbalance typical of real-world antenatal screening data and motivates the use of evaluation metrics sensitive to minority-class detection. The dataset size and feature composition provide sufficient statistical power to train and evaluate machine-learning models while remaining representative of routine antenatal care contexts in low-resource settings. To preserve methodological validity and prevent information leakage, variables that are diagnostic or post-diagnostic in nature were carefully assessed prior to modelling. Only predictors plausibly available before or at the time of early antenatal screening were considered for inclusion in the predictive feature set, consistent with the study's objective of supporting early risk stratification rather than retrospective diagnosis.

Because the dataset originates from a national referral hospital, the study population may include a higher proportion of high-risk pregnancies than community-level antenatal clinics. Consequently, the model was developed as a proof-of-concept early risk stratification tool within a referral-care context rather than as a universally generalisable screening model. Evaluation across additional healthcare facilities and population groups will be required before broader deployment.

The variable "prediabetes status" represents a documented medical history recorded in the patient chart prior to the index pregnancy or at the initial antenatal interview. It does not originate from glucose testing performed during the current pregnancy. Therefore, it reflects pre-existing metabolic risk information rather than a diagnostic measurement used to establish gestational diabetes in the study cohort.

3.1.1 Study timing and definition of early prediction

The objective of this study was early risk prediction rather than diagnostic classification. "Early" was operationally defined as prediction using information available at or shortly after the first antenatal care (ANC) visit and prior to routine gestational diabetes diagnostic testing. All predictor variables were obtained from measurements routinely recorded during the initial antenatal assessment, including demographic information, obstetric history, anthropometric measurements, vital signs, and standard booking laboratory investigations. These measurements are typically collected during the first clinical encounter and precede formal gestational diabetes screening in routine care pathways. The outcome variable (GDM diagnosis) was determined later in pregnancy according to the hospital's clinical screening protocol. Predictor variables were therefore temporally separated from the diagnostic outcome, ensuring that the model used only information available before confirmatory glucose testing. The variable prediabetes status was defined as a previously documented medical history recorded in patient records at or before the initial antenatal interview. It was not derived from glucose testing conducted during the index pregnancy. Consequently, this variable represents baseline metabolic risk rather than diagnostic evidence of gestational diabetes. No glucose measurements used to diagnose gestational diabetes, including fasting plasma glucose or oral glucose tolerance testing values, were included as predictive features. This temporal ordering ensures that the modelling task reflects prospective early risk stratification rather than retrospective reconstruction of clinical diagnosis.

3.2 Feature selection and leakage control

Careful feature selection was undertaken to ensure methodological validity and to prevent information leakage between predictors and the outcome variable. Because gestational diabetes mellitus (GDM) is a clinically diagnosed condition, particular attention was paid to excluding any variables that could directly encode diagnostic decisions or post-diagnosis information, which would artificially inflate model performance and compromise generalisability. An initial audit of all variables was performed to classify features into three categories: (i) pre-diagnostic screening variables, (ii) diagnostic or post-diagnostic variables, and (iii) non-informative identifiers. Only variables plausibly available prior to or at early antenatal screening were retained as candidate predictors. Variables serving as administrative identifiers (e.g., case numbers) were removed, as they carry no clinical relevance and risk unintended data leakage. Where laboratory-derived variables were present, their inclusion was evaluated in relation to the clinical objective of the study. Variables that are routinely used to define or confirm gestational diabetes diagnosis were excluded from the predictive feature set to avoid circular learning. Basic laboratory measurements retained in the dataset (e.g., haemoglobin concentration and lipid profile indicators such as HDL cholesterol) were part of routine antenatal booking investigations and were recorded prior to diagnostic glucose testing. These measurements are not used to establish gestational diabetes diagnosis and therefore do not encode the outcome label. Consequently, their inclusion reflects pre-diagnostic clinical information rather than post-diagnostic confirmation, preserving the predictive nature of the modelling task and reducing the risk of information leakage. The final feature set therefore comprised maternal demographic attributes (such as age and parity), anthropometric measurements (including body mass index), vital signs (systolic and diastolic

blood pressure), haematological indicators (haemoglobin concentration), and other routinely captured metabolic or cardiovascular markers recorded during antenatal visits. These variables reflect information commonly available in low-resource antenatal care settings and are suitable for early risk stratification. By explicitly separating label-defining information from predictive inputs, this study avoids the circular modelling pitfalls that can arise when outcome variables are deterministically derived from features used during training. This design ensures that predictive performance reflects genuine pattern learning rather than rule rediscovery, thereby strengthening the scientific validity, interpretability, and translational relevance of the proposed models.

Historical medical conditions (e.g., prior diagnosis of prediabetes) were retained because they represent patient history available at the first antenatal visit and are not derived from glucose testing conducted for gestational diabetes diagnosis during the same pregnancy.

3.3 Data preprocessing

A structured data preprocessing pipeline was applied to prepare the dataset for machine learning analysis while preserving clinical interpretability and preventing information leakage. All preprocessing steps were designed to reflect realistic deployment conditions in antenatal screening settings and were implemented in a reproducible manner. First, a completeness assessment was conducted across all candidate predictors. Variables exhibiting excessive missingness were evaluated for exclusion to minimise noise and reduce reliance on imputation. Features with more than 40% missing values were removed from further analysis, as such sparsity may introduce bias and instability in downstream models. This threshold balances information retention with data quality considerations commonly adopted in clinical prediction studies. For the remaining variables, missing values were handled using simple, clinically appropriate imputation strategies. Continuous features (e.g., body mass index, blood pressure, haemoglobin concentration, and lipid measures) were imputed using the median value computed from the training data only, which is robust to outliers and preserves the central tendency of skewed clinical distributions. Categorical variables, where present, were imputed using the mode. No outcome information was used during imputation to prevent target leakage. Categorical predictors were transformed using one-hot encoding, producing binary indicator variables for each category while avoiding ordinal assumptions. Continuous variables were retained in their original scale for tree-based models, which are invariant to monotonic transformations. For models sensitive to feature scaling, particularly logistic regression, standardisation (z-score scaling) was applied within the training pipeline to ensure that scaling parameters were learned exclusively from the training data and then applied to the test data. All preprocessing operations including feature exclusion, imputation, encoding, and scaling were embedded within model pipelines and executed after train–test splitting. This design ensured that no information from the evaluation data influenced feature transformation, thereby preserving the integrity of performance estimates. The resulting dataset was fully numeric, free of missing values, and suitable for robust supervised learning and cross-validation.

3.4 Handling class imbalance

The distribution of the outcome variable in the revised dataset reflects a moderate class imbalance, with 1372 cases (38.9%) diagnosed with gestational diabetes mellitus (GDM) and 2153 cases (61.1%) classified as non-GDM. This imbalance is representative of real-world antenatal screening populations and necessitates appropriate handling to ensure robust minority-class detection without inflating performance estimates. Rather than applying aggressive resampling techniques by default, this study adopted a conservative and model-aware strategy for addressing class imbalance. For baseline and linear models, including logistic regression, class-weighted loss functions were employed, assigning higher penalty to misclassification of GDM cases. This approach preserves the original data distribution while improving sensitivity to the minority class and is widely recommended for clinical prediction tasks. For tree-based and ensemble models, class imbalance was initially handled using internal class-weighting mechanisms (e.g., balanced class weights in Random Forest and adjusted loss weighting in gradient boosting). These methods allow the algorithms to account for imbalance during split optimisation without introducing synthetic data points. Synthetic Minority Over-sampling Technique (SMOTE) was considered only as a secondary strategy and applied exclusively within the training data when preliminary analyses indicated insufficient recall for GDM cases using weighting alone. When used, SMOTE was implemented inside the cross-validation folds to avoid information leakage, ensuring that no synthetic samples influenced model evaluation. This restrained application prevented the artefactual performance inflation associated with indiscriminate oversampling. Model performance was assessed using metrics sensitive to class imbalance, including recall, F1-score, precision–recall area under the curve (PR AUC), and receiver operating characteristic area under the curve (ROC AUC), rather than accuracy alone. This ensured that improvements in minority-class detection were not achieved at the expense of excessive false positives, which would reduce clinical utility in screening contexts. By prioritising weighting-based approaches and limiting oversampling to justified scenarios, the imbalance-handling strategy aligns with best practices in clinical machine learning and mitigates risks of overfitting, synthetic bias, and misleading performance estimates.

3.5 Model development

To evaluate the feasibility of predicting gestational diabetes mellitus (GDM) using routinely collected antenatal data, multiple supervised learning models with complementary inductive biases were developed. Model selection was guided by three principles: clinical interpretability, the ability to capture non-linear relationships among maternal risk factors, and robustness to moderate class imbalance commonly observed in antenatal datasets. Logistic regression was implemented as the baseline model due to its established use in clinical risk prediction and its transparent coefficient-based interpretation. The model was trained using a class-weighted loss function to account for imbalance between GDM and non-GDM cases. Continuous predictors were standardised within the training pipeline to ensure numerical stability and enable meaningful comparison of coefficient magnitudes. Logistic regression therefore served as a clinically interpretable benchmark against which more complex non-linear models could be assessed. To model non-linear interactions and threshold effects characteristic of metabolic risk factors, tree-based learning methods were also employed. A Random Forest classifier

was trained using an ensemble of decision trees built on bootstrap samples, with class-balanced splitting criteria to improve sensitivity to GDM cases. Random Forest models are well suited to structured clinical datasets because they handle mixed feature types, capture interaction effects, and reduce variance through aggregation of multiple trees. Feature importance measures derived from Random Forest additionally support clinical interpretability. An Extreme Gradient Boosting (XGBoost) model was included to further examine non-linear risk patterns using a boosting-based approach. Conservative hyperparameter settings were used to prioritise generalisation and minimise overfitting, including constrained tree depth, shrinkage through a low learning rate, and regularisation of split complexity. Class imbalance was addressed using loss weighting rather than aggressive resampling to ensure that performance improvements reflected genuine signal learning rather than synthetic data effects. An ensemble strategy was explored to assess whether combining complementary model predictions could improve robustness and reduce variance in risk estimation. Ensemble learning is motivated by the principle that aggregating multiple classifiers may stabilise predictions when individual models capture partially different aspects of the data. In this study, probabilistic outputs from the base learners were combined using a soft-voting approach rather than complex stacking architectures in order to preserve interpretability and limit overfitting risk. The ensemble was therefore included as an exploratory analysis rather than the primary predictive approach. All models were implemented using pipeline-based workflows, ensuring that preprocessing steps including imputation, encoding, scaling, and imbalance handling, were learned exclusively from the training data. Hyperparameters were selected conservatively based on established clinical machine learning practice rather than extensive optimisation on the test set. This modelling framework enables transparent comparison between linear and non-linear approaches and supports a realistic assessment of the added value of machine learning techniques for early GDM risk stratification in resource-constrained antenatal care settings.

3.6 Experimental design and validation strategy

Model evaluation was conducted using a structured and leakage-aware experimental design to ensure reliable estimation of predictive performance and generalisability. The dataset was partitioned into training and testing subsets using a stratified split, preserving the proportion of gestational diabetes mellitus (GDM) and non-GDM cases in both sets. The test set was held out completely and used only for final model evaluation. To prevent data leakage, dataset partitioning was performed prior to any preprocessing operations. The raw dataset was first divided into training and test subsets, and all preprocessing procedures were learned exclusively from the training data. Missing value imputation, categorical encoding, feature scaling, and class-imbalance handling were fitted using only the training subset. The learned preprocessing parameters were subsequently applied to the test data without re-estimation. Therefore, no statistical information from the evaluation data influenced model training or feature transformation. Within the training data, stratified five-fold cross-validation was employed to assess model stability and reduce variability associated with a single data split. During cross-validation, preprocessing steps were executed independently inside each fold, ensuring that validation samples were never used to estimate preprocessing parameters. This design simulates prospective application in which new patients are evaluated using

parameters learned from previously observed data. Variables used directly in clinical diagnosis of gestational diabetes were excluded from the predictive feature set. In particular, glucose measurements obtained during diagnostic testing (e.g., oral glucose tolerance testing) were not included among predictors. The outcome label was recorded independently in clinical records and was not deterministically derived from any predictor variable. Consequently, the modelling task represents early risk prediction rather than reconstruction of diagnostic criteria.

Model training and evaluation were conducted using consistent data partitions and fixed random seeds to ensure reproducibility and fair comparison across modelling approaches. Hyperparameters were selected conservatively based on established clinical machine learning practices rather than extensive tuning on the test set, thereby reducing the risk of optimistic bias.

Performance was assessed using metrics appropriate for imbalanced clinical outcomes. The area under the receiver operating characteristic curve (ROC AUC) and the area under the precision–recall curve (PR AUC) were used as primary indicators of discrimination. Secondary metrics included accuracy, recall, precision, F1-score, and confusion matrices to provide clinically interpretable evaluation of screening behaviour. Particular emphasis was placed on recall and PR AUC, reflecting the clinical priority of identifying high-risk pregnancies while limiting missed cases. This experimental design provides a realistic approximation of prospective clinical deployment, where model parameters are learned from historical data and applied to new patients. By combining stratified splitting, cross-validation, and strict separation between training and evaluation data, the evaluation framework minimises risks of overfitting, leakage, and inflated performance estimates.

3.7 Evaluation metrics

Model performance was evaluated using a set of complementary metrics selected to reflect both statistical discrimination and clinical screening priorities in gestational diabetes mellitus (GDM) risk prediction. Because the outcome exhibits moderate class imbalance and the clinical cost of missed cases is high, evaluation emphasised metrics sensitive to minority-class detection rather than overall accuracy alone. The area under the receiver operating characteristic curve (ROC AUC) was used as a primary metric to quantify each model's ability to discriminate between GDM and non-GDM cases across all possible decision thresholds. In addition, the area under the precision–recall curve (PR AUC) was reported to provide a more informative assessment of performance for the GDM class, particularly in the presence of class imbalance. PR AUC reflects the trade-off between precision and recall and is therefore well suited to screening applications where the positive class is of primary interest. Secondary performance measures included accuracy, precision, recall (sensitivity), and F1-score, calculated at a default probability threshold of 0.50 unless otherwise stated. Confusion matrices were generated to visualise classification outcomes and to examine the distribution of false positives and false negatives, supporting interpretation of clinical trade-offs between over-referral and missed diagnoses. All metrics were computed on the held-out test set to provide an unbiased estimate of generalisation performance. Where cross-validation was applied during model development, mean performance and variability across folds were reported for the training data to assess model stability. Collectively, these evaluation

measures provide a balanced and clinically meaningful assessment of predictive performance, supporting informed comparison of modelling approaches for antenatal GDM risk stratification.

3.8 Explainability and model interpretability

Model interpretability was incorporated as a core component of the analytical framework to support clinical transparency, trust, and meaningful use in antenatal screening contexts. Given the potential impact of automated risk stratification on clinical decision-making, explainability analyses were designed to clarify how predictive signals were learned and to ensure alignment with established clinical knowledge of gestational diabetes mellitus (GDM) risk factors. For models with intrinsic interpretability, such as logistic regression and decision tree–based learners, global explanations were obtained directly from model parameters. Logistic regression coefficients were examined to assess the direction and relative magnitude of association between individual predictors and GDM risk, enabling straightforward clinical interpretation of risk-increasing and risk-reducing factors. For tree-based models, feature importance measures were extracted to provide an overall ranking of predictors contributing to classification decisions. To capture non-linear and interaction effects not readily observable from linear coefficients, SHapley Additive exPlanations (SHAP) were employed for the gradient-boosted tree model. SHAP values were computed to quantify each feature's contribution to individual predictions and to summarise global importance patterns across the dataset. Global SHAP summary plots were used to visualise how predictor values influenced risk estimates across the population, while local explanations were examined for representative high-risk and low-risk cases to illustrate patient-level decision logic. Explainability analyses were performed exclusively on models trained without diagnostic or label-defining variables, ensuring that interpretations reflected genuine risk learning rather than reconstruction of clinical rules. All explanatory outputs were evaluated for clinical plausibility, with particular attention to established GDM risk factors such as maternal age, anthropometric measures, blood pressure, and haematological indicators. This approach ensured that interpretability findings complemented predictive performance and supported the potential translation of the models into clinically acceptable decision-support tools.

3.9 Ethical considerations

This study was conducted in accordance with established ethical principles governing biomedical research and the secondary use of clinical data. Ethical approval was obtained from the relevant institutional review bodies prior to data access and analysis. The dataset used in this study comprised retrospective, fully de-identified antenatal and laboratory records, and no direct contact with patients occurred at any stage of the research. All personal identifiers, including patient names, registration numbers, and national identification details, were removed prior to data transfer to the research team. As a result, individual participants could not be identified directly or indirectly, and the analysis posed minimal risk to patient privacy and welfare. In line with institutional and national research ethics guidelines, the requirement for individual informed consent was waived because the study involved secondary analysis of anonymised routine clinical data and did not influence patient care. Data handling and storage procedures complied

with applicable data protection and privacy regulations, including Uganda's Data Protection and Privacy Act (2019). Access to the dataset was restricted to authorised researchers, and all analyses were conducted on secure computing platforms. The predictive models developed in this study are intended to support clinical decision-making rather than replace clinician judgement, and outputs are framed as risk stratification tools for early screening rather than definitive diagnostic determinations. By adhering to these ethical safeguards, the study ensures responsible use of clinical data while contributing evidence toward the development of equitable, data-driven decision-support systems for maternal health in low-resource settings.

3.10 Architecture of the proposed GDM risk prediction framework

The proposed gestational diabetes mellitus (GDM) risk prediction framework follows a structured, end-to-end machine learning architecture designed to support early antenatal screening using routinely collected clinical data. As illustrated in Fig. 1, the framework integrates data ingestion, preprocessing, model learning, validation, and explainability into a unified and reproducible pipeline, ensuring methodological rigor, transparency, and clinical relevance. At the data input stage, de-identified patient-level antenatal records are provided as structured tabular data. Each record represents a unique pregnancy episode and contains maternal demographic attributes, anthropometric measurements, vital signs, and basic laboratory indicators available during routine antenatal visits. Administrative identifiers and non-informative fields are excluded at this stage to prevent noise and unintended information leakage, ensuring that only clinically meaningful variables are propagated through the pipeline. The preprocessing layer, shown in the second block of Fig. 1, transforms raw clinical inputs into a machine-learning–ready representation. This layer includes feature filtering based on missingness thresholds, explicit removal of diagnostic or post-diagnostic variables, and separation of predictors from the outcome label. Missing values are handled using median or mode imputation learned exclusively from training data, categorical variables are encoded into binary indicators, and feature scaling is applied selectively for models sensitive to variable magnitude. All preprocessing operations are embedded within pipeline workflows to ensure reproducibility and leakage-free execution. Following preprocessing, the data flow enters the model learning layer, where multiple supervised learning algorithms are trained in parallel. A logistic regression model provides a transparent baseline for clinical interpretation, while tree-based models, including Random Forest and Extreme Gradient Boosting, capture non-linear interactions and threshold effects among maternal risk factors. Where applicable, probabilistic outputs from individual models are

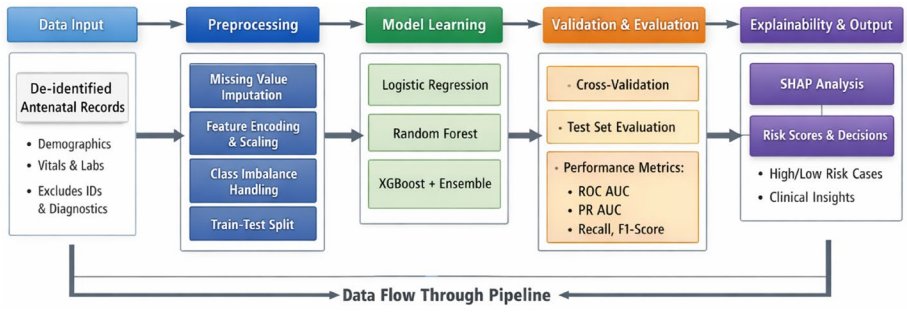

Fig. 1 Architecture of the proposed gestational diabetes mellitus (GDM) risk prediction framework

aggregated to form ensemble predictions, allowing the framework to assess whether complementary modelling strategies yield improved discrimination without sacrificing interpretability. The validation and evaluation layer, depicted in Fig. 1, implements stratified train–test splitting and stratified cross-validation to ensure robust estimation of generalisation performance. Models generate probability scores representing estimated GDM risk, which are evaluated using discrimination metrics such as the area under the receiver operating characteristic curve (ROC AUC) and the area under the precision–recall curve (PR AUC), alongside clinically interpretable measures including recall, precision, and confusion matrices. This layer ensures that predictive performance reflects realistic screening conditions rather than optimistic or leakage-driven estimates. To support transparency and clinical trust, the architecture incorporates a dedicated explainability and output layer. Global explanations are derived from model coefficients and feature-importance rankings, while local and non-linear effects are explored using SHapley Additive exPlanations (SHAP). The final outputs consist of calibrated GDM risk scores and binary risk classifications, enabling identification of high-risk and low-risk pregnancies for early antenatal triage and referral.

4 Experimental results

4.1 Descriptive characteristics of the study population

This section summarises the baseline demographic, anthropometric, and clinical characteristics of the antenatal cohort used for model development and evaluation. A total of 3525 antenatal records were included in the analysis, comprising 2153 women without gestational diabetes mellitus (Non-GDM) and 1372 women diagnosed with gestational diabetes mellitus (GDM). The relatively high proportion of GDM cases (38.9%) observed in this cohort should be interpreted in the context of the study setting. The data were obtained from Kawempe National Referral Hospital, a specialised obstetric referral centre that receives patients transferred from primary and secondary healthcare facilities, including women with suspected complications or high-risk pregnancies. Consequently, the study population represents a clinically enriched cohort rather than a general antenatal screening population. The reported prevalence therefore reflects characteristics of the referral-care population and should not be interpreted as population-level prevalence in Uganda. As shown in Table 1, women in the GDM group were older on average than those in the Non-GDM group (36.0 ± 6.0 years vs. 30.4 ± 5.2 years), indicating a clear age gradient associated with GDM risk. Body mass index (BMI) also differed substantially between groups, with women diagnosed with GDM exhibiting higher mean BMI values (31.1 ± 4.5 kg/m^2) compared with their Non-GDM counterparts (23.7 ± 4.2 kg/m^2). These findings are consistent with established epidemiological associations between advancing maternal age, increased adiposity, and gestational diabetes. Cardiovascular indicators demonstrated similar patterns. Mean systolic blood

Table 1 Baseline characteristics of the study population by gestational diabetes status

Variable	Overall ($N = 3525$)	Non-GDM ($n = 2153$)	GDM ($n = 1372$)
Age (years), mean ± SD	32.6 ± 6.2	30.4 ± 5.2	36.0 ± 6.0
Body mass index (kg/m^2), mean ± SD	27.9 ± 5.7	23.7 ± 4.2	31.1 ± 4.5
Systolic blood pressure (mmHg), mean ± SD	135.8 ± 22.7	121.4 ± 18.6	144.8 ± 20.4
Diastolic blood pressure (mmHg), mean ± SD	81.5 ± 11.4	76.3 ± 8.1	89.8 ± 10.8
Haemoglobin (g/dL), mean ± SD	14.0 ± 1.9	13.2 ± 1.5	15.1 ± 1.8
HDL cholesterol (mg/dL), mean ± SD	46.5 ± 10.8	49.7 ± 7.0	30.6 ± 12.3

pressure was higher among women with GDM (144.8 ± 20.4 mmHg) relative to the Non-GDM group (121.4 ± 18.6 mmHg), alongside elevated diastolic blood pressure values (89.8 ± 10.8 mmHg vs. 76.3 ± 8.1 mmHg). These differences suggest greater cardiometabolic burden among women diagnosed with GDM. Haematological and lipid profiles also varied across outcome groups. Women with GDM exhibited higher mean haemoglobin concentrations (15.1 ± 1.8 g/dL) than those without GDM (13.2 ± 1.5 g/dL). In contrast, mean high-density lipoprotein (HDL) cholesterol levels were markedly lower in the GDM group (30.6 ± 12.3 mg/dL) compared with the Non-GDM group (49.7 ± 7.0 mg/dL), reflecting dyslipidaemic patterns commonly associated with insulin resistance and impaired glucose metabolism. The multiple metabolic and cardiovascular indicators differed consistently between the two groups, suggesting substantial separation in cardiometabolic risk profiles within the studied cohort. Such multivariable clinical differences support the feasibility of early risk stratification using routinely collected antenatal variables while also indicating that model performance should be interpreted within the clinical context of a referral-care population.

4.2 Model performance comparison

This section compares the predictive performance of the evaluated machine learning models for gestational diabetes mellitus (GDM) risk prediction using a held-out test set. Performance metrics are summarised in Table 2 and include measures of discrimination (ROC AUC and PR AUC) and classification behaviour (accuracy, precision, recall, and F1-score), selected to reflect clinical screening requirements. Among the evaluated models, the tree-based methods demonstrated the strongest discrimination. Random Forest and XGBoost each achieved a ROC AUC of 0.997, indicating excellent ranking of GDM and non-GDM cases within the study cohort. Precision–recall performance was similarly high, with PR AUC values of 0.996 for Random Forest and 0.995 for XGBoost, supporting strong detection of the positive class under class imbalance. At the classification level, the Random Forest model achieved an accuracy of 0.971, recall of 0.994, and an F1-score of 0.963. The XGBoost model exhibited comparable performance, with an accuracy of 0.965, precision of 0.948, recall of 0.962, and an F1-score of 0.955. Compared with Random Forest, XGBoost demonstrated a slightly more conservative classification profile, characterised by higher precision and marginally lower recall. The ensemble model, constructed using soft-voting aggregation of base learners, achieved a ROC AUC of 0.996 and a PR AUC of 0.994, with an overall accuracy of 0.971. Precision (0.942) and recall (0.985) values were similar to those of the Random Forest model. However, the ensemble did not demonstrate a consistent improvement over the strongest individual learner, indicating that aggregation provided limited additional benefit in this dataset. This suggests that the primary predictive signal was already effectively captured by a single well-calibrated tree-based model. Logistic regression, used as an interpretable baseline, achieved a ROC AUC of 0.982 and a PR AUC of 0.965, with an accuracy of 0.961.

Table 2 Predictive performance of machine learning models on the test set

Model	ROC AUC	PR AUC	Accuracy	Precision	Recall	F1-score
Random Forest	0.997	0.996	0.971	0.934	0.994	0.963
XGBoost	0.997	0.995	0.965	0.948	0.962	0.955
Ensemble model	0.996	0.994	0.971	0.942	0.985	0.963
Logistic regression	0.982	0.965	0.961	0.938	0.965	0.951

Although discrimination was slightly lower than that of the non-linear models, logistic regression still demonstrated strong predictive capability, indicating that routinely collected antenatal variables contain meaningful predictive information even under linear modelling assumptions. All evaluated models achieved high internal discrimination. However, these results should be interpreted as performance within the studied cohort rather than as guaranteed performance in broader populations. The findings primarily demonstrate the feasibility of early GDM risk stratification using routinely collected antenatal variables, while external validation will be required to determine generalisable predictive accuracy.

4.3 Receiver operating characteristic and precision-recall curve analysis

To further evaluate discrimination beyond summary metrics, receiver operating characteristic (ROC) and precision–recall (PR) curve analyses were conducted for all evaluated models. The resulting curves are presented in Fig. 2, with Fig. 2a showing ROC curves and Fig. 2b showing PR curves for logistic regression, Random Forest, XGBoost, and the ensemble model. As illustrated in Fig. 2a, all models demonstrated strong discriminatory ability, with ROC curves approaching the upper-left region of the plot. Logistic regression achieved a ROC AUC of 0.982, indicating high but comparatively lower discrimination relative to the tree-based models. Random Forest and XGBoost achieved ROC AUC values of 0.997, while the ensemble model achieved 0.996. The substantial overlap among the Random Forest, XGBoost, and ensemble curves suggests similar ranking performance, indicating that combining models did not meaningfully improve separability beyond that achieved by a single tree-based learner. Precision–recall curves, shown in Fig. 2b, provide additional insight into model behaviour under class imbalance, particularly for the detection of GDM cases. Logistic regression achieved an average precision of 0.965, whereas Random Forest and XGBoost achieved higher average precision values of 0.996 and 0.995, respectively. The ensemble model achieved an average precision of 0.994, again closely matching the best individual model. Across most recall levels, tree-based models maintained higher precision than logistic regression, particularly in the clinically relevant moderate-to-high recall range. From a screening perspective, the PR curve is particularly informative because it reflects the trade-off between identifying high-risk pregnancies and generating unnecessary follow-up testing. The curves indicate that tree-based models maintain high precision even at elevated recall levels, supporting their suitability for triage-oriented risk stratification. However, these curves represent

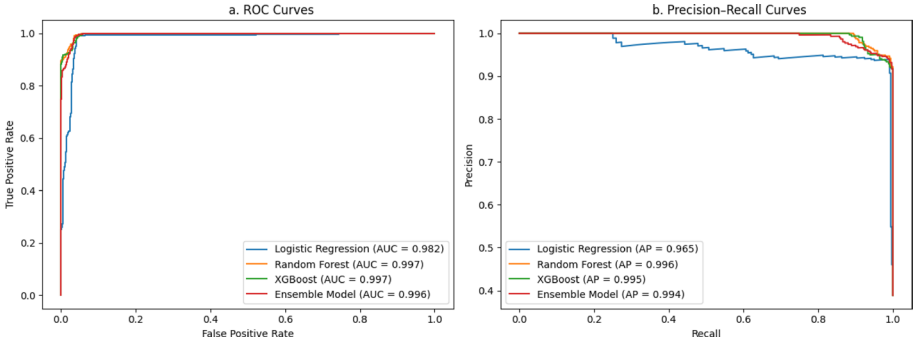

Fig. 2 Receiver operating characteristic (ROC) and precision-recall (PR) curves for gestational diabetes mellitus risk prediction models

internal evaluation within the study cohort and may differ in populations with different prevalence or risk-factor distributions. The curve analysis confirms that routinely collected antenatal variables contain sufficient predictive information to distinguish higher-risk from lower-risk pregnancies in this dataset. The findings should be interpreted as evidence of discriminatory signal within the cohort rather than proof of universal diagnostic accuracy, and external validation will be necessary to assess performance stability across clinical settings.

4.4 Confusion matrix analysis and screening trade-offs

To translate model performance into clinically interpretable outcomes, confusion matrix analyses were conducted for all evaluated models using the held-out test set. The resulting matrices are presented in Fig. 3, illustrating true and predicted classifications for gestational diabetes mellitus (GDM) and non-GDM cases across logistic regression, Random Forest, XGBoost, and the ensemble model. As shown in Fig. 3, the logistic regression model correctly classified 517 non-GDM and 331 GDM cases, while misclassifying 22 non-GDM cases as GDM (false positives) and 12 GDM cases as non-GDM (false negatives). This corresponds to a recall of 0.965, indicating that approximately 3.5% of GDM cases were missed at the operating threshold. Although this level of sensitivity is acceptable for baseline screening, the presence of false negatives highlights the limitations of linear decision boundaries for complex metabolic risk patterns. The Random Forest model demonstrated improved sensitivity, correctly identifying 341 GDM

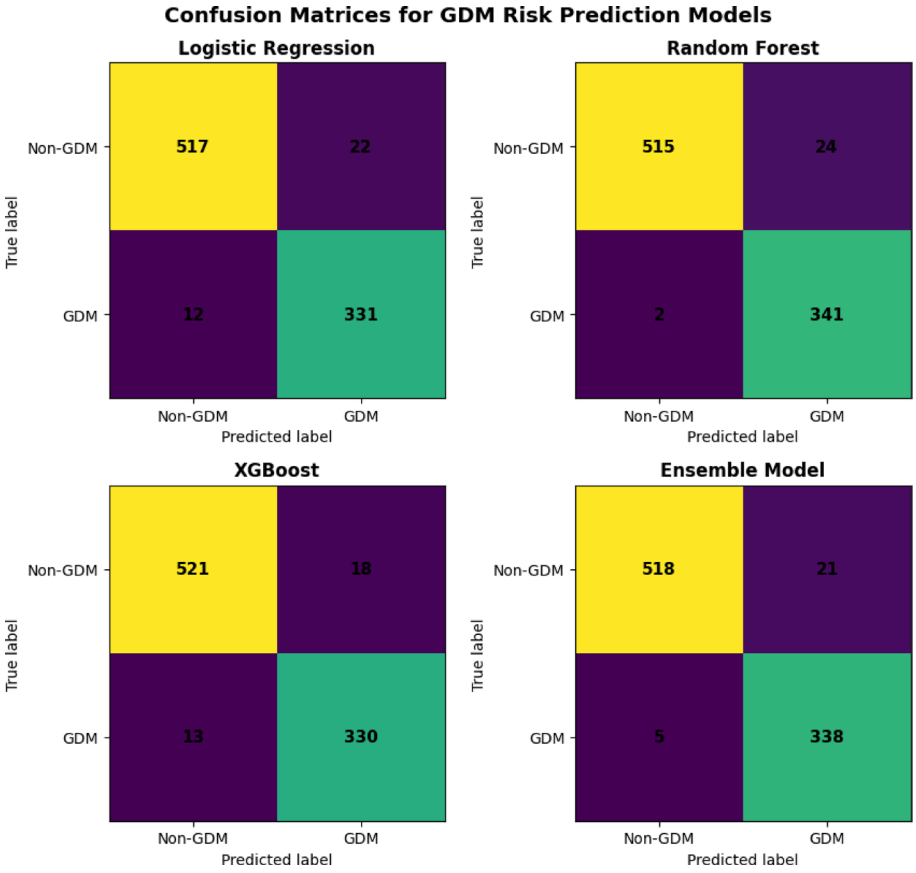

Fig. 3 Confusion matrices for gestational diabetes mellitus risk prediction models on the held-out test set

cases with only 2 false negatives, corresponding to a recall of 0.994. False positives were limited to 24 cases, resulting in a favourable balance between sensitivity and specificity. This behaviour is particularly advantageous for antenatal screening, where minimising missed high-risk pregnancies is prioritised over marginal increases in false-positive referrals. Similarly, the XGBoost model correctly classified 330 GDM cases with 13 false negatives, yielding a recall of 0.962. Compared with Random Forest, XGBoost exhibited slightly higher specificity, correctly identifying 521 non-GDM cases with 18 false positives, indicating a more conservative classification profile characterised by higher precision at the expense of modestly reduced sensitivity. The ensemble model achieved a strong balance between sensitivity and specificity, correctly identifying 338 GDM cases with 5 false negatives, corresponding to a recall of 0.985. False positives were limited to 21 cases, producing a confusion pattern closely aligned with that of the Random Forest model. Importantly, the ensemble did not demonstrate a substantial reduction in false negatives relative to Random Forest, reinforcing the conclusion that aggregation provided limited incremental benefit in this dataset.

4.5 Model explainability and feature contribution analysis

To enhance transparency and clinical interpretability, model explainability was assessed using Shapley Additive Explanations (SHAP) applied to the best-performing nonlinear model. Global feature importance and directionality of feature effects are illustrated in Fig. 4, which comprises a ranked global importance plot and a SHAP summary (beeswarm) plot. As shown in Fig. 4a, body mass index (BMI) emerged as the most influential predictor of gestational diabetes risk, exhibiting the highest mean absolute SHAP value. This indicates that variations in BMI contributed the largest average change in model output across the test population. High-density lipoprotein (HDL) cholesterol and systolic blood pressure were the next most influential variables, followed by prediabetes status and diastolic blood pressure, collectively highlighting the dominant role of metabolic and cardiovascular factors in GDM risk stratification. The SHAP summary plot in Fig. 4b provides further insight into the directionality of these effects. Higher BMI values were consistently associated with positive SHAP values, indicating increased predicted risk of GDM, whereas lower BMI values contributed negatively to risk predictions. In contrast, higher HDL levels were associated with negative SHAP values, reflecting a protective effect, while lower HDL values increased predicted GDM risk. Elevated systolic and diastolic blood pressure values similarly contributed positively to risk predictions, reinforcing their role as markers of underlying cardiometabolic stress. Binary clinical history variables, including prediabetes, previous gestational diabetes, and polycystic ovary syndrome (PCOS), exhibited clear directional effects, with positive feature values consistently shifting predictions toward higher GDM risk. Although maternal age was ranked lower than several metabolic indicators in terms of mean SHAP magnitude, higher age values still demonstrated a consistent positive contribution to risk, suggesting that age exerts an additive rather than dominant influence when combined with other risk factors. Several lifestyle and obstetric history variables such as family history of diabetes, sedentary lifestyle, number of pregnancies, history of large infant or birth defect, and unexplained prenatal loss exhibited smaller mean SHAP values. This indicates that, while these factors contribute to individual risk profiles, their influence is comparatively modest when evaluated alongside stronger metabolic predictors.

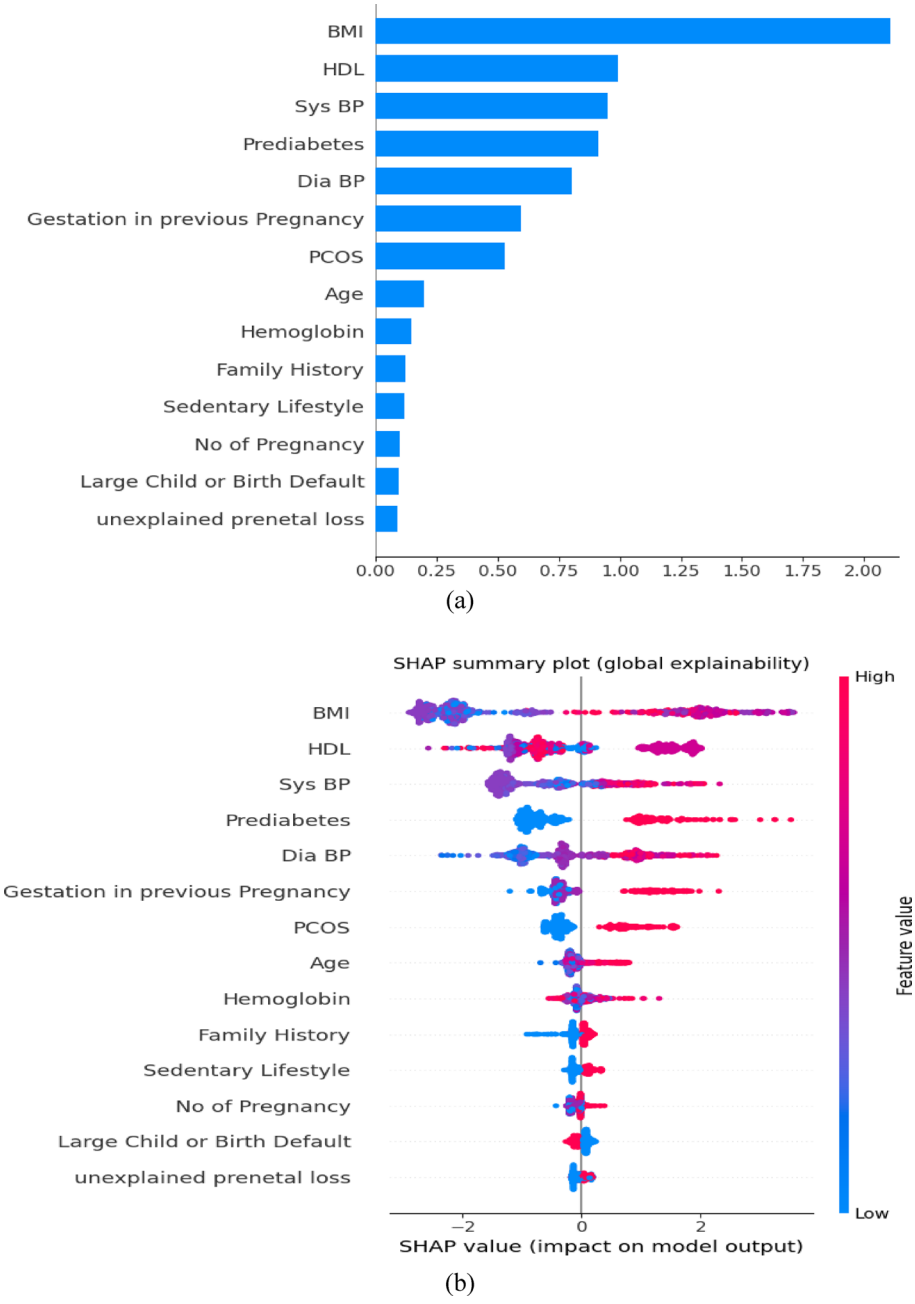

Fig. 4 **a** Global feature importance ranked by mean absolute SHAP value, indicating the average magnitude of each feature's contribution to model output across the test set. **b** SHAP summary (beeswarm) plot showing the distribution and direction of feature effects on predicted GDM risk

5 Discussion

This study evaluated the feasibility of machine learning–based gestational diabetes mellitus (GDM) risk prediction using routinely collected antenatal data in a low-resource setting. By combining descriptive cohort analysis, multi-metric model evaluation, and post-hoc explainability, the results provide a comprehensive assessment of both predictive performance and clinical interpretability.

5.1 Clinical relevance of baseline characteristics

The descriptive analysis revealed clear and clinically plausible differences between women diagnosed with GDM and those without GDM. Women in the GDM group were, on average, older and had substantially higher body mass index (BMI), systolic blood pressure, and diastolic blood pressure values. These findings align closely with established epidemiological evidence identifying advancing maternal age, increased adiposity, and cardiometabolic stress as key risk factors for gestational diabetes [45, 46]. The markedly lower HDL cholesterol levels observed among women with GDM further reflect dyslipidaemic profiles associated with insulin resistance and impaired glucose metabolism [47]. Importantly, despite these group-level differences, substantial overlap existed across all measured variables, underscoring the limitations of univariate or threshold-based screening approaches. This overlap highlights the necessity of multivariable models capable of capturing complex, non-linear interactions among risk factors, an observation that directly motivates the application of machine learning methods in this context.

An additional factor that may contribute to strong discrimination is the degree of separation between cardiometabolic profiles in the cohort. Women diagnosed with GDM in this dataset exhibited substantially higher body mass index and blood pressure values and markedly lower HDL cholesterol levels compared with the non-GDM group. When multiple correlated metabolic indicators differ consistently between outcome groups, machine learning models may identify clear multivariable patterns, leading to high discrimination without requiring deterministic diagnostic features. Therefore, the observed performance likely reflects strong underlying clinical signal within the studied population rather than reconstruction of diagnostic criteria. Nevertheless, this separation may be less pronounced in general antenatal populations, and model performance may accordingly decrease in broader screening settings.

5.2 Prevalence and cohort characteristics

The observed gestational diabetes prevalence in this cohort (38.9%) is higher than commonly reported population estimates of approximately 5–18% [3]. This difference is likely attributable to the referral nature of the study site. National referral hospitals typically manage more complex pregnancies and receive patients transferred due to clinical suspicion of complications, including metabolic disorders. As a result, the cohort represents a clinically enriched population rather than a population-based screening sample. Higher prevalence increases the separability between outcome groups and may contribute to stronger internal discrimination by machine learning models. However, this also implies that predictive performance may differ in primary care antenatal clinics where metabolic risk factors are less concentrated. Therefore, the findings should be interpreted as evidence of predictive feasibility within a referral-care setting rather than direct estimates of performance in general antenatal populations.

5.3 Predictive performance and screening suitability

Across the evaluated models, discrimination performance was consistently high, with Random Forest and XGBoost achieving ROC AUC values of 0.997 and precision–recall AUC values exceeding 0.995. Logistic regression demonstrated slightly lower but still strong discrimination (ROC AUC = 0.982). These results indicate that routinely collected

antenatal variables contain substantial predictive information related to gestational diabetes risk within the study cohort. However, predictive performance should be interpreted in the context of clinical screening rather than diagnosis. The objective of the proposed model is not to confirm gestational diabetes mellitus (GDM), which remains dependent on established diagnostic testing, but to identify pregnancies that may benefit from earlier confirmatory testing and closer follow-up. In this context, sensitivity is particularly important because missed high-risk pregnancies may delay intervention. The Random Forest model demonstrated the highest sensitivity, missing only two GDM cases in the test set (recall = 0.994), suggesting potential usefulness as a triage support tool. XGBoost exhibited a slightly more conservative classification profile, producing fewer false positives but at the cost of modestly lower sensitivity. The confusion matrix analysis highlights the clinical trade-off inherent in screening systems. Increasing sensitivity reduces missed cases but increases false-positive referrals, which may lead to additional laboratory testing and clinic visits. In antenatal care settings, such a trade-off may be acceptable because confirmatory testing (e.g., oral glucose tolerance testing) provides definitive diagnosis, and the clinical consequences of missed GDM are greater than those of unnecessary follow-up. Therefore, the observed classification behaviour is compatible with a pre-screening application rather than a standalone diagnostic tool. The ensemble model did not consistently outperform the strongest individual learner, indicating that performance gains from aggregation were limited in this dataset. This suggests that the predictive signal present in the available variables is captured effectively by a single well-regularised non-linear model. Logistic regression also demonstrated robust performance, implying that part of the risk signal is approximately linear, although non-linear models better captured complex interactions among metabolic and cardiovascular indicators. Importantly, these results reflect internal validation within a single cohort and should not be interpreted as evidence of universal performance. Clinical datasets often vary in prevalence, measurement practices, and patient characteristics, all of which can influence model behaviour. External validation across additional healthcare facilities and prospective evaluation within clinical workflows are therefore necessary before deployment. The findings instead support feasibility: routine antenatal data can be used to stratify risk and prioritise diagnostic testing in resource-constrained settings, where universal early OGTT screening is difficult to implement.

5.4 Robustness and risk of overfitting

The exceptionally high discrimination observed in internal evaluation warrants careful methodological consideration. In clinical prediction research, near-ceiling performance may arise from information leakage, overly optimistic validation procedures, or dataset-specific patterns. To mitigate these risks, several safeguards were implemented. Predictor variables were restricted to measurements plausibly available prior to diagnostic testing, and variables used to establish GDM diagnosis were excluded from the feature set. All preprocessing steps, including imputation and scaling, were performed within training pipelines after train-test splitting to prevent information from the evaluation set influencing model training. In addition, stratified cross-validation was applied within the training data, and a completely held-out test set was reserved for final evaluation. Hyperparameters were selected conservatively rather than exhaustively tuned to avoid fitting to noise. Despite these precautions, internal validation alone cannot exclude

the possibility that performance partly reflects characteristics of the study population, including referral-hospital case mix and class distribution. Therefore, the results should be interpreted as evidence of feasibility rather than definitive predictive accuracy. External validation across independent healthcare facilities, or temporal validation using later patient cohorts, will be necessary to determine whether similar discrimination can be achieved in broader antenatal populations and under routine clinical conditions.

5.5 Generalisability and population transportability

The present study utilised data from a single national referral hospital. Referral facilities typically receive patients with more complex clinical profiles, including higher prevalence of metabolic risk factors and obstetric complications. As a result, the distribution of predictors and outcome prevalence in this cohort may differ from those observed in primary or community-level antenatal clinics. Machine learning models are sensitive to population shift, and predictive relationships learned in one clinical environment may not remain stable in another healthcare setting. Differences in patient demographics, nutritional status, screening practices, and healthcare access can influence both predictor distributions and outcome definitions. Therefore, the reported performance should not be interpreted as guaranteed performance in all Ugandan antenatal settings. Instead, the findings demonstrate that routine antenatal variables contain sufficient information to support early risk stratification within the studied clinical context. Future work should include external validation in additional hospitals, rural health centres, and prospective cohorts to evaluate model transportability and calibration across different screening environments.

5.6 Expected performance in general antenatal populations

Because the study cohort represents a referral-care population, predictive performance may differ in lower-risk antenatal settings. Clinical prediction models are influenced by the spectrum of disease severity and risk-factor distribution within the population in which they are developed. When outcome groups exhibit clearer separation of metabolic characteristics, discrimination metrics such as ROC AUC may appear higher. In community antenatal clinics, where risk factors are less concentrated and patients are more heterogeneous, the distinction between GDM and non-GDM cases may be subtler. Therefore, model performance in primary care or rural antenatal clinics may be lower than observed in this study. However, this does not negate clinical usefulness. In screening applications, even moderate discrimination can meaningfully support triage decisions and prioritisation for confirmatory testing. The present findings demonstrate that routinely collected antenatal variables contain predictive information; future external validation will determine the degree of performance change across different healthcare levels.

5.7 Comparison with clinical machine learning studies

The predictive performance observed in this study appears higher than that reported in many clinical machine learning prediction studies and therefore warrants careful interpretation. Multi-algorithm benchmarking investigations in healthcare datasets typically report ROC AUC values ranging from approximately 0.75 to 0.95 depending on dataset characteristics, outcome definition, and feature availability [24, 48, 49]. Similar findings

have been reported in comparative studies evaluating multiple machine learning algorithms across clinical prediction tasks, where tree-based ensemble methods often outperform linear models but rarely achieve near-perfect discrimination.

Several recent medical machine learning benchmarking studies have shown that algorithm choice influences performance less than dataset characteristics, feature quality, and population structure [50, 51]. In many healthcare datasets, Random Forest and gradient boosting approaches consistently demonstrate strong performance due to their ability to model non-linear interactions and threshold effects among physiological variables. The present findings follow this general pattern, with tree-based models outperforming logistic regression while maintaining comparable interpretability through feature attribution.

The magnitude of performance observed in this study should therefore not be attributed solely to algorithmic superiority. Instead, it likely reflects properties of the available predictors and the clinical cohort. The dataset contains multiple cardiometabolic indicators (e.g., BMI, blood pressure, and lipid measures) that are physiologically related to insulin resistance and may provide substantial discriminative signal when measured prior to diagnosis. In addition, the data originate from a referral-hospital population, which may include a higher proportion of high-risk pregnancies and clearer separation between outcome groups than community-level antenatal populations.

For these reasons, the results should be interpreted as demonstrating feasibility rather than universal predictive accuracy. The study shows that routinely collected antenatal variables can support meaningful early risk stratification when analysed using appropriate machine learning methods, but performance estimates are cohort-dependent. External validation across multiple healthcare facilities and prospective evaluation are necessary to determine generalisability under varying prevalence, screening protocols, and demographic conditions.

5.8 Explainability and clinical interpretability

The SHAP-based explainability analysis provides strong evidence that model predictions are driven by clinically meaningful factors rather than spurious correlations. BMI emerged as the most influential predictor, consistent with its central role in insulin resistance and GDM pathophysiology [46, 52]. HDL cholesterol and blood pressure metrics followed in importance, reinforcing the contribution of cardiometabolic health to GDM risk. Importantly, higher HDL levels exhibited a protective effect, while elevated BMI and blood pressure values consistently increased predicted risk. Binary clinical history variables such as prediabetes, previous gestational diabetes, and polycystic ovary syndrome (PCOS) demonstrated clear directional effects, aligning with established clinical risk stratification frameworks [53]. Maternal age, although not the dominant predictor, contributed additively to risk predictions, suggesting that its influence is mediated through interactions with metabolic factors rather than acting in isolation. The relatively lower importance of lifestyle and obstetric history variables does not diminish their clinical relevance; rather, it indicates that their predictive contribution is context-dependent and secondary to stronger metabolic indicators in this dataset. Overall, the explainability results enhance trust in the model by demonstrating alignment with existing clinical knowledge, a critical requirement for adoption in healthcare settings [54].

The variable labelled prediabetes status represents previously documented medical history recorded in patient charts prior to or at the first antenatal assessment rather than glucose testing performed during the current pregnancy. It therefore reflects baseline metabolic susceptibility rather than a laboratory measurement used to establish gestational diabetes diagnosis. Its importance in the model is clinically plausible because women with prior impaired glucose regulation are known to have elevated risk of developing GDM in subsequent pregnancies.

Importantly, all predictors included in the explainability analysis were obtained from information available before diagnostic glucose testing for gestational diabetes. The model therefore identifies early metabolic risk patterns rather than reconstructing diagnostic criteria. This temporal separation between predictors and outcome supports the validity of the early risk prediction objective and reduces the likelihood that the observed feature importance is driven by information leakage.

5.9 Implications for low-resource antenatal care

From a practical perspective, the findings suggest that machine learning models particularly Random Forest and XGBoost can support early GDM risk stratification using data already collected in routine antenatal care. The strong sensitivity achieved by these models is well-suited for screening workflows, where early identification enables targeted diagnostic testing and preventive interventions. Moreover, the use of explainable models addresses a key barrier to adoption in low-resource settings by facilitating clinician understanding and accountability. Nevertheless, while internal performance is strong, external validation across additional healthcare facilities and populations is required to assess generalisability. Future work should also explore temporal validation and prospective evaluation to determine the real-world impact of integrating such models into antenatal care pathways.

5.10 Clinical implementation in resource-constrained antenatal care

To understand the potential utility of the proposed model, it is necessary to consider the workflow of antenatal care in many Ugandan public health facilities. During the first antenatal visit, basic information such as maternal age, parity, weight, blood pressure, and medical history is routinely recorded. However, confirmatory laboratory investigations, particularly the oral glucose tolerance test (OGTT), are not always performed at the same visit. The OGTT requires patient fasting, laboratory reagents, trained personnel, and several hours of clinic attendance. In busy public facilities, this often leads to delayed scheduling or missed testing when patients cannot return for repeat visits. Consequently, clinicians frequently rely on clinical judgement to decide which women should undergo further metabolic evaluation. This selective screening approach may fail to identify asymptomatic high-risk pregnancies early in gestation. The proposed model is therefore intended to function as a pre-screening triage tool, not a diagnostic system. By estimating risk using routinely collected measurements, the model could help prioritize women for confirmatory testing, counselling, and closer follow-up. In practice, implementation would not require additional laboratory infrastructure. The predictor variables used in this study are already recorded during standard antenatal assessment and could be entered into a simple digital calculator integrated into a clinic computer, tablet, or mobile device. The model output would be a risk score indicating whether

early glucose testing or referral is advisable. Such decision support may be particularly useful in high-volume clinics where midwives must rapidly assess many patients. Importantly, the system is designed to assist rather than replace clinical judgement. Final decisions regarding diagnosis and management would remain with healthcare providers, and confirmatory testing using established clinical guidelines would still be required. The model therefore complements existing screening protocols by improving early identification rather than substituting standard care.

5.11 Operational feasibility and deployment considerations

For clinical adoption, predictive performance alone is insufficient; operational feasibility within routine antenatal workflows is essential. The models developed in this study are computationally lightweight and operate on structured tabular data rather than medical imaging or continuous monitoring signals. After training, prediction requires only a small number of arithmetic operations and decision tree traversals, allowing inference to be performed in real time on standard computing hardware. In practical deployment, model execution would not require dedicated servers or graphical processing units. The trained model can be exported as a serialized object and executed on a standard desktop computer, laptop, or tablet commonly used for electronic medical records. For each patient, the input consists of routinely recorded variables (e.g., age, parity, body mass index, blood pressure, and selected booking laboratory measurements). Prediction time is effectively instantaneous, enabling risk scoring during a clinical visit without disrupting workflow. Integration into antenatal care could be implemented through a simple digital risk calculator embedded within an electronic medical record system or a standalone mobile application. During the first antenatal visit, a midwife or clinician would enter the already-recorded measurements into the interface, and the system would return a risk estimate. High-risk patients could then be prioritised for confirmatory glucose testing, dietary counselling, or closer monitoring. The tool therefore functions as a triage decision-support system rather than an automated diagnostic mechanism. Threshold selection may be adjusted according to clinical priorities. In high-volume urban clinics where laboratory testing capacity exists, a lower threshold could be used to maximise sensitivity and minimise missed cases. In settings with limited follow-up testing capacity, a higher threshold may reduce unnecessary referrals while still identifying patients at greatest risk. Future prospective studies should evaluate threshold calibration, workflow integration, and user acceptance among midwives and clinicians.

6 Conclusion, limitations, and future work

This study investigated the feasibility of early gestational diabetes mellitus (GDM) risk prediction using routinely collected antenatal care data in a resource-constrained clinical setting. Using a cohort of 3525 pregnancies, machine learning models demonstrated strong discrimination between women later diagnosed with GDM and those without the condition. Tree-based methods, particularly Random Forest and XGBoost, achieved the highest performance, while logistic regression also showed robust predictive capability. These findings indicate that routinely recorded demographic, anthropometric, and basic clinical measurements contain meaningful information related to metabolic risk during pregnancy. Importantly, the proposed system is intended as a risk-stratification and triage support tool rather than a diagnostic model. Gestational diabetes diagnosis remains

dependent on established clinical testing, including oral glucose tolerance testing. The model instead estimates early risk using information available at or near the first antenatal visit, enabling prioritisation of confirmatory testing, counselling, and follow-up in settings where universal early screening is difficult to implement. In this context, high sensitivity is desirable because early identification may support preventive interventions and reduce adverse maternal and neonatal outcomes. The explainability analysis demonstrated that model predictions were driven by clinically plausible factors, including body mass index, blood pressure, lipid indicators, and prior metabolic history. The alignment between model explanations and established clinical knowledge supports the credibility of the predictions and may facilitate clinician acceptance. Rather than functioning as a black-box classifier, the model provides interpretable risk estimates that can complement clinical judgement.

Several limitations should be considered when interpreting the findings. First, the study used retrospective data from a single referral-hospital cohort. Referral populations may differ from community antenatal populations in disease prevalence and risk distribution, and this may influence performance estimates. Second, although strict leakage-prevention procedures were applied, internal validation cannot fully establish generalisability. Third, the analysis relied on variables routinely recorded during antenatal assessment; while this supports practicality, additional biomarkers or longitudinal measurements may further improve prediction accuracy. Finally, the study evaluated classification performance at a fixed threshold and did not optimise thresholds for specific clinical workflows or resource constraints.

Future work should therefore focus on external and prospective validation across multiple healthcare facilities and levels of care. Evaluating model performance in community clinics, district hospitals, and diverse patient populations will be necessary to determine real-world applicability. Implementation studies are also required to assess how risk-prediction tools influence clinician decision-making, screening uptake, and patient outcomes. Integration into digital antenatal care systems, accompanied by user-centred interface design and workflow testing, will be essential for safe deployment.

Acknowledgements
None.

Author contributions
Emmanuel Ahishakiye (EA) designed the study. EA wrote the initial manuscript. Justine Nakirijja (JN) and Shallon Ahimbisibwe (SA) edited and extensively reviewed the manuscript. The author approved the manuscript.

Funding
We acknowledge the funding from Kyambogo University 10th Competitive grant.

Data availability
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability
Code is available upon request from the corresponding author.

Declarations

Ethics approval and consent to participate
Ethical approval was obtained from the Kyambogo University Research Ethics Committee. All methods were performed in accordance with the WMA Declaration of Helsinki (2013), the CIOMS International Ethical Guidelines for Health-related Research Involving Humans (2016), and the UNCST National Guidelines for Research involving Humans as Research Participants (July 2014). The requirement for informed consent for study participation was waived by the Kyambogo University Research Ethics Committee because the research involved retrospective analysis of fully anonymised secondary data, posed minimal risk to participants, and involved no direct interaction with human subjects. For participants under 16 years of age, the ethics committee waiver similarly applied to the use of de-identified routine antenatal care records; therefore, parental or legally authorised representative informed consent was not required for this secondary analysis.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 January 2026 / Accepted: 23 April 2026
Published online: 10 May 2026

References

1. Wang H, et al. IDF Diabetes Atlas: estimation of global and regional gestational diabetes mellitus prevalence for 2021 by International Association of Diabetes in Pregnancy Study Group's Criteria. Diabetes Res Clin Pract. 2022. https://doi.org/10.1016/j.diabres.2021.109050.
2. Catalano PM. Obesity, insulin resistance and pregnancy outcome. Reproduction. 2010;140(3):365. https://doi.org/10.1530/REP-10-0088.
3. Nakshine VS, Jogdand SD. A comprehensive review of gestational diabetes mellitus: impacts on maternal health, fetal development, childhood outcomes, and long-term treatment strategies. Cureus. 2023. https://doi.org/10.7759/cureus.47500.
4. Zalwango F, Seeley J, Namara A, Kinra S, Nyirenda M, Oakley L. Diagnosis of gestational diabetes in Uganda: the reactions of women, family members and health workers. Women Health. 2021. https://doi.org/10.1177/17455065211013769.
5. Mwanri AW, Kinabo J, Ramaiya K, Feskens EJM. Gestational diabetes mellitus in sub-Saharan Africa: systematic review and metaregression on prevalence and risk factors. Trop Med Int Health. 2015;20(8):983–1002. https://doi.org/10.1111/TMI.12521.
6. Saham N, et al. Barriers to routine antenatal syphilis screening in Uganda: provider perspectives and practices. Glob Qual Nurs Res. 2025;12:23333936251375456. https://doi.org/10.1177/23333936251375457.
7. Jamieson EL, et al. Oral glucose tolerance test to diagnose gestational diabetes mellitus: impact of variations in specimen handling. Clin Biochem. 2023;115(S3):33–48. https://doi.org/10.1016/j.clinbiochem.2022.10.002.
8. Reddi Rani P, Begum J. Screening and diagnosis of gestational diabetes mellitus, where do we stand. J Clin Diagn Res. 2016;10(4):QE01. https://doi.org/10.7860/JCDR/2016/17588.7689.
9. Boutilier JJ, Chan TCY, Ranjan M, Deo S. Risk stratification for early detection of diabetes and hypertension in resource-limited settings: machine learning analysis. J Med Internet Res. 2021;23(1):e20123. https://doi.org/10.2196/20123.
10. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Summits Transl Sci Proc. 2020;2020:191.
11. AlSaad R, et al. Artificial intelligence in gestational diabetes care: a systematic review. J Diabetes Sci Technol. 2025. https://doi.org/10.1177/19322968251355967.
12. Cubillos G, et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregn Childb. 2023;23(1):1–18. https://doi.org/10.1186/S12884-023-05766-4/TABLES/3.
13. Ji Q, et al. Early prediction of gestational diabetes mellitus using machine learning-integrated metabolomic and clinical features. Front Endocrinol. 2025;16:1687146. https://doi.org/10.3389/fendo.2025.1687146.
14. Yang J, et al. Early prediction of gestational diabetes mellitus based on systematically selected multi-panel biomarkers and clinical accessibility—a longitudinal study of a multi-racial pregnant cohort. BMC Med. 2025;23(1):430. https://doi.org/10.1186/s12916-025-04258-w.
15. Tessema ZT, et al. Determinants of accessing healthcare in sub-Saharan Africa: a mixed-effect analysis of recent Demographic and Health Surveys from 36 countries. BMJ Open. 2022;12(1):e054397. https://doi.org/10.1136/bmjopen-2021-054397.
16. Mapari SA, et al. Revolutionizing maternal health: the role of artificial intelligence in enhancing care and accessibility. Cureus. 2024;16(9):e69555. https://doi.org/10.7759/cureus.69555.
17. Abdelwanis M, Simsekler MCE, Gabor AF, Sleptchenko A, Omar M. Artificial intelligence adoption challenges from healthcare providers' perspectives: a comprehensive review and future directions. Saf Sci. 2026;193(1):107028. https://doi.org/10.1016/j.ssci.2025.107028.
18. Kahimakazi I, et al. Prevalence of gestational diabetes mellitus and associated factors among women receiving antenatal care at a tertiary hospital in South-Western Uganda. Pan Afr Med J. 2023. https://doi.org/10.11604/pamj.2023.46.50.38355.
19. Global Prevalence of Gestational Diabetes (GDM) | IDF Atlas. Available: https://diabetesatlas.org/data-by-indicator/hyperglycaemia-in-pregnancy-hip-20-49-y/prevalence-of-gestational-diabetes-mellitus-gdm/. Accessed 11 Dec 2025.
20. International Diabetes Federation. IDF Diabetes Atlas. 2019. https://doi.org/10.1016/S0140-6736(55)92135-8.
21. Hinneh T, Jahn A, Agbozo F. Barriers to screening, diagnosis and management of hyperglycaemia in pregnancy in Africa: a systematic review. Int Health. 2021;14(3):211. https://doi.org/10.1093/INTHEALTH/IHAB054.
22. Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy. Available https://www.who.int/publications/i/item/WHO-NMH-MND-13.2. Accessed 11 Dec 2025
23. Habibi N, et al. Maternal metabolic factors and the association with gestational diabetes: a systematic review and meta-analysis. Diabetes Metab Res Rev. 2022;38(5):e3532. https://doi.org/10.1002/DMRR.3532.
24. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16. https://doi.org/10.1016/j.csbj.2016.12.005.
25. Sarno L, et al. Use of artificial intelligence in obstetrics: not quite ready for prime time. Am J Obstet Gynecol MFM. 2023;5(2):100792. https://doi.org/10.1016/J.AJOGMF.2022.100792.
26. Yeganegi M, et al. Research advancements in the use of artificial intelligence for prenatal diagnosis of neural tube defects. Front Pediatr. 2025;13:1514447. https://doi.org/10.3389/FPED.2025.1514447.
27. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. Cureus. 2023;15(10):e46454. https://doi.org/10.7759/CUREUS.46454.

28. García-Luque R, Pimentel E, Durán F, Aranda-Gallardo M, Morales-Asencio JM. A machine learning-based clinical prediction rule for adverse outcomes in multimorbid patients. Intell Med. 2025;5(4):300–9. https://doi.org/10.1016/j.imed.2025.08.006.
29. Noroozi Z, Orooji A, Erfannia L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Sci Rep. 2023;13(1):22588. https://doi.org/10.1038/s41598-023-49962-w.
30. Stenwig E, Rossi PS, Salvi G, Skjærvold NK. The impact of feature combinations on machine learning models for in-hospital mortality prediction. Sci Rep. 2025;15(1):39119. https://doi.org/10.1038/s41598-025-26611-y.
31. Song MH, Lee YH, Kang UG. Comparison of machine learning algorithms for classification of the sentences in three clinical practice guidelines. Healthc Inform Res. 2013;19(1):16. https://doi.org/10.4258/hir.2013.19.1.16.
32. Beunza JJ, et al. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). J Biomed Inform. 2019;97(3):103257. https://doi.org/10.1016/j.jbi.2019.103257.
33. Performance evaluation methods for evolving artificial intelligence (AI)-enabled medical devices | FDA. Available: https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/performance-evaluation-methods-evolving-artificial-intelligence-ai-enabled-medical-devices. Accessed 02 Mar 2026.
34. Park DJ, Baik SM, Hong KS, Yi H, Lee JG, Lee JM. Development and external validation of an artificial intelligence model for predicting mortality and prolonged ICU stay in postoperative critically ill patients: a retrospective study. World J Emerg Surg. 2025;20(1):79. https://doi.org/10.1186/s13017-025-00650-2.
35. Al-Nafjan A, Aljuhani A, Alshebel A, Alharbi A, Alshehri A. Artificial intelligence in predictive healthcare: a systematic review. J Clin Med. 2025. https://doi.org/10.3390/jcm14196752.
36. Khokhar PB, Gravino C, Palomba F. Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review. Artif Intell Med. 2025;164:103132. https://doi.org/10.1016/J.ARTMED.2025.103132.
37. El Falah Z, Marfak A, Asmaa MA, Thimou A. External generalizability and internal accuracy of predictive models for perinatal mortality: a systematic review and meta-analysis. Clin Epidemiol Glob Health. 2025;36:102227. https://doi.org/10.1016/j.cegh.2025.102227.
38. Lee S, et al. External validation of an artificial intelligence model using clinical variables, including ICD-10 codes, for predicting in-hospital mortality among trauma patients: a multicenter retrospective cohort study. Sci Rep. 2025;15(1):1100. https://doi.org/10.1038/s41598-025-85420-5.
39. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2015;2015:1721–30. https://doi.org/10.1145/2783258.2788613;PAGE:STRING:ARTICLE/CHAPTER.
40. Zhang R, Liu Y, Zhang Z, Luo R, Lv B. Interpretable machine learning model for predicting postpartum depression: retrospective study. JMIR Med Inform. 2025;13(1):e58649. https://doi.org/10.2196/58649.
41. Chng SY, et al. Ethical considerations in AI for child health and recommendations for child-centered medical AI. npj Digit Med. 2025;8(1):152. https://doi.org/10.1038/s41746-025-01541-1.
42. Digital Health | WHO | Regional Office for Africa. Available: https://www.afro.who.int/health-topics/digital-health. Accessed 11 Dec 2025.
43. Ddamba A, et al. Factors influencing the availability and use of electronic medical records systems in public health facilities in Uganda: a cross-sectional assessment. BMC Med Inform Decis Mak. 2025;25(1):372. https://doi.org/10.1186/S12911-025-03190-6.
44. Mennickent D, et al. Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications. Front Endocrinol. 2023;14:1130139. https://doi.org/10.3389/FENDO.2023.1130139.
45. Gestational Diabetes Mellitus. In Special Collection: ADA Medical Affairs Article Collection. American Diabetes Association. Diabetes Care 2004;27(suppl_1):s88–s90. https://doi.org/10.2337/DIACARE.27.2007.S88.
46. Kim C. Gestational diabetes: risks, management, and treatment options. Int J Womens Health. 2010;2(1):339. https://doi.org/10.2147/IJWH.S13333.
47. Preda A, et al. Dyslipidemia in pregnancy: a systematic review of molecular alterations and clinical implications. Biomedicines. 2024;12(10):2252. https://doi.org/10.3390/BIOMEDICINES12102252.
48. Artzi NS, et al. Prediction of gestational diabetes based on nationwide electronic health records. Nat Med. 2020;26(1):71–6. https://doi.org/10.1038/s41591-019-0724-8.
49. Hassan A, Ahmad SG, Iqbal T, Munir EU, Ayyub K, Ramzan N. Enhanced model for gestational diabetes mellitus prediction using a fusion technique of multiple algorithms with explainability. Int J Comput Intell Syst. 2025;18(1):1–33. https://doi.org/10.1007/S44196-025-00760-4/FIGURES/1.
50. Baqraf YKA, Keikhosrokiani P, Al-Rawashdeh M. Evaluating online health information quality using machine learning and deep learning: a systematic literature review. Digit Health. 2023;9:20552076231212296. https://doi.org/10.1177/20552076231212296.
51. Labory J, Njomgue-Fotso E, Bottini S. Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. Comput Struct Biotechnol J. 2024;23:1274–87. https://doi.org/10.1016/j.csbj.2024.03.016.
52. Catalano PM, Shankar K. Obesity and pregnancy: mechanisms of short term and long term adverse consequences for mother and child. BMJ. 2017. https://doi.org/10.1136/BMJ.J1.
53. Quaresima P, Myers SH, Pintaudi B, D'Anna R, Morelli M, Unfer V. Gestational diabetes mellitus and polycystic ovary syndrome, a position statement from EGOI-PCOS. Front Endocrinol. 2025;16:1501110. https://doi.org/10.3389/FENDO.2025.1501110.
54. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;2017-Decem(Section 2):4766–75.

Publisher's note