Early Prediction of Gestational Diabetes Using Machine Learning Models with Non Invasive Clinical Features

Khadiza Akter[1], Most. Jannatul Ferdows[1], Mohammed Motaher Hossain[2*]

[1] Department of Nursing, International University of Business Agriculture and Technology Dhaka, Bangladesh

[2] Department of Statistics, International University of Business Agriculture and Technology Dhaka, Bangladesh

{khadiza.cn, jannatul.rn}@iubat.edu, motaher@iubat.edu*

Abstract. Gestational diabetes mellitus is linked to adverse maternal and neonatal outcomes, which can be improved through early prediction. This current study evaluates the predictive performance of several machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, Light Gradient Boosting Machine, Extreme Gradient Boosting Tree, and Extreme Gradient Boosting by using a publicly available dataset consisting of 1013 records. The dataset included six predictive features: age, pregnancy number, weight, height, body mass index, heredity, and the outcome variable indicate the presence or absence of gestational diabetes. Data pre-processing steps included outlier detection, standardization, and Synthetic Minority Oversampling Technique for class balancing. Model performance was evaluated using accuracy, precision, recall, specificity, and F1 score. LightGBM achieved the highest overall accuracy (89.62%), followed by XGBTree (88.46%) and RF (87.69%). Our findings align with prior research showing the superiority of ensemble models in capturing complex feature interactions.

Keywords: Gestational Diabetes Mellitus, Early Detection, Machine Learning, Predictive Modeling

1. Introduction

Gestational Diabetes Mellitus (GDM) is a common pregnancy complication characterized by elevated blood glucose levels which commonly diagnosed between 24–28 weeks of gestation [1]. The consequences of GDM are concerning due to its significant risks for both the mother and the child. It increases the likelihood of complications such as preeclampsia, preterm birth, cesarean delivery, and an intensified risk of developing type 2 diabetes in the future. For the child, GDM increases the risk of obesity, glucose intolerance, and cardiovascular diseases later in life future [2], [3], [4].

The global prevalence of GDM is estimated to be 14% which has a significant regional variation. The highest prevalence rates were observed in the Middle East and North Africa (27.6%) and South-East Asia (20.8%), while the prevalence is lower in most developed countries, ranging from 7%

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_22

to 10% [2], [5]. It is estimated that 1 in every 6 live births is complicated by GDM globally [6]. Several predisposing factors, such as older maternal age, obesity, multiparity, and a family history of diabetes, have been identified as increasing the likelihood of developing GDM [7], [8]. It is commonly diagnosed by using oral glucose tolerance tests [1]. However, there is no single global standard for diagnosing GDM. Early prediction of GDM is highly important to improve maternal and child health outcomes as it enhances timely intervention. In this context, researchers have increasingly turned to computational classifiers to predict GDM at earlier stages of pregnancy.

This study aims to explore a simplified and cost-effective approach using only non-invasive variables such as age, weight, height, BMI, number of pregnancies, and family history to provide an effective means of early GDM detection. Unlike many prior studies which mostly rely on extensive clinical data, our work emphasizes early prediction using minimal invasive data. We also comprehensively evaluated six different machine learning algorithms including ensemble methods and evaluates their predictive accuracy on publicly available dataset. Thus, our study uniquely contributes to the existing literature by leveraging non- invasive clinical features to avoid the dependency on invasive biochemical markers in diverse clinical contexts including the low resource settings.

2. Literature Review

Gestational Diabetes Mellitus (GDM) is a prevalent metabolic disorder in pregnancy associated with adverse maternal and neonatal outcomes. In recent times machine learning classifiers are showing better predictive accuracy compared to traditional statistical methods. Researchers are continuously exploring different machine learning approaches for early detection of gestational diabetes mellitus.

Ye et al. compared conventional logistic regression with multiple ML models, including DT, RF, and gradient boosting methods, using clinical and laboratory data from Chinese pregnant women. Their findings indicated that ensemble models, particularly RF and gradient boosting, achieved superior predictive performance, with Area Under Curve (AUC) values exceeding those of logistic regression, highlighting the advantage of non-linear decision boundaries in capturing complex interactions in biomedical datasets [9]. Another study by Artzi et al. utilized nationwide electronic health record (EHR) data to develop GDM prediction models with RF, XGBoost, and LightGBM. LightGBM emerged as a strong performer, demonstrating high computational efficiency and robustness to missing values, while maintaining competitive predictive accuracy [10].

Similarly, in a large-scale cohort of Asian women, Zhang et al. implemented XGBoost and RF models, achieving AUC scores above 0.85, which significantly outperformed traditional approaches [11]. Meta-analytical evidence has further reinforced the value of ML in GDM prediction. A systematic review and meta-analysis by Zhao et al

synthesized findings from multiple studies, concluding that tree-based ensemble methods such as RF and XGBoost consistently outperformed single classifiers like DT or KNN in terms of sensitivity, specificity, and balanced accuracy [12].

In Iran, a study by Bigdeli et al. employed RF, DT, and KNN models to predict GDM risk factors. RF achieved the highest accuracy (~86%), benefiting from its ensemble nature and ability to handle high-dimensional data [13]. Similarly, Mennickent et al. assessed multiple ML-based models for GDM prediction and found that gradient boosting algorithms delivered more reliable predictions across diverse datasets [14]. Zhang et al. developed a novel nomogram integrating ML-derived features from LightGBM with clinical risk factors, improving interpretability without sacrificing predictive power. This hybrid approach offers clinicians a user-friendly tool that leverages both statistical transparency and ML performance [15].

It is evidenced that there is variation in the predictive outcomes depending on the selected algorithm and data set. Overall, the literature shows that advance machine learning demonstrates superior predictive metrics than the simpler algorithms. In our current paper we aimed to show the differences of predictive outcomes of six machines learning outcome in which logistic regression will be the base line.

3. Methodology

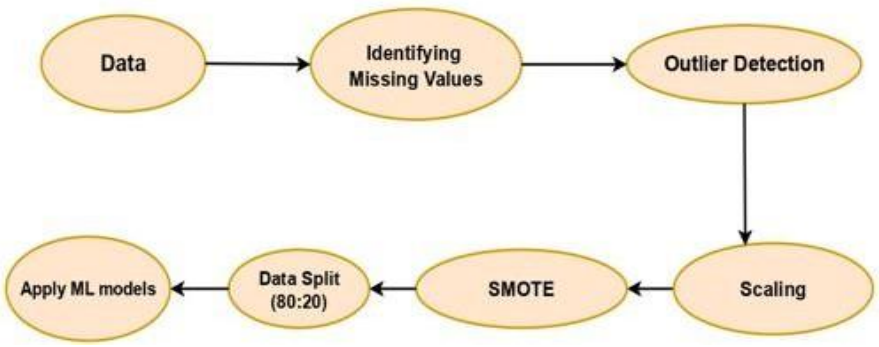

Fig.1. Work flow diagram

The methodology diagram presented in Fig. 1 and it demonstrates the data collection to apply ML models for analyzing the early predictions of diabetes.

3.1. Data Collection

In our study we used kaggle gestational diabetes dataset which has been obtained from the Kurdistan region laboratories containing 1013 sample. This dataset contains six variables: age, pregnancy number, weight, height, BMI, hereditary, and prediction (GDM present or absent). The categorical variable was coded as 0 and 1 in which 1 indicates presence of the condition and 0 indicates absence of the disease [16].

3.2. Data Preprocessing

Data quality plays an important role in assessing the performance of a machine learning model in which data preprocessing is the key to achieving this. The Data preprocessing steps employed in this paper has discussed in the following section.

3.2.1. Data Visualization and Cleaning

Our data set consists no missing values. Outliers were detected through IQR method to identify and handle extreme values, as seen in Fig. 2. We have used the following equation.

$$L = Q1 - (1.5 * IQR) \quad (1)$$

$$U = Q3 + (1.5 * IQR) \quad (2)$$

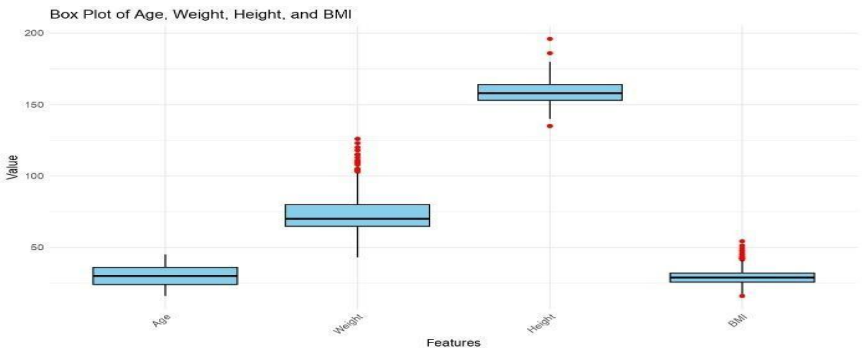

Fig.2. Box plots with outliers of selected continuous variables.

The correlation coefficient matrix was obtained to observe the relation between the different attributes and the output. Fig. 3 shows the correlation matrix where the coefficient indicates both the strength of the relationship between the variables as well as the direction.

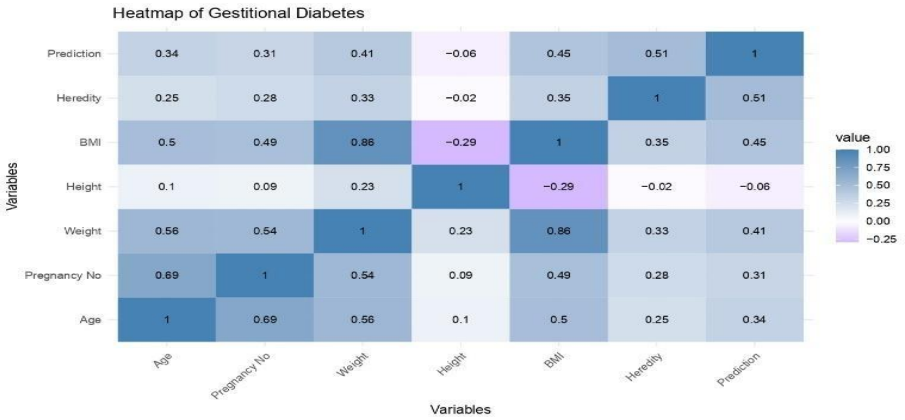

Fig.3. Correlation Matrix of gestational diabetes datase

Fig. 4 figure provides an exploratory analysis of Age, BMI, Weight, and Height across prediction classes. Diagonal plots show the distribution of each variable, with clear shifts in BMI and Weight between groups. Pairwise scatter plots and Pearson correlation coefficients reveal a strong positive relationship between BMI and Weight, with statistical significance marked by asterisks. Boxplots highlight group differences and the class distribution plot confirms a balanced dataset for prediction.

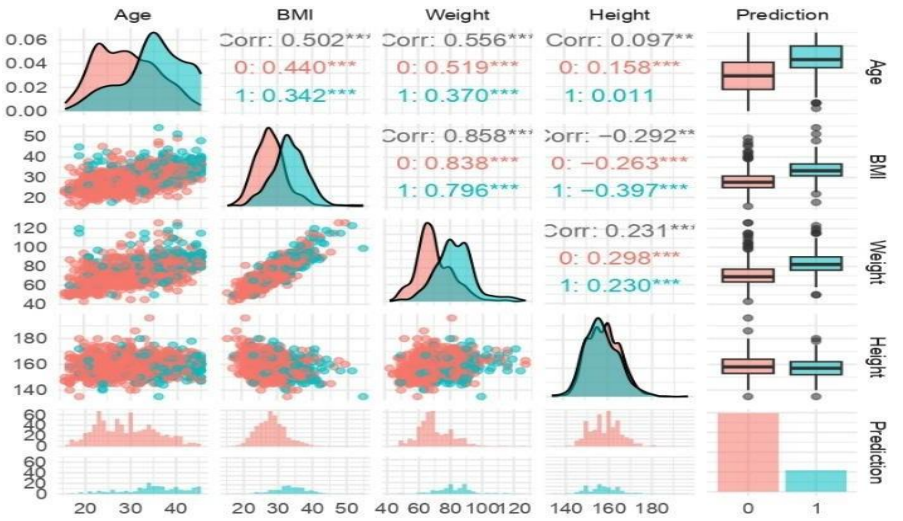

Fig.4. Pair Plot of Selected Features Grouped by gestational diabetes.

3.2.2. Checking for Imbalances

Fig. 5 shows KDE plots of Age, Weight, Height, and BMI for two prediction classes (0 = blue, 1 = red). Age and Weight distributions indicate that class 1 generally has higher values than class 0. BMI also shows a clear shift toward higher values in class 1, while Height distributions overlap more closely. These patterns suggest potential variable importance in distinguishing between the two classes.

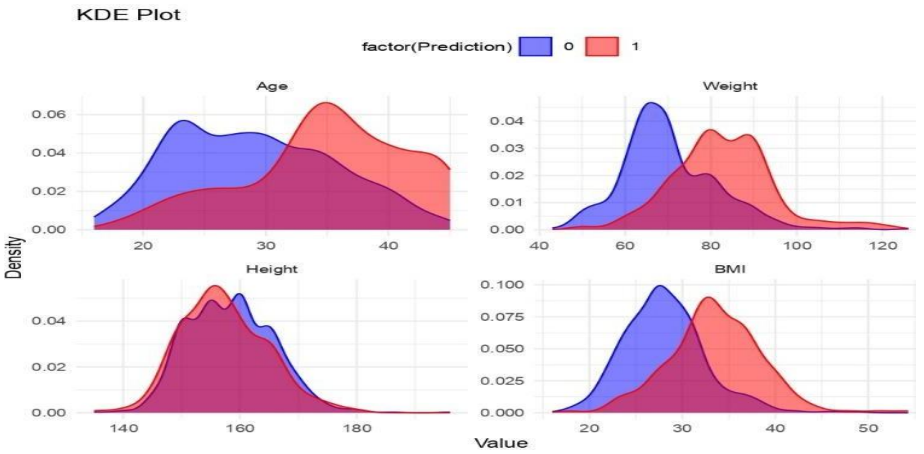

Fig.5. KDE plots of Age, Weight, Height, and BMI for two prediction classes (0 = blue, 1 = red).

3.2.3. Z-score Scaling

All numerical features were standardized using Z-score scaling to ensure comparability and prevent features with larger numeric ranges from dominating model training. Standardization was specifically applied for algorithms sensitive to feature magnitude, such as logistic regression. Tree-based models are generally insensitive to feature scale, so scaling was not necessary for them, but it was applied across all models for consistency and easier comparison.

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

3.2.4. Data balance by SMOTE (Synthetic Minority Over- sampling Technique)

To address class imbalance, we have applied the Synthetic Minority Oversampling Technique (SMOTE) to synthesize minority class samples, thereby reducing bias toward the majority class and improving generalization. Fig.6. compares class distributions before and after applying SMOTE. Initially, the dataset had a significant imbalance in which class 0 having far more samples than class 1. Synthetic Minority Oversampling Technique (SMOTE) forms synthetic examples for the minority class to gain balance. It prevents model bias toward the majority class and improves predictive performance on underrepresented categories.

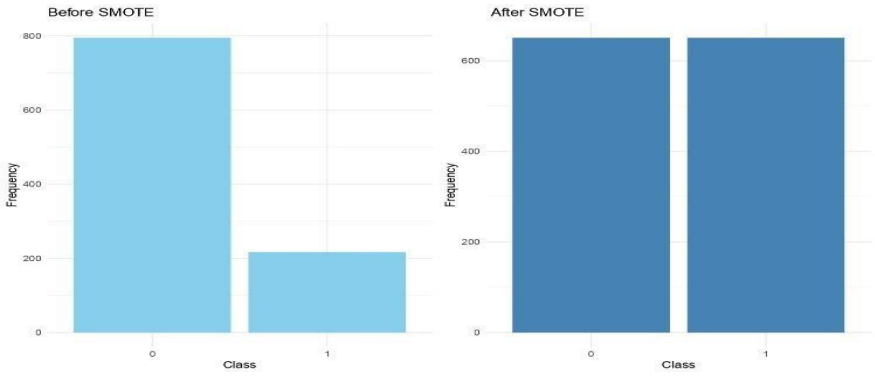

Fig.6. Class distribution before and after applying SMOTE.

3.3. Applied Algorithms

3.3.1. Random Forest (RF)

Random forest is considered as an effective machine learning algorithm which is commonly used in classification and regression. It combines multiple models like a forest consists of a lot of trees to achieve a better prediction [17].

$$G = 1 - \sum_{i=1}^{n} p_i^2 \quad (4)$$

3.3.2. Logistic Regression (LR)

Logistic Regression is a statistical model which is used for binary classification tasks for example, predicting the presence or absence of heart disease. It estimates the

probability of a binary response based on one or more predictor variables [18].

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (5)$$

3.3.3. Decision Tree (DT)

A Decision Tree is a widely used supervised machine learning algorithm for classification and regression. It makes prediction by splitting the data into branches based on specific decision rules, creating a tree-like structure that is easy to interpret [19], like Fig. 7.

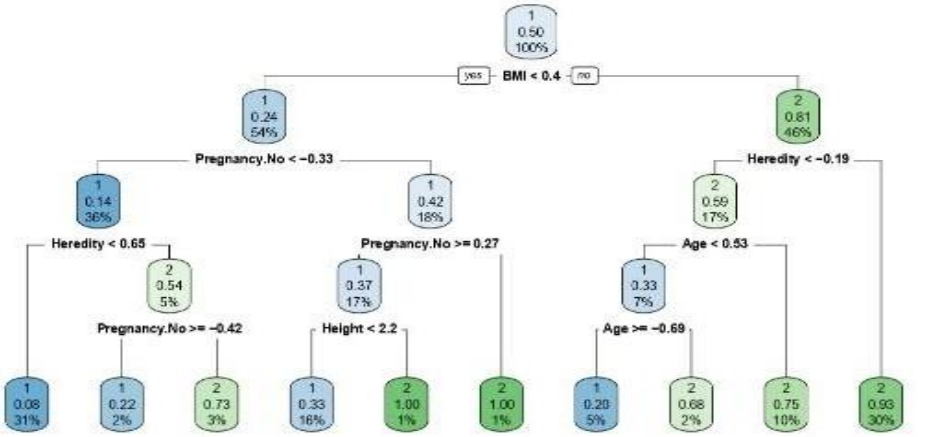

Fig.7. Used model in GDM prediction.

3.3.4. Light Gradient Boosting Machine (LGBM)

LGBM is a gradient boosting framework particularly suitable for large datasets. It uses a novel tree-based learning algorithm that grows tree leaf-wise instead of level-wise because it often provides better accuracy. It is also suitable in handling categorical features directly without the need for one-hot encoding [20].

The objective function of Light GBM is-

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \lambda \sum_{k=1}^{K} |\theta_k|^2 \quad (6)$$

3.3.5. Extreme Gradient Boosting Tree (XGBtree)

XGBtree, or Extreme Gradient Boosting Tree, is a powerful machine learning algorithm that builds an ensemble of decision trees in sequence, where each new tree corrects the errors of the previous ones, resulting in fast and highly accurate predictions [21].

The objective function of XGBtree is-

$$L = \sum_{i=1}^{n} (y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (7)$$

3.3.6. Extreme Gradient Boosting (XGB)

XGBoost is one of the advanced implementations of gradient boosting which use decision trees for its base learner. Regularization, handling missing data and parallelization are commonly used by XGBoost for an efficient and improved predictive performance. It aims to minimize the loss function by iteratively adding new trees that correct the errors of previous trees [21].

The objective function of XGBoost is-

$$Obj(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (8)$$

3.4. Evaluation Metrics

In our paper, we assess the quality and performance of machine learning models using a confusion matrix (N×N), where N represents the number of predicted classes. For a binary classification task, this matrix has dimensions of 2×2, as seen in Table 1.

Table 1. Confusion Matrix for Classification Model

Predicted Class	Actual Class: 1	Actual Class: 0
1	False Positive (FP) True Positive (TP)	
0	True Negative (TN)	False Negative (FN)

The confusion matrix presents the number of correct and incorrect predictions for each class. For example, True Positives (TP) present positive cases correctly identified while True Negatives (TN) represent negative cases correctly identified. The model's performance was calculated using standard metrics which includes accuracy, precision, recall (sensitivity), specificity, and F1 score. Accuracy presents the proportion of correct predictions, precision presents the proportion of true positive results among all positive predictions, recall measures the model's ability to identify actual positives, specificity measures the correct identification of negatives, and the F1 score provides a balance between precision and recall. Following equations (see Table 2) have been applied to assess the performance of the selected models.

Table 2. Evaluation Metrics for Classification Model

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

Sensitivity (Recall)	$RE = \frac{TP}{TP + FN}$
Specificity	$SP = \frac{TN}{TN + FP}$
F1 Score	$F1 = \frac{2TP}{2TP + FP + FN}$

4. Result and Discussion

From Table 3 and Fig. 8, among all tested models, LightGBM showed the highest accuracy (89.62 %) followed by XGBTree with an accuracy of 88.46%. LightGBM also outperformed in recall (90.77%) and F1 score (89.73%). XGBTree showed the highest specificity (90.00%) which indicates its effectiveness in correctly identifying non-GDM cases. The performance of Random Forest (RF) model was also well, showing an accuracy of 87.69% with balanced precision and recall but its specificity was slightly lower than that of XGBTree. Logistic Regression (LR) achieved moderate performance with an accuracy of 80.77%, while Decision Tree (DT) showed lower predictive capability including the weakest recall (73.08%) and F1 score (77.55%). XGBoost showed balanced metrics overall (accuracy: 84.62%, F1: 84.50%). Our finding is consistent with prior study indicating ensemble methods demonstrated clear advantages over single classifiers in modeling complex, non-linear relationships within biomedical datasets.

With earlier studies on GDM prediction, Ye et al. found that ensemble approaches for example RF and gradient boosting outperformed logistic regression, particularly in capturing non-linear feature interactions [9]. Using nationwide EHRdata, Artzi et al. reported that LightGBM not only achieved high predictive accuracy but was also computationally efficient and robust to missing data [10]. Similar trends were observed by Kang et al. in a large Asian cohort, where XGBoost and RF models achieved AUC scores above 0.85, significantly outperforming traditional models [11]. Zhao et al. further confirmed these advantages in a meta-analysis, showing that tree-based ensembles consistently delivered superior sensitivity, specificity, and balanced accuracy compared to single classifiers [12].

Evidence from other populations further supports our findings. Bigdeli et al. obtained approximately 86% accuracy with RF when predicting GDM risk factors in an Iranian dataset, while Mennickent et al. reported that gradient boosting models produced reliable predictions across diverse settings [13], [14]. The LightGBM-based nomogram developed by Li et al. demonstrated that interpretability can be improved without compromising accuracy [15]. In addition, Rahman et al. showed that fine-tuned gradient boosting models like XGBTree can surpass 90% accuracy when incorporating longitudinal glucose data and demographic variables [10].

Overall, our results reinforce the growing evidence that ensemble models particularly, LightGBM, XGBTree, and RF offer strong predictive capability for GDM. Their ability to manage high-dimensional, heterogeneous datasets and uncover complex feature relationships makes them promising tools for timely identification of at-risk pregnant women and clinical decision support in managing GDM risk. Further the use of only non-invasive features makes our models highly suitable in low-resource settings where the laboratory testing may not be immediately available. Thus, our work may complement traditional diagnostic methods and act as scalable early screening tools to contribute earlier interventions in lowering adverse maternal and neonatal outcomes.

Table 3. Details of model's performance using the testing set after feature selection

Model	Accu-racy (%)	Prec-ision (%)	Recall (%)	Speci-ficity (%)	F1 Score (%)
RF	87.69	88.28	86.92	88.46	87.60
LR	80.77	82.78	77.69	83.85	80.16
DT	78.85	82.60	73.08	84.62	77.55
LGBM	89.62	88.72	90.77	88.46	89.73
XGBtree	88.46	89.68	86.92	90.00	88.28
XGBoost	84.62	85.16	83.85	85.38	84.50

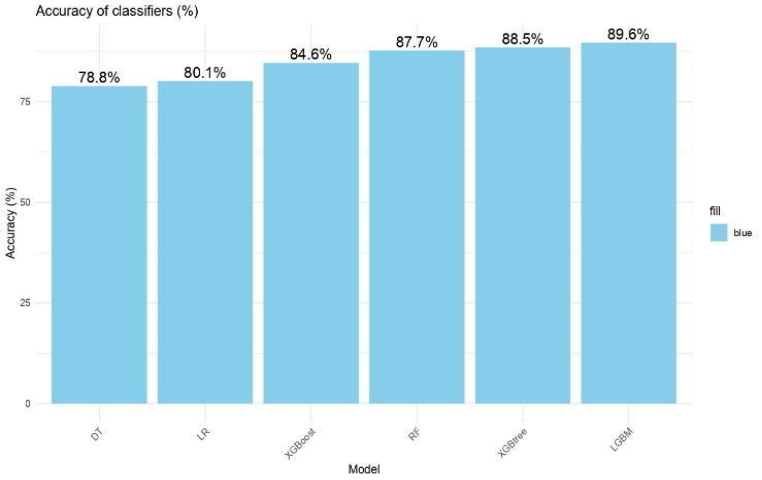

Fig.8. Accuracy comparison of applied models.

Table 4 and Fig. 9 presented the confusion matrix results of all models on the testing dataset. Among them, LGBM achieves the best performance with the highest true positives (118) and the lowest false negatives (12), making it most effective for detecting AD cases. XGBtree records the lowest false positives (13), indicating better control of false alarms. Random Forest shows balanced performance, while XGBoost performs moderately. In contrast, Logistic Regression and Decision Tree exhibit higher false negatives, making them less suitable for accurate medical diagnosis. Overall, boosting-based models, particularly LGBM, provide more reliable results.

Table 4. Details of confusion using the testing set

Model	False Positive	True Positive	True Negative	False Negative

RF	15	113	115	17
LR	21	101	109	29
DT	20	95	110	35
LGBM	15	118	115	12
XGBtree	13	113	117	17
XGBoost	19	109	111	21

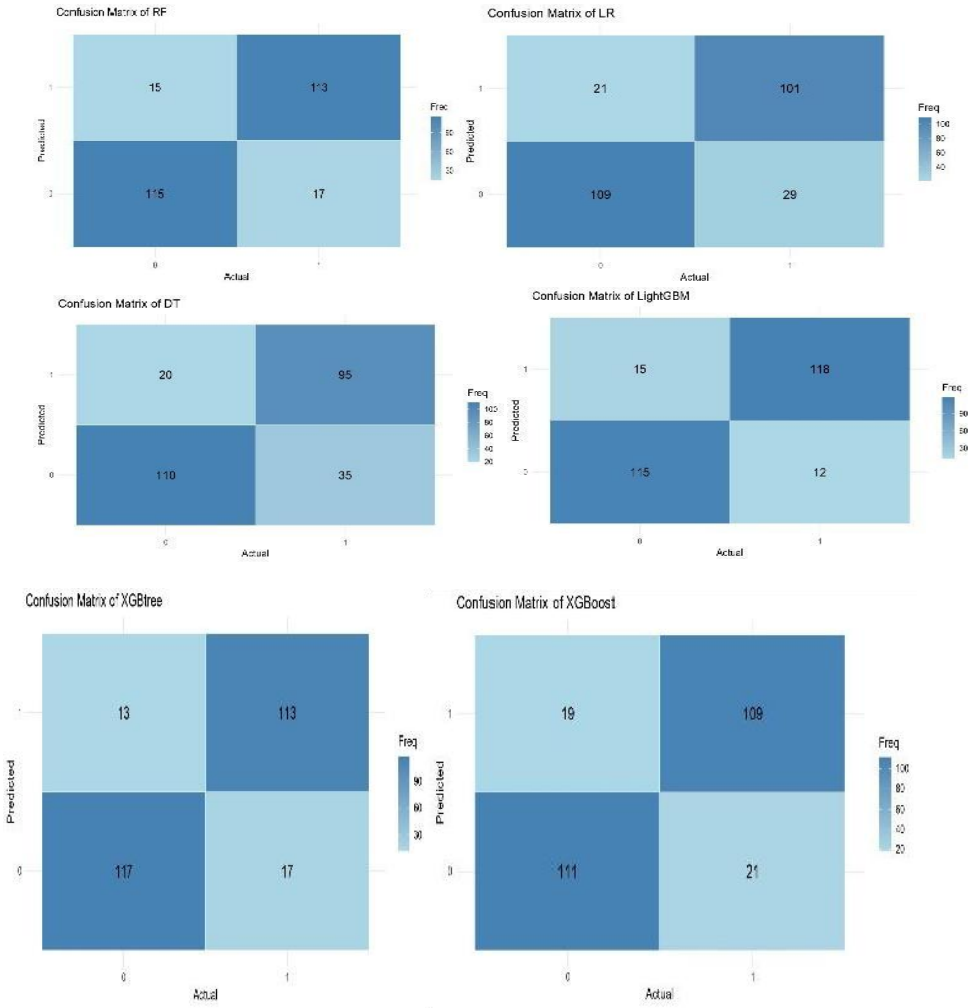

Fig.9. The confusion matrix of the studied model RF, LR, DT, LGBM, XGBtree and XGBoost.

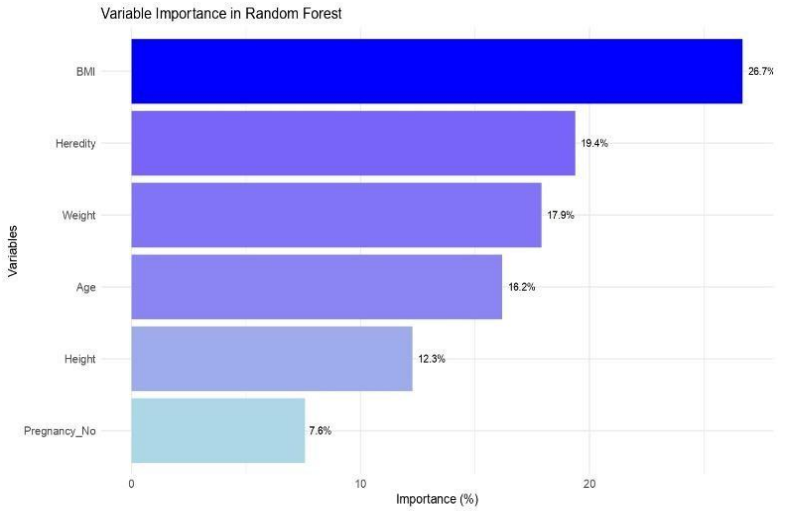

Fig.10. Variable importance with Random Forest.

The Fig. 10 presents the variable importance from a random forest model, showing the relative contribution of each feature to the model's predictive performance. BMI is the most influential predictor at 26.7%, followed by heredity (19.4%), weight (17.9%), and age (16.2%). Height (12.3%) and pregnancy_no (7.6%) contribute less. Overall, the results indicate that BMI, heredity, weight, and age are the key drivers in the model.

5. Conclusion and Future Work

Among six applied algorithms LightGBM achieved the highest accuracy (89.62%), followed by XGBTree (88.46%) in early prediction of gestational diabetes. Our study findings emphasize the effectiveness of ensemble methods like LightGBM and XGBTree in handling complex biomedical data and their suitability for resource-strain settings. These models offer a scalable solution for early GDM detection and may contribute in reducing obstetric complications. Future work may focus on applying these models to larger datasets by integrating both clinical and non-invasive data as well as incorporating longitudinal data to enhance predictive accuracy.

Acknowledgment

Thanks to the Kaggle gestational diabetes dataset for its significant contribution in the success of our work.

References

1. Mack LR, Tomich PG. Gestational diabetes. Obstetrics and gynaecology clinics of North America. 2017 Jun;44(2):207-17.
2. Sweeting A, Hannah W, Backman H, Catalano P, Feghali M, Herman WH, Hivert MF, Immanuel J, Meek C, Oppermann ML, Nolan CJ. Epidemiology and management of gestational diabetes. The Lancet. 2024 Jul 13;404(10448):175-92.
3. Johns EC, Denison FC, Norman JE, Reynolds RM. Gestational diabetes mellitus:

mechanisms, treatment, and complications. Trends in Endocrinology & Metabolism. 2018 Nov 1;29(11):743-54.
4. Modzelewski R, Stefanowicz-Rutkowska MM, Matuszewski W, Bandurska-Stankiewicz EM. Gestational diabetes mellitus recent literature review. Journal of clinical medicine. 2022 Sep 28;11(19):5736.
5. Wang H, Li N, Chivese T, Werfalli M, Sun H, Yuen L, Hoegfeldt CA, Powe CE, Immanuel J, Karuranga S, Divakar H. 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 research and clinical practice. 2022 Jan 1;183:109050.
6. Arianne Sweeting, Jencia Wong, Helen R Murphy, Glynis P Ross, A Clinical Update on Gestational Diabetes Mellitus, Endocrine Reviews, Volume 43, Issue 5, October 2022, Pages 763 793,
7. Yong HY, Mohd Shariff Z, Mohd Yusof BN, Rejali Z, Tee YY, Bindels J, van Der Beek EM. Independent and combined effects of age, body mass index and gestational weight gain on the risk of gestational diabetes mellitus. Scientific Reports. 2020 May 22;10(1):8486.
8. Fu Q, Chen R, Xu S, Ding Y, Huang C, He B, Jiang T, Zeng B, Bao M, Li S. Assessment of potential risk factors associated with gestational diabetes mellitus: evidence from a Mendelian randomization study. Frontiers in Endocrinology. 2024 Jan 8;14:1276836.
9. Ye Y, Xiong Y, Zhou Q, Wu J, Li X, Xiao X. Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: a retrospective cohort study. Journal of diabetes research. 2020;2020(1):4168340.
10. Artzi NS, Shilo S, Hadar E, Rossman H, Barbash-Hazan S, Ben- Haroush A, Balicer RD, Feldman B, Wiznitzer A, Segal E. Prediction of gestational diabetes based on nationwide electronic health records. Nature medicine. 2020 Jan;26(1):71-6.
11. Kang BS, Lee SU, Hong S, Choi SK, Shin JE, Wie JH, Jo YS, Kim YH, Kil K, Chung YH, Jung K. Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. Scientific Reports. 2023 Aug 16;13(1):13356.
12. Zhao M, Yao Z, Zhang Y, Ma L, Pang W, Ma S, Xu Y, Wei L. Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: a systematic review and meta-analysis. BMC Medical Informatics and Decision Making. 2025 Jan 13;25(1):18.
13. 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 Jan 3;25(1):3.
14. Mennickent D, Rodr⟨guez A, Far⟨as-Jofr⌈ M, Araya J, Guzm⟨n- Guti⌈rrez E. Machine learning-based models for gestational diabetes mellitus prediction before 24 28 weeks of pregnancy: A review. Artificial Intelligence in Medicine. 2022 Oct 1;132:102378.
15. Zhang R, Li Z, Xilifu N, Yang M, Dai Y, Zang S, Liu J. A nomogram to predict gestational diabetes mellitus: a multi-center retrospective study. Journal of Molecular Cell Biology. 2025 Mar 10:mjaf008.
16. https://www.kaggle.com/datasets/rasooljader/gestational- diabetes?resource=download
17. M. Sheykhmousa et al., "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., vol. 13, pp. 6308-6325, Sep. 25, 2020.
18. P. Schober and T. R. Vetter, "Logistic regression in medical research," Anesth. Analg., vol. 132, no. 2, pp. 365-366, Feb. 1, 2021.
19. B. Charbuty and A. Abdulazeez, Classification based on decision tree algorithm for machine learning, Journal of Applied Science and Technology Trends, vol. 2, no. 1, pp. 20 28, Mar. 2021.
20. G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," Adv. Neural

14 Inf. Process. Syst., vol. 30, Dec. 2017.
21. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, Aug. 13, 2016, pp. 785-79.