

Development of a Cyclic Gradient Boosting Model for Early Prediction of Gestational Diabetes Risk

^{1*}Obianozie Ifunanya Mirian, ²Oluwatobi Noah Akande, ³Prema Kirubakaran

^{1,3}Department of Information Technology, Faculty of Computing, Nile University of Nigeria, Abuja, Nigeria

²Computer Science Department, Nile University of Nigeria, Abuja, Nigeria

*Corresponding Author: anaekwe.mirian@gmail.com

ABSTRACT

Gradient Boosting Model is a notable public health challenge in Nigeria, especially in the low resource care facilities lacking adequate facilities for standard GDM diagnosis. This study attempts to develop a machine learning technique capable of accurately interpreting GDM prediction results through the use of clinical data. To accomplish this objective, a Cyclic Gradient Boosting (CGB) algorithm was proposed and analyzed using Shapley Additive exPlanations (SHAP). A total of 918 antenatal patient records were used in this investigation. Based on the results obtained from training and validation process, it was possible to obtain a good performance of the designed CGB model, with 82.6% accuracy and ROC-AUC equal to 0.822 and high specificity (>90%). In addition, it is possible to mention the high precision (62.9%), and the analysis of important features revealed that family history of diabetes (~0.60) and body mass index (BMI=0.42) were the most influential predictors. More detailed investigation of the SHAP values shows that the interaction between heredity and BMI is significant (interaction weight: 0.05), as well as the influence of heredity on weight (0.04). Cyclic boosting allowed obtaining consistent and balanced feature importances. In addition, SHAP helped interpret the model by providing easily comprehensible patient specific explanations. It was found that the combination of predictive power with interpretability is possible using non-invasive routinely collected data. This approach offers a practical solution for early identification of GDM risk in resource-constrained settings as well as support targeted screening and timely intervention.

KEYWORDS: Gestational Diabetes Mellitus (GDM), Explainable Machine Learning, Cyclic Gradient Boosting (CGB), SHapley Additive exPlanations (SHAP), Risk Prediction, Low-Resource Settings, Nigeria, Antenatal Care.

I. INTRODUCTION

GDM is a glucose intolerance disease which is diagnosed for the first time during pregnancy among women who are previously non-diabetic (Choudhury & Devi Rajeswari, 2021; World Health Organization, 2013). The incidence rate is not universal and fluctuates between 1% and 28%, according to different populations and criteria used (Muche et al., 2019). Nevertheless, the prevalence remains significantly higher among low and middle-income nations, in which approximately 80% of cases occur despite limited healthcare resources (Kanguru et al., 2014). In Nigeria, the situation is particularly complex. According to a systematic review, the overall prevalence rate was 11% out of more than 46,000 participants, but individual studies showed wide ranges, starting from 0.5% up to 55% (Azeez et al., 2021; Odoh et al., 2025). Nonetheless, despite the prevalence statistics provided above, GDM remains underdiagnosed. For example, the Oral Glucose Tolerance Test (OGTT), which is considered a gold standard in diagnosis, is carried out late or not done at all. Moreover, some studies revealed that most of the times, GDM

screening is based on risk-based approaches, and such screening methods cannot detect the disease in individuals without clinical risk factors such as obesity or older mothers (Azeez et al., 2021; Odoh et al., 2025). Late or improper screening for GDM poses significant health threats, such as an increased risk of pre-eclampsia, caesarean delivery, macrosomia, and metabolic abnormalities in children born to GDM mothers (Azeez et al., 2021). Risk scoring systems based on conventional statistics do not have the capability to identify complicated relationships between clinical variables. Meanwhile, many machine learning models are opaque and difficult to implement in clinical settings (Elfadel Magboul et al., 2025). However, recent advancements in explainable artificial intelligence provide a possible solution to these concerns. The use of cyclic gradient boosting (CGB), which is an ensemble learning model coupled with cyclic feature-updating techniques, has been found to possess predictive capabilities. Combined with Shapley Additive exPlanations (SHAP), the model provides the ability to make interpretable explanations at the level of individual patients (Alamadi et al., 2023; Lundberg & Lee, 2017). Applying such an approach to Nigerian antenatal data could address several critical gaps in many ways because it will be helpful in overcoming many challenges that include early detection of risks, better decision-making due to interpretability, and effective distribution of scarce testing resources (Kang et al., 2025; Li et al., 2025). This is particularly important in settings where the challenge is not only limited resource but also inconsistent screening practices. Against this backdrop, this study therefore, aims to develop a CGB model, supported by SHAP-based interpretation, for early prediction of GDM risk among Nigerian pregnant women.

II. RELATED WORKS

There have been several studies done in machine learning applications for GDM predictions where predictive and methodological advances have been considered. The use of artificial intelligence for diabetes predictions was examined by Awasthi et al. in 2025. In this regard, it is essential to note that the study of Awasthi et al. (2025) advanced development in model complexity but not in model interpretability. On the other hand, the studies of Silva et al. (2020) revealed the effectiveness of machine learning algorithms. Nonetheless, there was a lack of clarity in terms of model transparency. In comparative analysis, machine learning algorithms have been found to be superior to traditional risk-scoring algorithms for predicting GDM. For instance, Zhang et al. (2022) showed better predictive accuracy in the use of machine learning algorithms. Nonetheless, issues of generalisability have continued to exist. In this regard, it should be noted that most studies involved datasets of high quality and not representative of the resource-limited environments. For instance, Li et al. (2025) predicted GDM accurately retrospectively. However, they failed to produce clinically accurate real-time predictions. Prediction of GDM in the early stages of pregnancies has yielded positive results. Cubillos et al. (2023) improved the accuracy in predicting GDM earlier but depended on high-quality datasets not typical of resource-limited environments. Similarly, Li et al., (2025) combined it with Auto ML to enhance performance, albeit with higher computational demands.

Cyclic Gradient Boosting has recently emerged as a promising technique. Kamatchi & maniraj (2025) demonstrated its ability to balance accuracy and interpretability, while Hasan et al., (2025) combined it with Auto ML to enhance performance, although with higher computational demands. SHAP has also gained attention as a tool for model interpretability. Zhang et al., (2025) showed that SHAP explanations improve clinician's trust by clearly attributing prediction outcomes to specific features. However, computational cost and sensitivity to data quality remains the challenge. Within Nigeria, Kokori et al., (2024) explored machine learning for GDM detection and found promising results, though issues such as data imbalance and limited contextual relevance were noted. In general, existing literature reviewed in this study has shown three key gaps. These are limited interpretability, lack of context-specific models

for low-resource settings and insufficient integration of explainable techniques like SHAP with robust algorithms such as CGB. Thus, this study addresses these gaps by combining both approaches in a low-resource setting like Nigeria.

III. METHODOLOGY

This study adopts a structured and reproducible machine learning framework for the early prediction of gestational diabetes mellitus (GDM). The methodology consists of sequential phases, from data collection to model interpretation, ensuring both predictive performance and clinical relevance. The architecture of the proposed cyclic boosting model for early detection of gestational diabetes risk is shown in Figure 1.

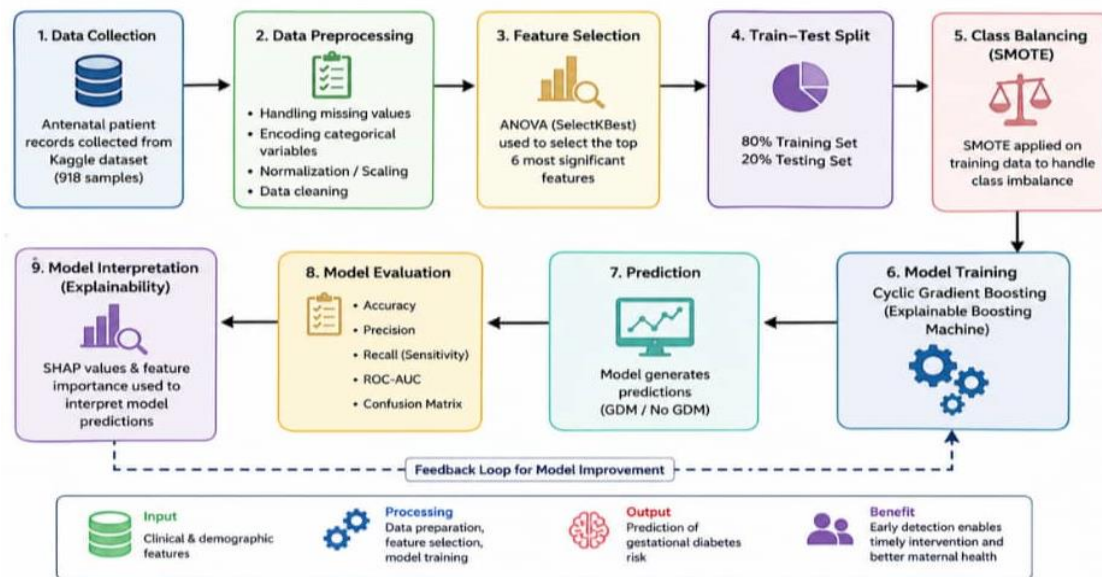


Figure 1: Architecture of the Cyclic Gradient Boosting Model.

A. Data Collection

A total of 918 antenatal patient records was obtained from clinical sources. The dataset includes routinely collected maternal health variables which are relevant to GDM risk assessment, such as Age, Body Mass Index (BMI), Weight, Hereditary history, and Pregnancy-related indicators.

B. Data Preprocessing

Data pre-processing was conducted to ensure quality and consistency. This involved:

- Handling missing values
- Encoding categorical variables into numerical representations
- Normalizing continuous variables to ensure uniform scale across features

C. Feature Selection

Feature selection was performed using the ANOVA F-test through the SelectKBest method. This approach evaluates the statistical relationship between each feature and the target variable. Six features with the highest significance scores were retained. These features were also supported by clinical evidence as key predictors of gestational diabetes, including maternal age, BMI, weight, hereditary history, and pregnancy-related factors. This step reduces model complexity while maintaining predictive performance.

D. Train-Test Split

The dataset was partitioned into training and testing subsets using an 80:20 ratio, 80% for Training and 20% for Testing. This ensured unbiased evaluation of model performance on unseen data.

E. Class Balancing Using SMOTE

To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was used to the training dataset. This method generates synthetic samples for the minority class, improving the model's ability to detect positive GDM cases.

F. Model Development

The Cyclic Gradient Boosting (CGB) model is an interpretable machine learning approach that builds predictions in an additive manner by iteratively updating feature-specific functions. Unlike traditional boosting methods that grow complex trees, CGB updates one feature at a time in a cyclic sequence, making the model more transparent and easier to interpret. The CGB model represents the prediction function as an additive model:

$$f(x) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_n(x_n)$$

where β_0 is the intercept term, $f_j(x_j)$ represents the contribution of feature j and n is the number of features. The model is trained iteratively by minimizing a loss function using gradient descent. At each iteration, the model updates one feature function at a time while keeping others fixed. The term "cyclic" refers to the process in which the model updates feature contributions sequentially. Instead of updating all features simultaneously, the algorithm cycles through each feature and adjusts its contribution based on the residual errors from previous iterations. This process continues until convergence is achieved.

Algorithm

Initialize model with constant prediction:

$$f(x) = \beta_0$$

Repeat until convergence:

For each feature j in $\{1, 2, \dots, n\}$:

Compute residuals: $r = y - f(x)$

Fit a simple function $f_j(x_j)$ to residuals

Update model:

$$f(x) = f(x) + \text{learning rate} * f_j(x_j)$$

Return final model

This cyclic feature updating mechanism allows the model to maintain high predictive performance while preserving interpretability, making it particularly suitable for clinical decision-support systems. Cyclic Gradient Boosting builds an additive model by updating one feature at a time in a repeating cycle, which improves accuracy while maintaining interpretability. This is different from traditional boosting methods that build complex trees.

G. Model Evaluation

Model performance was evaluated using:

- (a). **Accuracy:** This measures the overall proportion of correctly classified instances (both positive and negative) out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- (b). **Precision (Positive Predictive Value):** This quantifies how many of the instances predicted as *positive* (high risk of gestational diabetes) are actually correct.

$$Precision = \frac{TP}{TP + FP}$$

- (c). **Recall (Sensitivity or True Positive Rate):** Recall measures the ability of the model to correctly identify actual positive cases.

$$Recall = \frac{TP}{TP + FN}$$

- (d). **ROC-AUC (Receiver Operating Characteristic – Area Under Curve):** ROC-AUC evaluates the model's ability to distinguish between classes across all classification thresholds. It is the area under the ROC curve, which plots True Positive Rate (Recall) against False Positive Rate (FPR)

$$FPR = \frac{FP}{FP + TN}$$

If AUC = 1.0 then it's a perfect classifier, if AUC = 0.5 then it is random guessing, AUC > 0.8 is a strong discriminative ability. ROC-AUC is threshold-independent, making it particularly useful when tuning your Cyclic Gradient Boosting model for optimal sensitivity-specificity trade-offs. True Positives (TP) is the Correctly predicted gestational diabetes cases, True Negatives (TN) is the correctly predicted non-cases, False Positives (FP) is the incorrectly predicted as high risk and False Negatives (FN) is the missed true cases. Additionally, a 5-fold cross-validation strategy was employed to assess robustness and generalizability.

H. Model Interpretation

Model interpretability was achieved using SHAP (Shapley Additive Explanations). Both global and local explanations were generated to understand feature importance and individual prediction behaviour.

IV. RESULTS AND DISCUSSION

A. Model Performance Before and After SMOTE

The results shown in Table 1 demonstrate the impact of class balancing using SMOTE on model performance. While overall accuracy remained constant at 0.826, a substantial improvement was observed in recall, which increased from 0.537 to 0.732. This indicates that the model became significantly more effective in identifying true cases of gestational diabetes mellitus. The number of false negatives was also reduced from 19 to 11, further confirming enhanced detection capability. This improvement is particularly important in clinical settings, where false negatives represent missed diagnoses that may lead to severe maternal and neonatal complications. The use of SMOTE therefore directly addresses the class imbalance problem (77.7% vs 22.3%) and improves the model's sensitivity to minority class cases. Although precision decreased from 0.629 to 0.588, this trade-off is expected when improving recall in imbalanced datasets.

Table 1: Model Performance Before and After SMOTE

| Metric | Before SMOTE | After SMOTE |
|----------------------|--------------|-------------|
| Accuracy | 0.826 | 0.826 |
| ROC-AUC | 0.822 | 0.825 |
| Precision | 0.629 | 0.588 |
| Recall | 0.537 | 0.732 |
| False Negatives (FN) | 19 | 11 |

In clinical applications, especially for early disease detection, prioritizing recall is essential to minimize missed diagnoses. Additionally, the slight increase in ROC-AUC from 0.822 to 0.825 suggests marginal improvement in the model’s ability to distinguish between positive and negative cases. Overall, SMOTE contributed to a more clinically reliable model.

B. Model Performance After Threshold Adjustment

The model performance after adjusting the classification threshold from 0.5 to 0.4 is presented in Table 4.2. This adjustment increased recall to 0.756, indicating further improvement in the model’s ability to detect true GDM cases. This step was introduced specifically to address the limitation of low sensitivity identified during initial evaluation. In medical diagnostics, sensitivity is often prioritized over overall accuracy to reduce missed cases. However, this gain came at the expense of precision, which decreased to 0.500, and accuracy, which dropped to 0.777. This outcome reflects a typical trade-off in classification problems, where increasing sensitivity results in a higher number of false positives.

Table 2: Model Performance After Threshold Adjustment

| Metric | Value |
|-----------|-------|
| Accuracy | 0.777 |
| Precision | 0.500 |
| Recall | 0.756 |
| ROC-AUC | 0.825 |

Despite this, the adjustment is justified within a clinical context, where failing to detect a true case of gestational diabetes poses greater risk than incorrectly flagging a healthy individual. The ROC-AUC remained stable at 0.825, indicating that the model’s overall discriminative ability was not compromised. These results confirm that threshold tuning is an effective strategy for optimizing models for medical screening purposes.

C. Confusion Matrix After Threshold Adjustment

The confusion matrix in Table 3 shows a detailed breakdown of the model’s classification outcomes. The model correctly identified 31 true positive cases and 112 true negative cases, demonstrating its capability to detect both GDM and Non-GDM cases. Importantly, false negatives were reduced to 10, reinforcing the effectiveness of both SMOTE and threshold tuning in improving case detection. This aligns with clinical priorities where minimizing missed diagnoses is critical.

Table 3: Confusion Metrix After Threshold Adjustment

| Actual / Predicted | Predicted Negative | Predicted Positive |
|--------------------|--------------------|--------------------|
| Actual Negative | 112 (TN) | 31 (FP) |
| Actual Positive | 10 (FN) | 31 (TP) |

However, the model produced 31 false positives, reflecting the trade-off associated with increasing sensitivity. While this may lead to additional follow-up testing for some healthy individuals, it is acceptable in clinical settings where missing a diagnosis poses greater risks. Overall, the confusion matrix confirms that the model achieves a balance between detection and general classification performance, with a clear emphasis on minimizing missed cases.

D. Cross-Validation Result (5-Fold ROC-AUC)

Table 4 shows the results of the 5-fold cross-validation used to evaluate model robustness. The ROC-AUC scores range from 0.738 to 0.953, with a mean value of 0.840. This indicates that the model maintains strong predictive performance across different subsets of the data. However, the standard deviation of 0.075 suggests some variability in performance, indicating sensitivity to data distribution. This variation may be attributed to class imbalance or the relatively limited dataset size. To further quantify model reliability, a 95% confidence interval (CI) for the ROC-AUC was estimated using the t-distribution: 95% CI: 0.747 – 0.933.

Table 4: Cross-Validation Result (5-Fold ROC-AUC)

| Fold | ROC-AUC |
|------------------------------|---------------|
| Fold 1 | 0.953 |
| Fold 2 | 0.879 |
| Fold 3 | 0.848 |
| Fold 4 | 0.738 |
| Fold 5 | 0.783 |
| Mean | 0.840 |
| Standard Deviation | 0.075 |
| 95% Confidence Interval (CI) | 0.747 – 0.933 |

This interval indicates that the true model performance is likely to fall within this range, reflecting strong but moderately variable predictive ability. The relatively wide CI suggests some sensitivity to data partitioning, likely due to dataset size and inherent variability. However, the lower bound (0.747) still indicates acceptable discriminative performance. Overall, the performance remains acceptable, and the results demonstrate that the model generalizes reasonably well. Further improvements could be achieved through increased dataset size or additional feature engineering.

E. Cross Model Comparison

The comparative results in Table 5 show that all models achieved strong predictive performance, with Logistic Regression having the highest mean ROC-AUC (0.863), followed by the EBM model (0.840) and Random Forest (0.780). However, variability across folds was observed for all models.

Table 5: Cross-Validation Performance Across Models

| Fold | EBM (CGB) | Random Forest | Logistic Regression |
|--------------------|--------------|---------------|---------------------|
| Fold 1 | 0.953 | 0.877 | 0.966 |
| Fold 2 | 0.879 | 0.802 | 0.922 |
| Fold 3 | 0.848 | 0.744 | 0.863 |
| Fold 4 | 0.738 | 0.702 | 0.750 |
| Fold 5 | 0.783 | 0.777 | 0.813 |
| Mean | 0.840 | 0.780 | 0.863 |
| Standard Deviation | 0.075 | 0.067 | 0.084 |

To determine whether these differences are statistically significant, a one-way ANOVA test was conducted (Table 6). The results ($F = 1.454$, $p = 0.272$) indicate no statistically significant difference in performance among the models at the 0.05 significance level. Post-hoc analysis using Tukey HSD further confirmed that no pairwise differences were statistically significant, as all p-values exceeded 0.05.

Table 6: ANOVA and Post-hoc Statistical Results

| Test | Statistic | Value |
|--------------------------------------|---------------------|-----------------|
| ANOVA F-statistic | F | 1.454 |
| ANOVA p-value | p | 0.272 |
| EBM vs Logistic Regression | Mean Diff / p-value | 0.023 / 0.891 |
| EBM vs Random Forest | Mean Diff / p-value | -0.059 / 0.480 |
| Logistic Regression vs Random Forest | Mean Diff / p-value | -0.082 / 0.263 |
| Significance Level | α | 0.05 |
| Overall Conclusion | — | Not Significant |

These findings suggest that the observed variations in ROC-AUC are likely due to random variation rather than true performance differences. Despite similar predictive performance, the proposed Cyclic Gradient Boosting model offers a key advantage in interpretability, distinguishing it from more complex models such as XGBoost and Random Forest, which are often considered black-box approaches.

F. Model Interpretation (SHAP)

The SHAP method was applied to the trained Cyclic Gradient Boosting model to interpret predictions at both Global level (overall feature importance) and Local level (individual patient prediction). Figure 2 illustrates a SHAP Bee swarm plot in which each point represents an individual, positioned by its SHAP value to show the magnitude and direction of its contribution to the prediction, and coloured by feature value, with blue indicating lower values and red indicating higher values.

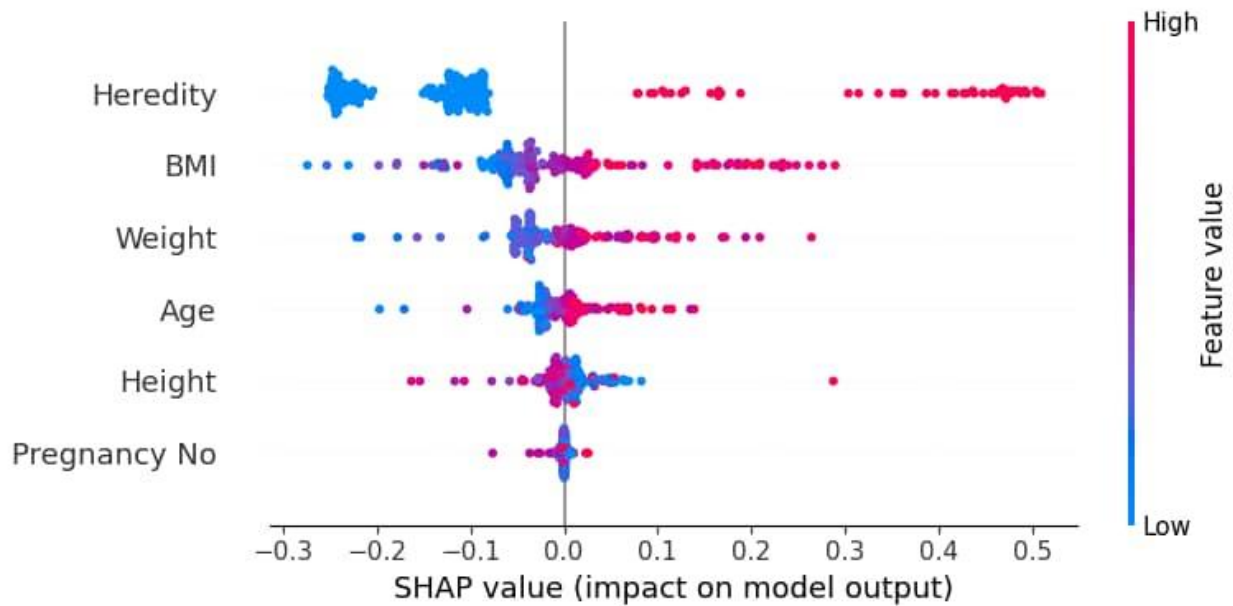


Figure 2: SHAP Value Beeswarm Plot

The vertical spread of the points captures variability in each feature's impact across the population. Family history of diabetes produced the widest range of SHAP values, with higher values clustering on the positive side at approximately +0.2 to +0.5, confirming a strong and consistent increase in predicted GDM risk when heredity is present. BMI produced a similar but slightly narrower spread, with higher values generating positive prediction shifts from +0.1 to +0.4 and reflecting a nonlinear pattern in which greater adiposity substantially raises GDM likelihood. Weight and maternal age followed comparable patterns with more restricted dispersion. Higher body weight contributed positively to predicted risk at approximately +0.1 to +0.3, and advancing maternal age produced incremental increases in GDM probability. Height and number of pregnancies showed limited SHAP variability, with most points clustered near zero and no consistent directional effects. Lower height values linked to a slight increase in predicted risk through their contribution to BMI rather than as independent factors, and their net influence was minimal. Parity similarly produced no strong or consistent prediction shift. Negative SHAP values appeared for lower BMI, and younger maternal age combined with the absence of family history of diabetes corresponded to reduced predicted risk, capturing their protective roles. The beeswarm plot also captures the directionality of each predictor's effect: higher values of BMI, weight, age, and heredity consistently produced positive SHAP values pointing to greater risk, while lower values corresponded to risk reduction. Non-linear patterns emerged for BMI in particular, where values above certain thresholds clustered sharply along the positive axis. The SHAP analysis also served as a bias assessment tool. Although the strong contribution of family history aligns with clinical knowledge, its prominence raises questions about equitable application across Nigeria's genetically and ethnically diverse population. Further validation in specific population subgroups remains necessary to confirm generalisable and equitable model performance.

V. CONCLUSION

This study presents a Cyclic Gradient Boosting model which demonstrates competitive performance when compared with existing machine learning approaches for gestational diabetes prediction. While models such as XGBoost and Random Forest may achieve slightly higher or similar accuracy, they lack interpretability, which is critical in clinical applications. Unlike black-box models, the proposed approach provides feature-level explanations, making it more suitable for medical decision-making environments. In healthcare applications, model interpretability is as important as predictive performance. Therefore, although some existing models achieve marginally higher accuracy, the proposed model offers a better balance between performance and explainability, making it more practical for real-world clinical deployment. Even though some models like XGBoost may have slightly higher accuracy, our model was chosen because it is interpretable and suitable for clinical use, where understanding predictions is critical.

REFERENCES

- [1] R. Awasthi et al., "Artificial intelligence in diabetes mellitus prediction: Advancements and challenges - A review," *Curr. Diabetes Rev.*, vol. 20, no. 10, pp. 853–868, 2025. doi: 10.2174/0115733998289291240221113912.
- [2] T. A. Azeez, T. Abo-Briggs, and A. S. Adeyanju, "A systematic review and meta-analysis of the prevalence and determinants of gestational diabetes mellitus in Nigeria," *Indian J. Endocrinol. Metab.*, vol. 25, no. 3, pp. 182–190, 2021. doi: 10.4103/ijem.ijem_301_21.
- [3] Y.-N. Chan et al., "A machine learning approach for early prediction of gestational diabetes mellitus using elemental contents in fingernails," *Sci. Rep.*, vol. 13, no. 1, art. no. 4184, 2023. doi: 10.1038/s41598-023-31270-y.
- [4] A. A. Choudhury and V. Devi Rajeswari, "Gestational diabetes mellitus—A metabolic and reproductive disorder," *Biomed. Pharmacother.*, vol. 143, art. no. 112183, 2021. doi: 10.1016/j.biopha.2021.112183.
- [5] N. Dubois and I. Giroux, "Gestational diabetes mellitus: Efficacy of non-pharmacological interventions for management and prevention," *Healthcare*, vol. 13, no. 18, art. no. 2261, 2025. doi: 10.3390/healthcare13182261.
- [6] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient boosting with categorical features support," *arXiv*, preprint arXiv:1810.11363, 2018.
- [7] N. M. Elfadel Magboul et al., "Machine learning for predicting the transition from gestational diabetes to type 2 diabetes: A systematic review," *Cureus*, vol. 17, no. 5, art. no. e84314, 2025. doi: 10.7759/cureus.84314.
- [8] X. Hu, X. Hu, Y. Yu, and J. Wang, "Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm," *Front. Endocrinol.*, vol. 14, art. no. 1105062, 2023. doi: 10.3389/fendo.2023.1105062.
- [9] A. Kamatchi and D. V. Maniraj, "Interpretable predictive models for Alzheimer's disease using Explainable Boosting Machines," *Int. J. Res. Publ. Rev.*, vol. 6, no. 3, pp. 4345–4352, 2025.
- [10] L. Kang, D. Luo, W. Xie, X. Luo, J. Mei, and J. He, "An explainable machine learning model for predicting preterm birth in pregnant women with gestational diabetes mellitus and hypertensive disorders of pregnancy: Development and external validation," *Front. Endocrinol.*, vol. 16, art. no. 1665935, 2025. doi: 10.3389/fendo.2025.1665935.
- [11] L. Kanguru, N. Bezawada, J. Hussein, and J. Bell, "The burden of diabetes mellitus during pregnancy in low- and middle-income countries: A systematic review," *Glob. Health Action*, vol. 7, art. no. 23987, 2014. doi: 10.3402/gha.v7.23987.
- [12] E. Kokori et al., "The role of machine learning algorithms in detection of gestational diabetes; a narrative review of current evidence," *Clin. Diabetes Endocrinol.*, vol. 10, no. 1, art. no. 18, 2024. doi: 10.1186/s40842-024-00176-7.