DOI: https://doi.org/10.61841/kmnfeh71 **Publication URL:** https://jarcse.com/index.php/CSE/article/view/11

A DATA-DRIVEN MACHINE LEARNING APPROACH FOR EARLY PREDICTION OF GESTATIONAL DIABETES MELLITUS

[1]Noor Sami Razzaq Najjar

[1]The General Directorate of Education, the Ministry of Education of Iraq Najaf, Iraq

Corresponding Author :
noor.najjar232@gmail.com

To Cite This Article: A DATA-DRIVEN MACHINE LEARNING APPROACH FOR EARLY PREDICTION OF GESTATIONAL DIABETES MELLITUS. (2026). *Journal of Advance Research in Computer Science & Engineering (ISSN2456-3552), 11(1),1-7.* https://doi.org/10.61841/kmnfeh71

ABSTRACT

Gestational diabetes mellitus (GDM) is a common pregnancy complication that can have serious consequences for both the mother and the fetus if not detected early. In this study, we propose an machine learning to predict gestational diabetes mellitus using the patient's clinical characteristics. The dataset was preprocessed to handle missing values, normalize numerical variables, and identify relevant clinical features to improve predictive performance. The data were divided into two sets: a test set and a training set. SMOTE was applied only to the training set. Methods: We conducted a retrospective model development and internal validation study using 10,000 pregnancy records. Candidate predictors included age, pre-pregnancy body mass index (BMI), ethnicity, family history of diabetes, previous GDM Blood Pressure Previous Pregnancies Diet & Lifestyle Missing-data handling and categorical encoding were nested within stratified 10-fold cross-validation to prevent data leakage. Three algorithms were compared: gradient boosted trees (GBT), random forest (RF), and neural network (NN). The primary metric was area under the receiver operating characteristic curve (ROC-AUC); secondary metrics included precision, recall, specificity, F1-score, accuracy, and precision-recall AUC (PR-AUC).

Results: GBT achieved the highest overall discrimination, and tree-based models outperformed the neural network by a small but consistent margin in this dataset. GBT offered the most balanced overall performance profile. This study supports the feasibility of early GDM prediction from routinely available antenatal variables, although temporal or external validation is still required before clinical implementation.

The experimental results indicate that the proposed framework showed promising predictive performance across several evaluation metrics of accuracy, precision, F1-score, and recall. This approach provides an effective and scientific solution for predicting gestational diabetes and can support clinical decision-making systems.

KEYWORDS: *Gestational diabetes mellitus; machine learning; early prediction; gradient boosted trees; random forests; neural networks.*

INTRODUCTION

Gestational diabetes mellitus is a major public health problem and a significant health challenge that threatens the health of both the mother and the fetus during pregnancy and affects many pregnancies worldwide [1, 2]. Routine screening is commonly performed at 24-28 weeks of gestation in many clinical settings [3,4], which limits the opportunity for early risk stratification, targeted counseling, and closer metabolic surveillance in the first trimester [5].

Gestational diabetes is characterized by high insulin levels and occurs primarily as a result of insulin resistance and relative insulin deficiency, which are the two main causes of gestational diabetes.[6] If managed carefully, it usually returns to normal after delivery [7]. Pregnancy is a state of physiological changes that represents a unique metabolic state, exposing women to a range of health problems, including gestational diabetes [8]. Gestational diabetes is diagnosed as varying degrees of glucose intolerance during pregnancy, where insulin production is required to maintain blood sugar levels. When the pancreas is unable to secrete the necessary amount of insulin, gestational diabetes develops. [9] Gestational diabetes poses significant risks to both the mother and the fetus if not diagnosed early, leading to missed opportunities for early intervention [10]. The increased use of medical services and the abundance of data available during pregnancy are driving the development of machine learning models using artificial intelligence that enable the early detection of women at risk of developing gestational diabetes. This can be achieved by relying on electronic health records. Early detection ensures optimal care and helps avoid associated risks such as cesarean delivery, fetal growth variability, preeclampsia, and type 2 diabetes mellitus [11].

Machine learning methods are well suited to this task because they can accommodate non-linear relationships and interactions among demographic, obstetric, and laboratory predictors [12].

This study aims to improve early prediction of gestational diabetes by using artificial intelligence algorithms for early detection. This will enhance screening, guide treatment and improve follow-up strategies. It represents a significant opportunity to improve detection rates and, consequently, outcomes before resorting to traditional pregnancy tests.

In this research, we compared three machine learning models for the early prediction of gestational diabetes using a number of routinely available prenatal variables. We also assessed the models using a set of multiple classification and discrimination measures and then identified the most significant indicators associated with the risk of developing gestational diabetes in the analyzed dataset.

MATERIALS AND METHODS

DATASET DESCRIPTION

The dataset used in this research was obtained from a general collection on gestational diabetes available on the Kaggle platform.[13] This dataset includes clinical records of pregnant women. It contained 10,000 pregnancy records and 17 columns, including one identifier field (patient_id), 15 candidate predictors, and one binary outcome label (gdm_outcome). and a classification of gestational diabetes presence (1) or absence (0) of GD. The dataset also includes variables such as age, body mass index (BMI), blood pressure, glucose and lipid profiles, family history of diabetes, and others. Stratified sampling was applied during data partitioning so that class proportions were preserved across training and testing sets.

Table 1: Dataset Features (Gestational Diabetes – Kaggle)

Feature	Type	Description
Age	Numerical	Age of pregnant woman
BMI	Numerical	Body Mass Index
Blood Glucose Levels	Numerical	Glucose measures
Blood Pressure	Numerical	Systolic & Diastolic BP
Family History of Diabetes	Categorical	Presence of familial diabetes
Previous Pregnancies	Numerical	Number of previous pregnancies
Diet & Lifestyle	Categorical	Lifestyle risk indicators
Target (GDM)	Binary	0 = Non-GDM, 1 = GDM

DATA PROCESSING

Missing numerical values were imputed using the median, while missing categorical values were imputed using the most frequent category. We also used the most frequent category to replace missing values and, to reduce scale variances, we standardized numerical features and created selected derived variables, including age groups and blood pressure categories, where appropriate We also collected glucose analysis tests.

The original dataset showed a class imbalance, with the number of cases without gestational diabetes exceeding the number of cases with it. To address this imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE) exclusively to the training subset.

MACHINE LEARNING MODELS

In this research, we used several machine learning algorithms for evaluation: gradient boosted trees (GBT), random forest (RF), and a feed-forward neural network (NN).

PROPOSED FRAMEWORK

Figure 1, shows the proposed framework consists of data preprocessing, class balancing, model development, validation, and final performance evaluation

The pipeline begins with data preprocessing applied to the input dataset, followed by a two-stage feature selection module The proposed framework consisted of data preprocessing, class balancing using SMOTE on the training data only, model development using GBT, RF, and NN, hyperparameter tuning within stratified 10-fold cross-validation, and final evaluation on the held-out test set,using accuracy, precision, recall, specificity, F1-score, ROC-AUC, and PR-AUC

All experiments were implemented in RapidMiner Studio. The same train-test partitioning and validation strategy were applied across all models to ensure fair comparison.

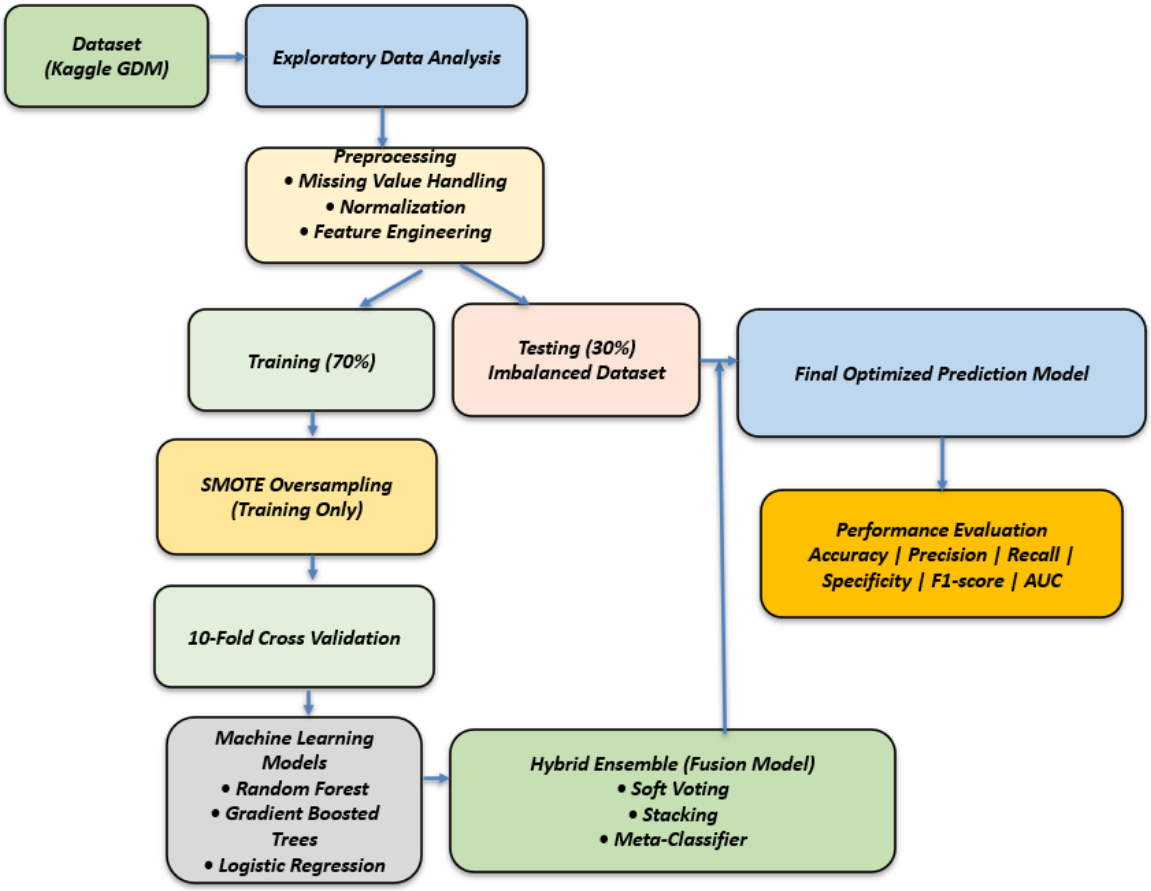

Figure 1: Workflow of the proposed machine learning framework for GDM prediction

RESULTS AND DISCUSSION

APPLYING CLASSIFICATION MODELS

The proposed framework was implemented using the RapidMiner Studio Model development and internal validation were performed using stratified k-fold cross-validation for the ANN, RF, and GBT models by using cross-validation, depending on the K-fold technique to minimise error rate. This technique divides the dataset into K equal subsets for training and testing. On the other hand, we trained the models by configuring their hyperparameters for three models as well as checking confusion matrix accuracy.

We applied machine learning classification models to the following steps, which divide and validate the data into training and test data.

The model was then optimized for best performance, and the accuracy of class detection in the artificial neural network was improved using a confusion matrix.

We performed the default configuration for the first step to achieve accuracy. For the neural network, we evaluated different architectures by adjusting the number of hidden neurons and monitoring performance using the confusion matrix and other evaluation metrics.

COMPARATIVE MODEL PERFORMANCE

All three algorithms achieved similar accuracy, a result in the range of 72.63% to 72.70%. However, more informative differences emerged in discrimination and the sensitivity-specificity trade-off. GBT achieved the highest ROC-AUC, PR-AUC, precision, and specificity, indicating the best overall balance between ranking ability and false-positive control. RF achieved the highest F1-score and recall, which favors its use in a screening context where missing fewer true GDM cases is the primary clinical goal The neural network remained competitive across all metrics.

Table 2 presents the comparative performance of the three machine learning models under stratified 10-fold cross-validation. As shown in Table 2, all models achieved similar accuracy, while GBT obtained the highest ROC-AUC, PR-AUC, precision, and specificity. In contrast, RF achieved the highest recall and F1-score, which may make it more suitable in screening contexts where reducing false negatives is especially important.

Table 2: Comparative performance of the three machine-learning models under stratified 10-fold cross-validation.

Model	Accuracy %	Precision %	Recall %	Specificity %	F1-score %	ROC-AUC	PR-AUC
Gradient boosted trees	72.70 ± 1.39	76.64 ± 1.32	83.52 ± 1.00	52.55 ± 3.36	79.93 ± 0.94	0.777 ± 0.016	0.851 ± 0.015
Random forest	72.63 ± 1.17	75.84 ± 1.14	85.06 ± 1.18	49.49 ± 3.27	80.17 ± 0.79	0.773 ± 0.017	0.848 ± 0.015
Neural network	72.64 ± 1.47	75.98 ± 1.43	84.80 ± 1.30	50.00 ± 4.05	80.13 ± 0.97	0.769 ± 0.016	0.843 ± 0.015

The performance differences between GBT and RF were modest; therefore, model selection should also consider clinical priorities such as sensitivity versus specificity,As shown in Figure 2, the performance differences between the models were modest, although GBT performed best in precision, specificity, and ROC-AUC, while RF led in recall and F1-score.

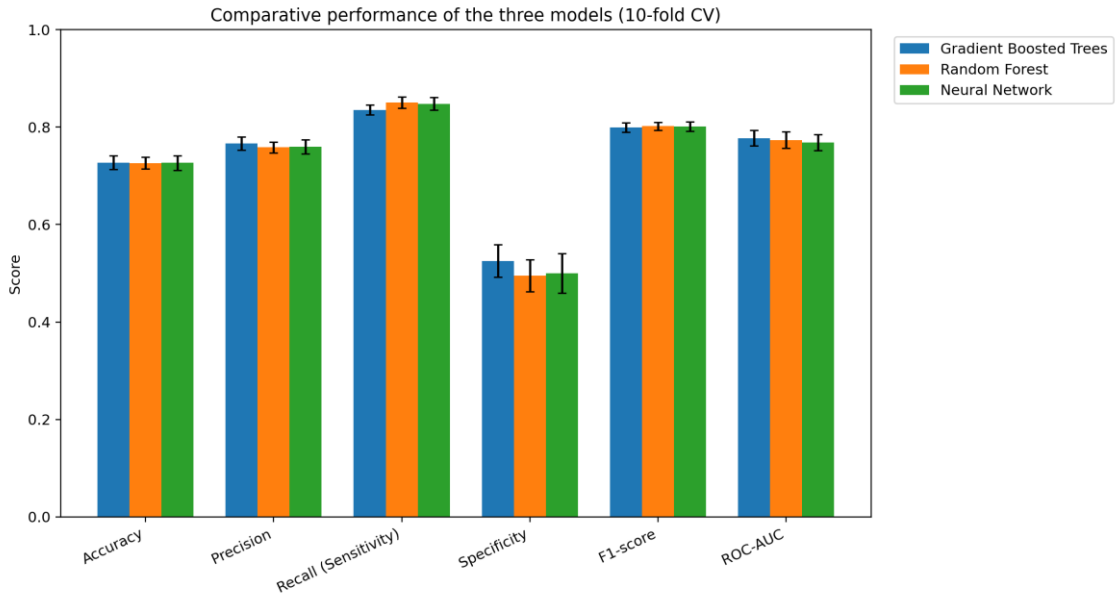

Figure 2: Grouped comparison of classification metrics across the three models. GBT had the highest precision, specificity, and ROC-AUC, while RF had the highest recall and F1-score.

Figures 3 and 4 further illustrate the discriminative performance of the evaluated models. Figure 3 presents the out-of-fold ROC curves, showing that GBT achieved the highest ROC-AUC, while Figure 4 presents the precision-recall curves, where GBT also obtained the highest average precision.

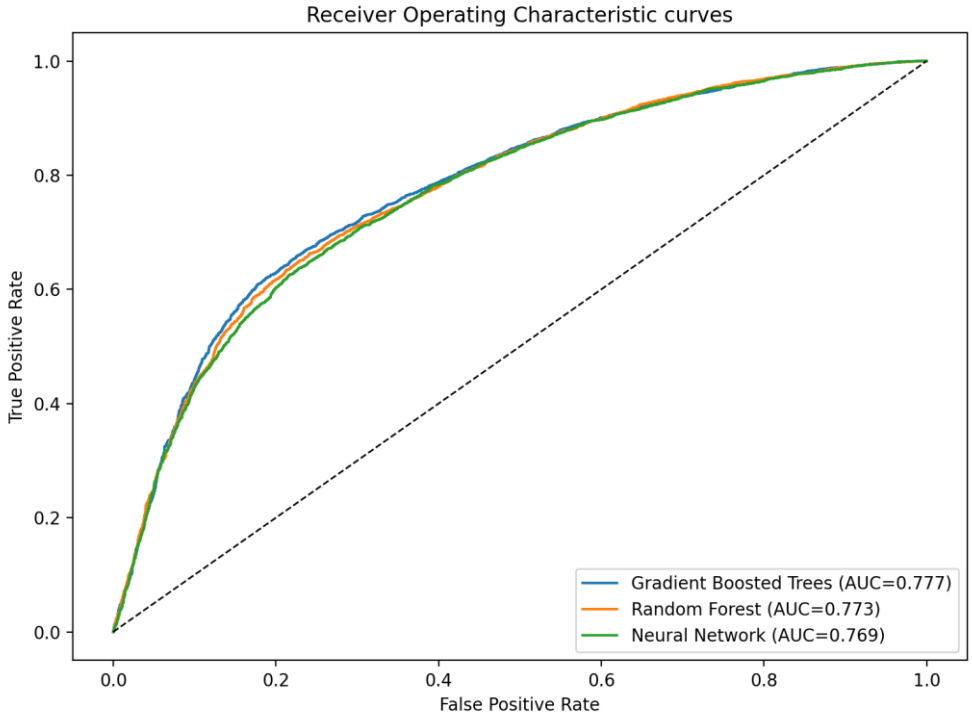

Figure 3: Out-of-fold receiver operating characteristic curves for gradient boosted trees, random forest, and neural network models.

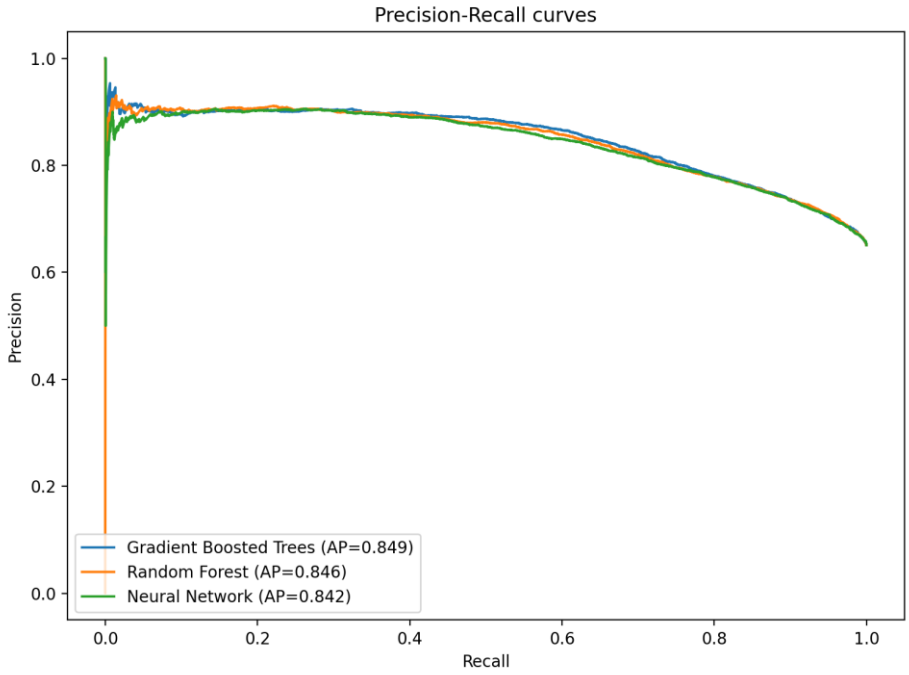

Figure 4: Out-of-fold precision-recall curves for the three models. The average precision was highest for gradient boosted trees.

CONFUSION MATRIX ANALYSIS AND VARIABLE IMPORTANCE

The confusion matrix analysis provided additional insight into the performance of the evaluated models. GBT produced fewer false positives and more true negatives, which explains its higher specificity. In contrast, RF achieved more true positives and fewer false negatives, which is consistent with its stronger recall and F1-score. The neural network showed competitive but slightly less balanced performance compared with the two tree-based models.

These results indicate that early glycemic indicators and maternal metabolic condition are pivotal in forecasting GDM risk. As shown in Figure 5, the aggregated confusion matrices provide further insight into the classification performance of the three models.

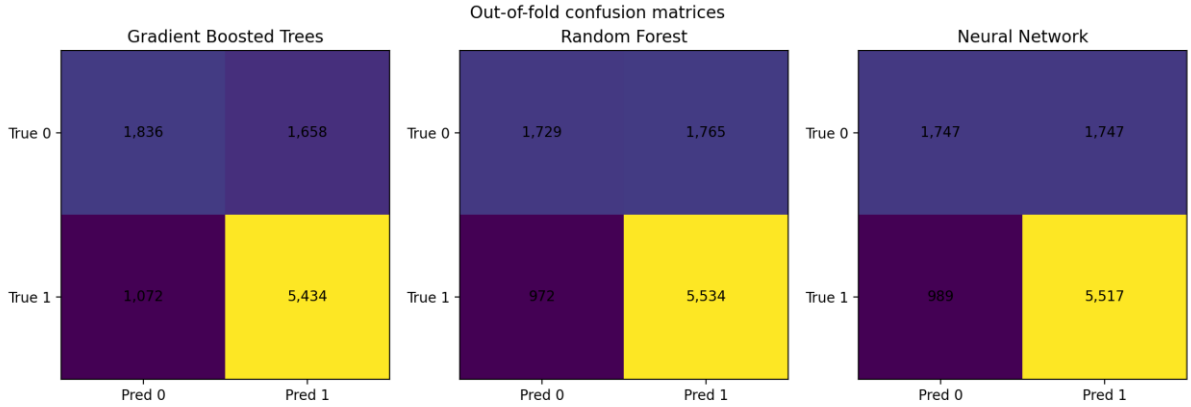

Figure 5: Aggregated out-of-fold confusion matrices for gradient boosted trees, random forest, and neural network models.

Figure 6 presents the top 10 predictors in the best-performing gradient boosted trees model according to permutation importance, expressed as reduction in ROC-AUC.

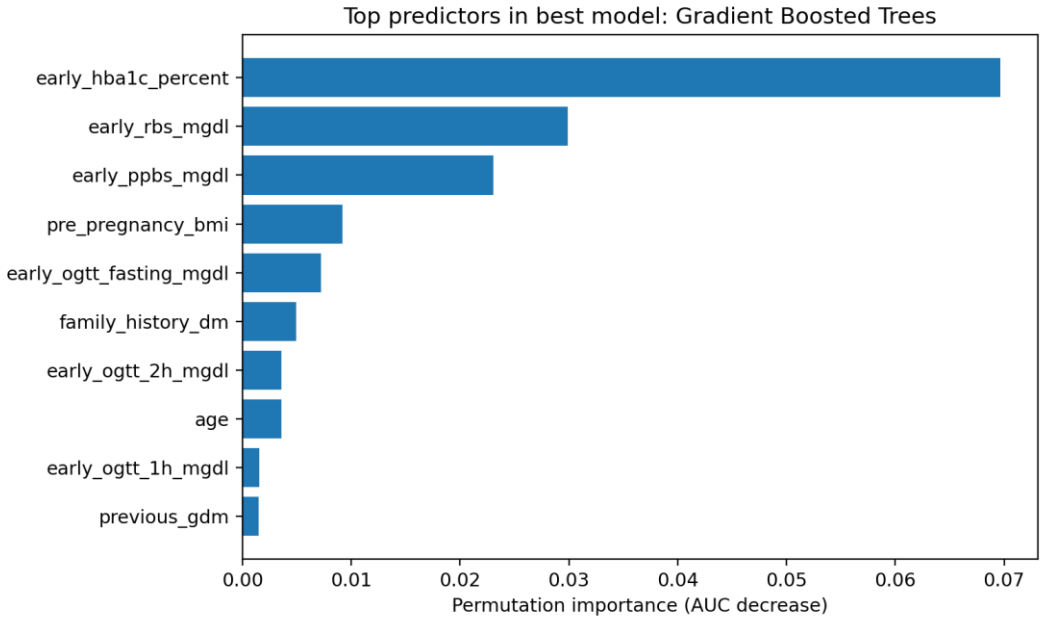

Figure 6: Top 10 predictors in the best-performing gradient boosted trees model according to permutation importance expressed as reduction in ROC-AUC.

CONCLUSION

This study compared three machine learning models—gradient boosted trees, random forest, and neural network—for the early prediction of gestational diabetes mellitus using routinely available antenatal variables. The findings showed that all three models achieved similar overall accuracy, while GBT provided the best overall discrimination and specificity. RF, on the other hand, achieved the highest recall and F1-score, which may be advantageous in screening-oriented settings where identifying as many true cases as possible is a priority.

The results also showed that early glycemic indicators and maternal metabolic characteristics were among the most important predictors of GDM risk. Overall, these findings support the feasibility of using machine learning to assist in early GDM risk stratification. Nevertheless, external validation is still required before such models can be considered for routine clinical use.

REFERENCES

1. Hassan A, et al. Enhanced model for gestational diabetes mellitus prediction using a fusion technique of multiple algorithms with explainability. *International Journal of Computational Intelligence Systems*. 2025;18(1):1–33.
2. Zaky H, et al. Machine learning based model for the early detection of gestational diabetes mellitus. *BMC Medical Informatics and Decision Making*. 2025;25(1):130.
3. Plows JF, Stanley JL, Baker PN, Reynolds CM, Vickers MH. The pathophysiology of gestational diabetes mellitus. *International Journal of Molecular Sciences*. 2018;19(11):3342. doi:10.3390/ijms19113342.

4. US Preventive Services Task Force, Davidson KW, Barry MJ, Mangione CM, et al. Screening for gestational diabetes: US Preventive Services Task Force recommendation statement. *JAMA*. 2021;326(6):531–538. doi:10.1001/jama.2021.11922.
5. Bhattacharya S, Nagendra L, Dutta D, et al. First-trimester fasting plasma glucose as a predictor of subsequent gestational diabetes mellitus and adverse fetomaternal outcomes: a systematic review and meta-analysis. *Diabetes & Metabolic Syndrome*. 2024;18(6):103051. doi:10.1016/j.dsx.2024.103051.
6. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
7. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
8. Bigdeli SK, Ghazisaedi M, Ayyoubzadeh SM, Hantoushzadeh S, Ahmadi M. Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: a cross-sectional study at a hospital fertility health center in Iran. *BMC Medical Informatics and Decision Making*. 2025;25(1):3. doi:10.1186/s12911-024-02799-3.
9. Zhou F, Ran X, Song F, et al. A stepwise prediction and interpretation of gestational diabetes mellitus: foster the practical application of machine learning in clinical decision. *Heliyon*. 2024;10(12):e32709. doi:10.1016/j.heliyon.2024.e32709.
10. Yang Z, Shi X, Wang S, et al. An early prediction model for gestational diabetes mellitus created using machine learning algorithms. *International Journal of Gynecology & Obstetrics*. 2025;170(2):665–674. doi:10.1002/ijgo.70055.
11. Gao J, Song S, Duo Y, et al. Establishment of an accurate prediction system for gestational diabetes mellitus based on the characteristics of metabolic kinetics in early pregnancy: a prospective two-center cohort study in population according to IOM criteria. *Endocrine*. 2025;90(3):1201–1220. doi:10.1007/s12020-025-04415-4.
12. Belsti Y, Moran L, Mousa A, Teede H, Enticott J. Evaluation of machine learning and logistic regression-based gestational diabetes prognostic models. *Journal of Clinical Epidemiology*. 2025;187:111957. doi:10.1016/j.jclinepi.2025.111957.
13. Gestational diabetes dataset. Kaggle. Available from: https://www.kaggle.com/datasets/akshaydattatraykhare/gestational-diabetes-dataset