International Journal of Innovative Information Systems & Technology Research 14(3):430-447, July-Sept, 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org ISSN: 2467-8562

DOI:10.5281/zenodo.22934211

Early Prediction of Gestational Diabetes Mellitus Using Ensemble Machine Learning

[1]Ayodeji O. Ayeni, [2]Oluwatosin L. Ogbeide, [2*]Olatunde D. Akinrolabu, [3]Victor O. Otokiti

[1]Department of Information and Communication Technology,
Faculty of Computing,
Adekunle Ajasin University, Akungba Akoko, Ondo State, Nigeria

[2]Department of Data Science and Artificial Intelligence,
Faculty of Computing,
Adekunle Ajasin University, Akungba Akoko, Ondo State, Nigeria.

[3]Department of Computer Science,
Faculty of Computing,
Adekunle Ajasin University, Akungba Akoko, Ondo State, Nigeria.

Corresponding author: Olatunde David Akinrolabu (olatunde.akinrolabu@aaua.edu.ng)

ABSTRACT

Gestational Diabetes Mellitus (GDM) is a prevalent pregnancy complication with significant maternal and neonatal risks. Conventional screening methods, often applied late in pregnancy, limit opportunities for early intervention. Recent studies have demonstrated the promise of machine learning models such as Random Forest, XGBoost, and ensemble approaches for earlier and more accurate GDM prediction, though challenges such as class imbalance, interpretability, and reliance on extensive clinical data remain. This project applied a structured methodology to overcome these challenges. Data pre-processing involved removing duplicates and outliers, followed by feature engineering (Age × BMI, BMI^2) to enrich predictive features. Class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE). Three classifiers Logistic Regression, Random Forest, and XGBoost, were developed, with hyperparameter tuning performed to optimise performance. Ensemble strategies, specifically Hard Voting and Soft Voting, were then implemented to combine model strengths. Models were evaluated using Accuracy, Precision, Recall, F1-score, Confusion Matrix, and AUC, with recall emphasised due to its clinical importance. Results showed that feature engineering and SMOTE substantially improved sensitivity across models. While optimised XGBoost achieved very high recall, it sacrificed precision. The Soft Voting Ensemble achieved the most balanced performance, with 82.96% accuracy, 93.02% recall, and an AUC of 0.933, outperforming baseline models and aligning with or exceeding findings in related literature. This study concludes that ensemble learning, particularly Soft Voting, can support early GDM prediction using accessible demographic and anthropometric features. The work contributes by demonstrating that careful pre-processing, feature engineering, and balanced resampling can yield competitive results with limited data, offering a cost-effective decision-support tool for healthcare providers.

Keywords: gestational diabetes mellitus, machine learning, ensemble learning, SMOTE, feature engineering, ensemble voting, early prediction

1.0. INTRODUCTION

Gestational Diabetes Mellitus (GDM) is a condition in which a hormone made by the placenta prevents the body from using insulin effectively. GDM is characterised by high blood sugar levels that develop during pregnancy; glucose builds up in the blood instead of being absorbed by the cells. Unlike type 1 diabetes, gestational diabetes is not caused by a lack of insulin, but by other hormones produced during pregnancy that can make insulin less effective, a condition referred to as insulin resistance. In most cases, gestational diabetes symptoms disappear following delivery. However, women who had GDM are at increased risk of developing type 2 diabetes within 5–10 years after pregnancy and recurrence of GDM in future pregnancies. GDM often has mild or no noticeable symptoms, but some women may experience increased thirst and urination, fatigue, blurred vision, and slow healing of cuts or wounds, among other symptoms. Although any woman can develop GDM during pregnancy, some factors that may increase the risk include overweight or obesity, with a Body Mass Index above 30 a significant predictor [1]; age over 25 years [2]; having previously given birth to an infant weighing greater than 9 pounds [3]; family history of diabetes [4]; previous history of GDM; prediabetes (impaired glucose tolerance); and race/ethnicity, with higher prevalence in certain ethnic groups such as Hispanic, African American, and Native American populations. Although increased glucose in the urine is often included in the list of risk factors, it is not believed to be a reliable indicator of GDM [5].

According to the World Health Organisation (WHO), GDM affects approximately 9.4% of pregnant women worldwide. The American Diabetes Association recommends screening for undiagnosed type 2 diabetes at the first prenatal visit in women with diabetes risk factors [6]. In pregnant women not known to have diabetes, GDM testing should be performed at 24 to 28 weeks of gestation, and women with diagnosed GDM should be screened for persistent diabetes 6 to 12 weeks postpartum. It is also recommended that women with a history of GDM undergo lifelong screening for diabetes or prediabetes at least every three years. Treatment for gestational diabetes is determined by a woman's age, overall health, medical history, extent of the disease, and tolerance for specific medications, and focuses on keeping blood glucose levels in the normal range through a special diet, exercise, daily blood glucose monitoring, and, where necessary, insulin injections. It is essential for women who had GDM to maintain a healthy lifestyle, including a balanced diet and regular exercise, to reduce their risk of developing type 2 diabetes and other related health issues. Unlike type 1 diabetes, gestational diabetes generally occurs too late to cause birth defects, since these usually originate during the first trimester, while insulin resistance from placental hormones typically does not occur until approximately the 24th week. Untreated or poorly managed GDM can lead to complications for both mother and baby, including pre-eclampsia, macrosomia, neonatal hypoglycemia, low serum calcium and magnesium levels in the infant, increased risk of cesarean delivery, and preterm labor.

In recent years, artificial intelligence has been increasingly applied to predict GDM, leveraging machine learning (ML) algorithms and electronic health records (EHRs), with the goal of identifying high-risk women early in pregnancy to enable timely interventions. A study published in the Journal of Clinical Epidemiology used a random forest algorithm to predict GDM, achieving an AUC-ROC of 0.85 [7]; researchers applied an SVM algorithm to predict GDM, reporting an AUC-ROC of 0.83 [8]; and a study published in the IEEE Journal of Biomedical and Health Informatics used a neural network to predict GDM, achieving an accuracy of 86.2% [9]. Common features used in GDM prediction models include demographic data (age, ethnicity, family history of diabetes), anthropometric data (BMI, weight gain during pregnancy), clinical data (blood pressure, glucose levels, previous obstetric history), and lifestyle data (physical activity, diet, smoking status).

1.1. Problem Statement

Despite advancements in prenatal care, the incidence of GDM continues to rise. Early identification and management of GDM can significantly reduce maternal and neonatal complications. The increasing prevalence of gestational diabetes mellitus among pregnant women poses a significant risk to maternal

and fetal health. GDM is characterised by insulin resistance and impaired insulin secretion, and this pathophysiological change is exacerbated by hormonal fluctuations during pregnancy, particularly the increased production of human placental lactogen (hPL) and cortisol [10]. Research has shown that women with GDM have altered glucose and lipid metabolism, increasing the risk of adverse pregnancy outcomes. Establishing a predictive model using machine learning techniques can enhance early diagnosis and facilitate timely interventions [11].

2.0. Literature Review

Recent studies have increasingly demonstrated the potential of machine learning (ML) to improve early prediction of gestational diabetes mellitus (GDM), particularly through the use of ensemble and advanced learning algorithms. These studies generally show that ML can achieve strong predictive performance while identifying clinically relevant risk factors such as maternal age, body mass index (BMI), fasting glucose, and family history of diabetes.

Zheqing et al. [19] conducted a meta-analysis of ML models for GDM prediction and reported a pooled AUROC of 0.8891 for non-logistic-regression models. Their findings identified maternal age, BMI, and family history of diabetes as important predictors, demonstrating the potential of ML to improve GDM risk stratification.

Adefuwa et al. [1] developed an explainable machine-learning system for diabetes detection using four models, with XGBoost achieving the best performance at 91.99% accuracy and 94.38% AUC-ROC. The system showed relatively consistent performance across demographic groups, while LIME identified HbA1c and fasting glucose as the most important predictors. The model was also tested on the Pima Indians Diabetes Dataset, achieving 72.53% accuracy, demonstrating potential for generalisation. Overall, the study highlights the potential of explainable and inclusive machine learning for accessible diabetes detection.

Kumar et al. [20] investigated population-specific risk prediction using demographic and clinical information from more than 5,000 women. Their model achieved an AUC of 0.90, highlighting the importance of developing prediction models that reflect the characteristics of the target population rather than relying solely on generalised models.

Du et al. [21] developed an explainable ML-based clinical decision support system for GDM prediction and achieved an AUC of 0.87. Their study emphasised that predictive performance alone is insufficient for clinical adoption, as healthcare professionals also require models whose predictions can be understood and interpreted.

Wang et al. [22] applied an ensemble learning framework to GDM prediction using Tianchi competition data. The study combined LightGBM, XGBoost, and Gradient Boosting Tree through a genetic-algorithm-optimised ensemble, while SHAP was used to assess feature importance. The ensemble achieved an AUC above 0.90 and outperformed the individual models, demonstrating the value of combining complementary algorithms.

Zhang et al. [23] reviewed ML approaches for GDM prediction and found that XGBoost and Random Forest generally achieved better predictive performance than traditional statistical approaches, with reported AUC values ranging from 0.85 to 0.92. Their findings further support the effectiveness of tree-based ML methods for GDM risk prediction.

Belsti et al. [24] evaluated the CatBoost classifier using data from 48,502 singleton pregnancies in Australia. CatBoost achieved an AUC of approximately 0.93 and significantly outperformed traditional logistic regression, demonstrating the potential of gradient-boosting techniques for large-scale GDM prediction.

Kang et al. [25] examined GDM prediction among 34,387 Korean pregnancies using LightGBM and XGBoost. XGBoost achieved an AUC of 0.804 during the second trimester and performed better than LightGBM. The study also incorporated SHAP and the Boruta algorithm to improve feature selection and model interpretation, reinforcing the importance of combining predictive modelling with explainability.

Gupta et al. [26] reviewed recent developments in ML-based GDM prediction and reported consistent AUC values between 0.85 and 0.92 across XGBoost, Random Forest, and deep learning models. The review identified fasting glucose, BMI, and family history of diabetes as recurring predictors, suggesting that these variables have substantial predictive value across different modelling approaches and populations.

Li et al. [27] investigated the application of Long Short-Term Memory (LSTM) networks to electronic health record data involving 52,781 pregnancies in China. The LSTM model achieved an AUC of 0.92 and outperformed traditional logistic regression. Maternal BMI, fasting blood glucose, age, and blood pressure were identified as important predictors, demonstrating the potential of deep learning to exploit longitudinal clinical information.

Rao et al. [28] examined explainable AI for GDM prediction in a low-resource rural setting in India. Their XGBoost model achieved an AUC of 0.86 while maintaining high interpretability through SHAP analysis. Importantly, the study relied on relatively simple predictors because of limited laboratory resources, demonstrating that effective ML-based prediction can potentially be developed even where advanced diagnostic resources are unavailable.

The reviewed works demonstrate a consistent advantage of ML over conventional statistical approaches for GDM prediction, with XGBoost, Random Forest, CatBoost, and ensemble models frequently achieving AUC values between 0.85 and 0.92 or higher. Meta-analytic evidence further supports this trend, showing that ML models can provide strong discriminatory performance across different datasets and populations. Studies on postpartum Type 2 Diabetes Mellitus risk prediction also indicate that XGBoost and Random Forest can outperform logistic regression in identifying women with GDM who are at increased long-term risk of developing Type 2 diabetes. Despite these promising findings, several limitations remain evident across the literature. Differences in datasets, populations, predictor availability, class distributions, and validation procedures create challenges in comparing model performance and transferring models across settings. Class imbalance and limited population-specific validation are also recurring concerns, while the growing emphasis on explainability highlights the need to make ML predictions clinically interpretable. These limitations provide a strong justification for the present study's focus on appropriate data preprocessing, comparative evaluation of classical ML models, and ensemble learning to improve the reliability and generalisability of GDM prediction.

3.0. METHODOLOGY

This section describes the processes of data acquisition, pre-processing, feature engineering, handling of class imbalance, model development, and evaluation. These steps are critical in ensuring that the data is clean, reliable, and appropriately structured for machine learning algorithms, and the methodological framework is designed to guarantee reproducibility, transparency, and scientific rigour in the subsequent stages of model development and evaluation. Figure 3.1 presents the proposed architecture for GDM prediction adopted in this study.

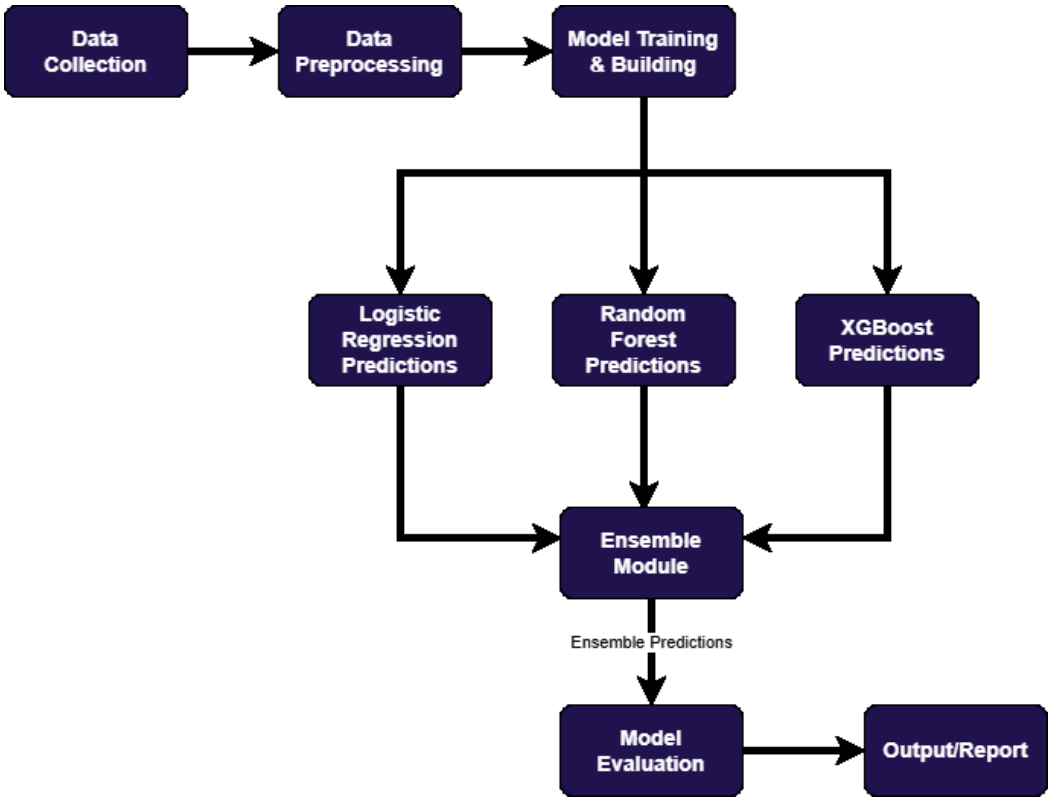

Figure 1: System Model for GDM Prediction

3.1. Data Acquisition and Pre-processing

Data acquisition and pre-processing constitute the foundational stage of the methodological framework, ensuring that the dataset is properly sourced, understood, and transformed into a form suitable for predictive modelling. The steps taken include dataset description, initial checks and cleaning, exploratory data visualisation, feature engineering, treatment of class imbalance, and feature scaling.

3.1.1. Data Source and Provenance

The dataset used in this study was obtained from Kaggle and contains clinical and demographic information related to gestational diabetes in women from laboratories within the Kurdistan region (available at: https://www.kaggle.com/datasets/rasooljader/gestational-diabetes). The dataset is openly accessible for research purposes and provides real-world clinical data suitable for predictive modelling.

3.1.2. Dataset Structure and Summary Statistics

The dataset comprises 1,012 records and 7 variables, including both predictor features and the target variable: Age (float, maternal age in years), Pregnancy No (float, number of pregnancies), Weight (float, maternal weight in kilograms), Height (float, maternal height in centimetres), BMI (float, body mass index calculated as weight in kg divided by height in m^2), Heredity (integer, family history of diabetes; 1 = Yes, 0 = No), and Prediction (integer, the target variable representing gestational diabetes status; 1 = Positive, 0 = Negative).

Table 1: Descriptive Statistics of the Dataset

Statistic	Age	Pregnancy No	Weight	Height	BMI	Heredity	Prediction
count	1012.000	1012.000	1012.000	1012.000	1012.000	1012.000	1012.000
mean	30.357	2.517	72.988	158.154	29.256	0.226	0.214
std	7.022	1.530	12.552	7.367	5.092	0.418	0.410
min	16.000	1.000	43.000	135.000	16.000	0.000	0.000
25%	24.000	1.000	64.750	153.000	25.700	0.000	0.000
50%	30.000	2.000	70.000	158.000	28.900	0.000	0.000
75%	36.000	3.000	80.000	164.000	32.000	0.000	0.000
max	45.000	9.000	126.000	196.000	54.300	1.000	1.000

Descriptive statistics revealed a mean maternal age of 30.36 years (range: 16–45) and a mean BMI of 29.26 (range: 16.0–54.3). The target variable is imbalanced, with approximately 21.4% positive cases ($n \approx 217$) and 78.6% negative cases ($n \approx 795$).

3.1.3. Data Cleaning

Prior to analysis, the dataset underwent essential cleaning steps to ensure reliability and consistency of results. First, duplicate entries were identified and removed, reducing the dataset size from 1,012 rows to 918 rows; retaining duplicates could have biased the learning algorithms by over-representing certain cases. Hence their removal was necessary to preserve data integrity. Next, outlier detection and removal were conducted on the numerical features (Age, Weight, Height, and BMI) using the Interquartile Range (IQR) method. This method calculates the first quartile (Q1) and third quartile (Q3), determines the interquartile range ($IQR = Q3 - Q1$), and defines lower and upper bounds as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively; data points falling outside these bounds were classified as outliers and removed. This procedure was crucial, as extreme values can distort statistical summaries and negatively influence model performance, particularly for algorithms sensitive to data distributions. After applying this procedure, the dataset size was further reduced from 918 rows to 892 rows, yielding a cleaner and more representative dataset for model training.

3.1.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the dataset and to uncover underlying trends, patterns, and potential challenges that could influence the modelling phase. Feature Distributions: Histograms comparing predictor variables across the two target classes (GDM and No GDM) revealed distinct patterns. Women diagnosed with GDM tended to fall within higher age brackets; higher pregnancy counts were slightly more common among women with GDM, though the difference was less pronounced; both weight and BMI showed elevated values among GDM cases, consistent with the well-established association between higher BMI and gestational diabetes risk; height did not differ substantially between the two groups, indicating limited discriminative power; and heredity displayed a clear separation, with a higher proportion of GDM cases associated with a family history of diabetes.

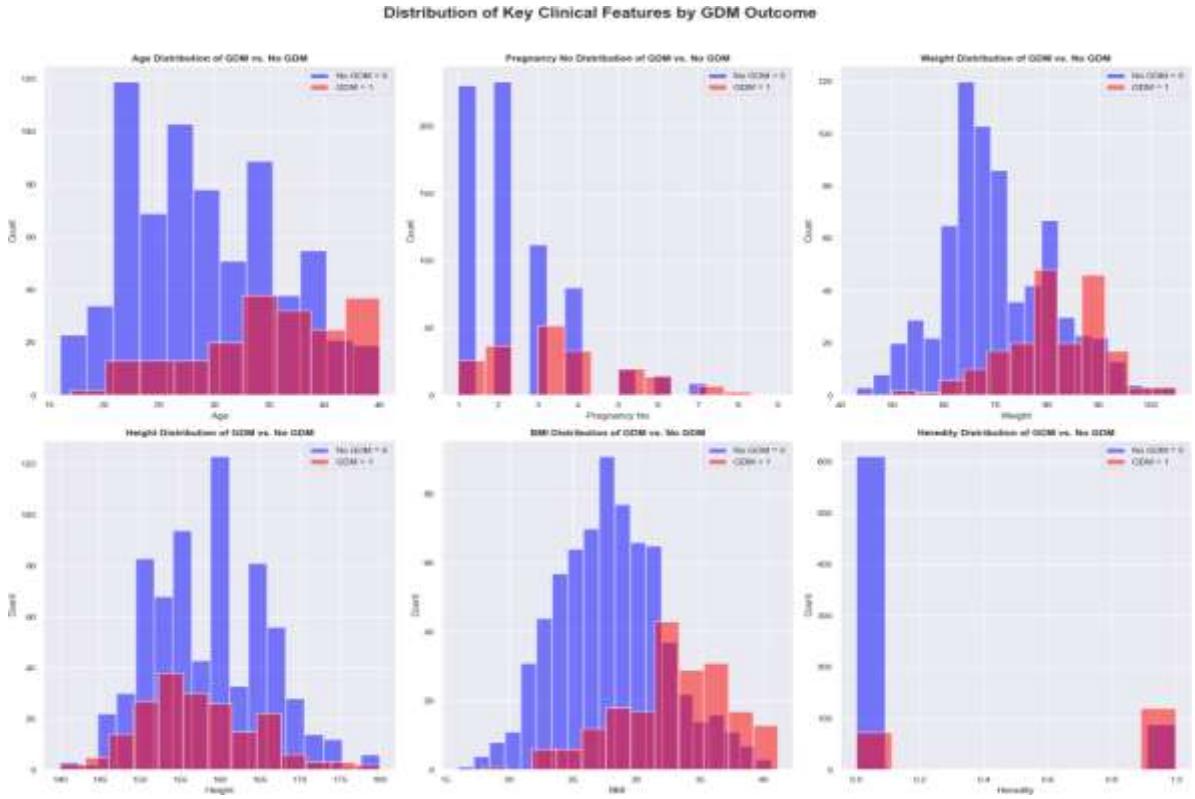

Figure 2: Distribution of Features across GDM and No GDM Classes

3.1.5. Feature Engineering

Feature engineering was performed to enhance the predictive capacity of the dataset. While the dataset includes individual attributes such as Age, BMI, Weight, Height, and Heredity, the interactions between these variables can provide more discriminative power. Two engineered features were created in this study to improve model performance:

a) Age × BMI (Interaction Term): This feature captures the combined effect of maternal age and body mass index, allowing the model to account for the compounded risk associated with being both older and overweight/obese, which may not be fully expressed when considering Age and BMI independently.

b) BMI^2 (Nonlinear Transformation): This feature represents the squared value of BMI, allowing the model to capture potential nonlinear relationships between BMI and the likelihood of GDM, since the risk of GDM may not increase linearly with BMI and very high BMI values could disproportionately increase risk.

These engineered features were integrated alongside the original variables in the training process. Overall, feature engineering in this project was not simply a mathematical exercise but a deliberate effort to incorporate domain knowledge from medical literature into the dataset, ensuring that the models could learn from both the raw measurements and the derived relationships relevant to gestational diabetes prediction.

3.1.6. Handling Class Imbalance

Exploratory Data Analysis revealed that the target variable was highly imbalanced, with 193 positive cases (GDM) compared to 699 negative cases (No GDM). Class imbalance is a critical issue in machine learning, as most algorithms assume a roughly balanced distribution of classes; in such situations, models often become biased toward the majority class, yielding high overall accuracy but poor sensitivity in

detecting the minority class. In this project, this would result in models frequently predicting "No GDM" while failing to identify actual cases of gestational diabetes, significantly reducing clinical usefulness. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was adopted. Unlike simple oversampling methods that duplicate existing minority samples, SMOTE generates synthetic instances by interpolating between minority class observations and their nearest neighbours in feature space, improving the diversity of the minority class, reducing overfitting risk, and enhancing generalisation. SMOTE was applied after feature engineering and feature scaling, but only on the training data, to prevent data leakage and ensure that evaluation metrics on the unseen test set remained unbiased. After applying SMOTE, the training data became balanced, with equal representation of both classes (519 GDM cases and 519 non-GDM cases), providing a stronger foundation for building models with improved recall and overall quality.

Figure 4: SMOTE Sampling

3.2. Model Development and Training

After preprocessing the dataset, the next step was to develop predictive models for GDM capable of distinguishing between positive (GDM = 1) and negative (GDM = 0) cases based on the provided features. This stage involved splitting the dataset into training and testing subsets, selecting appropriate algorithms, and optimising performance to achieve the best predictive accuracy and recall, in line with standard supervised learning best practices.

3.2.1. Dataset Splitting

To objectively evaluate model performance, the cleaned and pre-processed dataset was divided into training and testing subsets using a 75%–25% split: a training set of 669 samples (~75% of 892 total entries), used to train the models and perform hyperparameter tuning, and a testing set of 223 samples (~25% of 892 total entries), used exclusively to assess final model performance and generalisation ability. This strategy reduces the risk of overfitting and ensures that evaluation results more accurately reflect real-world applicability.

3.2.2. Description of Models

Three supervised machine learning models were selected for predicting GDM due to their complementary strengths in handling different data characteristics. Logistic Regression is a supervised learning technique that models a categorical (binary) target using the equation $\log(p(X) / (1 - p(X))) = \beta 0 + \beta 1X$, where $p(X)$ is the dependent variable, X the independent variable, $\beta 0$ the intercept, and $\beta 1$ the slope coefficient [29]. Input values are combined linearly and transformed via the logistic (sigmoid) function into a probability between 0 and 1. In this project, Logistic Regression predicts the likelihood of GDM based on maternal health features (Age, Pregnancy No, Weight, Height, BMI, Heredity) plus the engineered variables (Age × BMI, BMI^2), serving as a fast, interpretable baseline model. Random Forest is an ensemble learning method composed of multiple decision tree predictors, each trained on a random subset of the data and features [30]. By combining the outputs of many weak learners, Random Forests achieve strong predictive power while reducing overfitting risk, and are more robust in the presence of outliers and high-dimensional data [31], [32]. An important feature is the ability to assess variable importance via the Gini impurity criterion [33], [34]. In this project, Random Forest classifies patients based on the same maternal and engineered features, capturing complex, non-linear relationships while resisting overfitting. Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosted decision trees that builds models sequentially, with each new model correcting the errors of the previous ones. XGBoost incorporates a regularisation term that penalises model complexity, implements efficient handling of missing values and sparsity, and uses parallelised tree construction for computational efficiency [36]. For

this binary classification task, XGBoost minimises a logistic loss function and applies boosting iterations to refine the classification boundary, while also providing feature importance to indicate which maternal and clinical variables contribute most significantly to prediction. XGBoost is particularly relevant here for its ability to model complex non-linear relationships, handle class imbalance effectively when combined with SMOTE, and maintain high predictive accuracy while avoiding overfitting.

3.2.3. Hyperparameter Tuning

Hyperparameter tuning was conducted to optimise the performance of the predictive models. Unlike model parameters learned directly from the data, hyperparameters are external configurations that control the learning process; poorly chosen configurations can lead to underfitting or overfitting. Hyperparameter tuning was carried out using Grid Search with cross-validation, which systematically evaluates all possible combinations within a predefined set of hyperparameter values and selects the configuration yielding the best performance on the validation folds. Since Logistic Regression was already well-suited to the dataset and computationally efficient, it was retained with default parameters. Hyperparameter optimisation was instead applied to Random Forest, tuning the number of estimators (trees), maximum tree depth, minimum samples required to split an internal node, and the number of features considered for the best split; and to XGBoost, tuning the number of estimators, learning rate, maximum tree depth, and subsampling ratio. The best-performing hyperparameter sets identified during Grid Search were used to train the final models.

3.2.4. Ensemble Learning (Hard and Soft Voting)

In addition to training individual models, an ensemble approach was adopted to combine the strengths of multiple classifiers, since ensemble learning integrates the predictions of different models to achieve better performance and generalization than any single model alone [37], [38]. Both Hard Voting and Soft Voting ensembles were constructed using Logistic Regression, Random Forest, and XGBoost as base learners. Hard Voting aggregates the predicted class labels from each model and selects the majority vote as the final prediction; for binary classification, a sample is classified as positive (GDM) if at least two of the three models predict it as such. This leverages the diversity of the individual models: Logistic Regression provides a strong statistical baseline, Random Forest offers robustness against noise and feature interactions, while XGBoost contributes gradient-boosted precision [39]. Soft Voting averages the predicted probabilities from each model rather than relying solely on discrete class labels, allowing models with higher confidence to exert greater influence on the final decision, often resulting in improved performance compared to hard voting [40]. The rationale for employing these ensemble strategies lies in their ability to mitigate the weaknesses of individual models, since Logistic Regression may underperform on non-linear relationships while Random Forest and XGBoost, though powerful, may sometimes overfit. From a clinical perspective, the use of ensembles is particularly valuable, as they reduce the likelihood of misclassification in sensitive medical diagnoses such as GDM; Soft Voting in particular incorporates the probability estimates of each model, improving sensitivity to minority class cases and supporting more accurate identification of at-risk patients [41], [42]. The study evaluated the machine learning models using Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and AUC to provide a comprehensive assessment of performance. Because GDM prediction is a binary medical classification task and may involve class imbalance, relying on accuracy alone could be misleading. Recall (Sensitivity) was prioritized because failing to identify a true GDM case can have serious consequences for the mother and child. Precision assessed the correctness of positive predictions, while F1-Score balanced precision and recall. The Confusion Matrix provided insight into false positives and false negatives, and AUC measured the models' ability to distinguish between GDM and non-GDM cases across different thresholds.

4.0. RESULTS AND DISCUSSION

4.1. Baseline Model Performance

As an initial step, the dataset was used to train baseline models without the application of feature engineering, class balancing, or hyperparameter optimisation, establishing benchmark performance metrics against which later optimised models could be compared. The Logistic Regression model achieved relatively high accuracy (91.03%) but showed limitations in recall (62.79%), correctly classifying 176 non-GDM cases and 27 GDM cases while misclassifying 16 true GDM cases as non-GDM and 4 non-GDM cases as GDM. This indicates that while the model correctly identified the majority of negative cases, it failed to capture a substantial proportion of positive GDM cases, and the relatively high accuracy came at the cost of missing a substantial number of positive GDM cases.

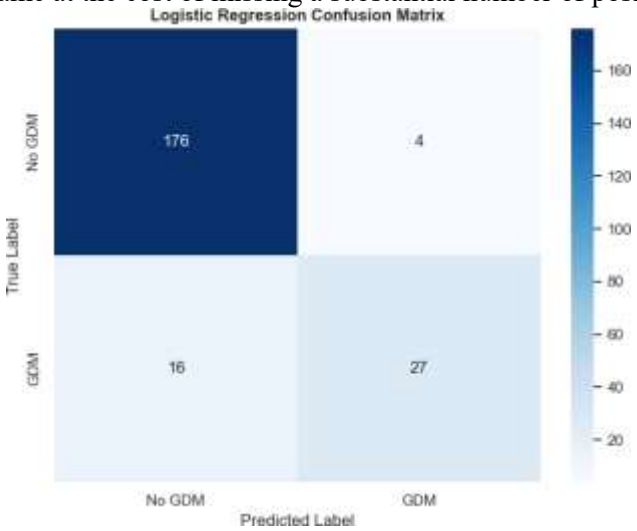

Figure 5: Confusion Matrix for Logistic Regression (Baseline*)*

The baseline Random Forest classifier produced an accuracy of 87.89%, with moderate precision (76.67%) and recall (53.49%), correctly classifying 173 non-GDM cases and 23 GDM cases while misclassifying 20 GDM cases as non-GDM and 7 non-GDM cases as GDM. Compared to Logistic Regression, Random Forest demonstrated a slightly better balance between precision and recall but still missed a considerable number of positive cases, highlighting the model's difficulty in capturing minority class instances despite its robustness.

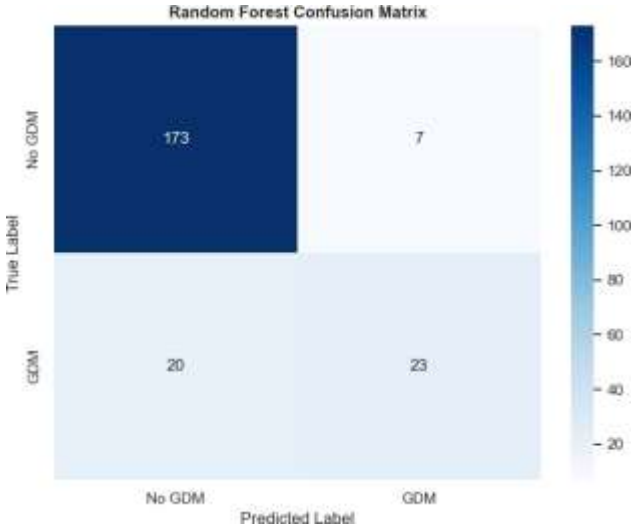

Figure 6: Confusion Matrix for Random Forest (Baseline)

The XGBoost model yielded an accuracy of 86.10%, precision of 65.79%, and recall of 58.13%, correctly classifying 167 non-GDM cases and 25 GDM cases while misclassifying 18 GDM cases as non-GDM and 13 non-GDM cases as GDM. Its performance was slightly better than Random Forest in terms of recall but still insufficient for clinical use, achieving better sensitivity to GDM cases than Logistic Regression, though at the cost of misclassifying more negative cases.

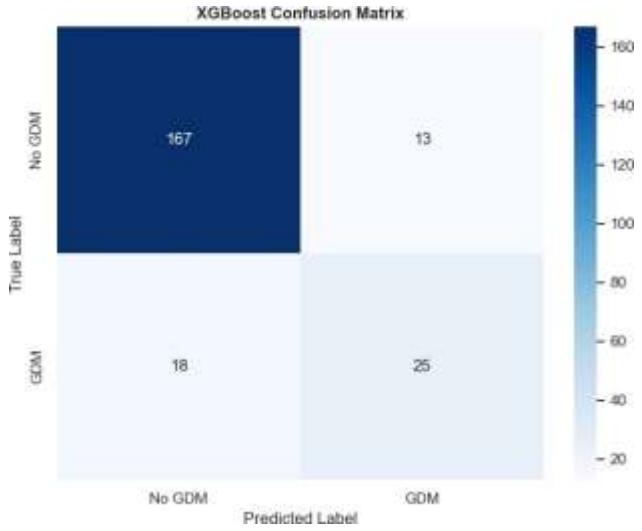

Figure 7: Confusion Matrix for XGBoost (Baseline)

The baseline performance of the three models demonstrated clear trade-offs between accuracy, precision, and recall, highlighting the challenges of predicting GDM in an imbalanced dataset. Logistic Regression produced the highest overall accuracy and precision but achieved only moderate recall, meaning more than one-third of true GDM cases were missed a problematic limitation in a clinical setting, since undetected GDM cases could lead to severe maternal and neonatal complications. Random Forest achieved lower accuracy and precision but slightly more balance, though its recall remained unsatisfactory. XGBoost yielded the lowest accuracy among the three models but provided a more favourable balance between precision and recall, showing somewhat greater sensitivity to positive GDM cases at the cost of misclassifying more non-GDM cases as positive - a trade-off that is more clinically acceptable, since false positives can be resolved through confirmatory testing while false negatives pose greater health risks.

Table 2: Comparative Performance of Baseline Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	91.03	87.10	62.79	72.97
Random Forest	87.00	73.33	51.16	60.27
XGBoost	86.10	65.79	58.14	61.73

The baseline results established clear performance limitations, particularly in terms of recall, which is critical in clinical prediction tasks such as GDM diagnosis. These findings emphasised the need for further improvements through feature engineering, class balancing with SMOTE, and hyperparameter optimisation.

4.2. Optimised Model Performance

Following the baseline evaluation, additional pre-processing techniques were implemented to enhance model performance, including the creation of the engineered features (Age × BMI and BMI^2), balancing the dataset with SMOTE, and hyperparameter tuning for Random Forest and XGBoost, aiming to improve sensitivity to minority cases while maintaining acceptable overall performance. After

incorporating feature engineering and applying SMOTE, Logistic Regression demonstrated a substantial improvement in recall, increasing to 88.37% from 62.79% in the baseline model, making it far more sensitive in detecting GDM cases; this improvement came at the expense of precision, which dropped to 60.32%, reflecting a higher number of false positives, while accuracy stood at 86.55% and the F1-score improved to 71.70%. The confusion matrix shows that Logistic Regression correctly classified 155 non-GDM cases and 38 GDM cases, missing only 5 positive cases, though 25 false positives explain the drop in precision.

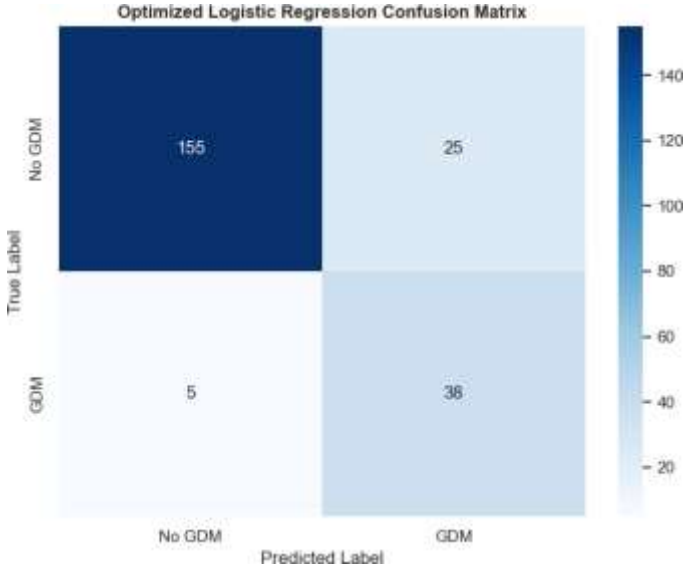

Figure 8: Confusion Matrix for Logistic Regression (Optimised)

Random Forest was further optimised using Grid Search, achieving an accuracy of 83.41%, a precision of 56.25%, and a recall of 62.79%. While accuracy slightly decreased compared to the baseline (86.99%), recall improved modestly, and the F1-score was 59.34%, a more balanced trade-off. The model correctly classified 159 non-GDM cases but struggled more than Logistic Regression in detecting GDM cases, with 16 false negatives and 21 false positives, though the improvement in recall compared to baseline suggests the model benefited from SMOTE and hyperparameter tuning.

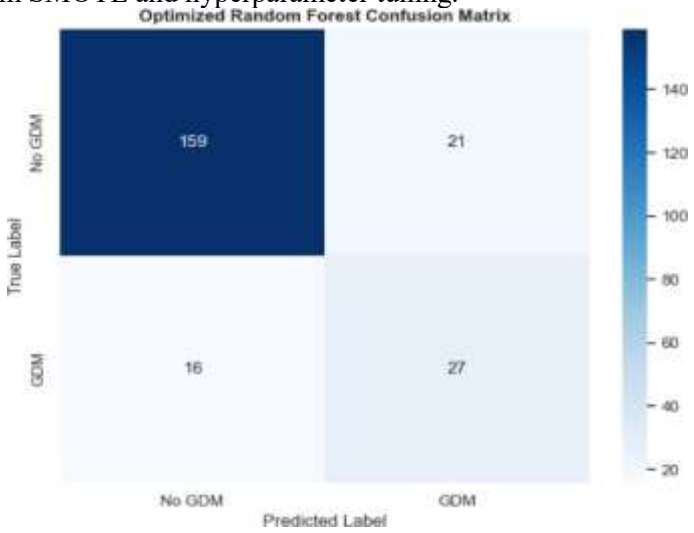

Figure 9: Confusion Matrix for Random Forest (Optimised)

Hyperparameter tuning for XGBoost produced unusual results: recall reached 100%, meaning all positive GDM cases were detected, but this came at the cost of an extremely low accuracy of 23.32% and precision of 20.09%. The model correctly classified only 9 non-GDM cases and 43 GDM cases while misclassifying 171 non-GDM cases as GDM; the F1-score was 33.46%, reflecting poor balance despite perfect sensitivity. Although such recall is theoretically ideal in clinical diagnosis, the overwhelming number of false positives indicates the model predicted GDM for almost all patients regardless of actual status, rendering it impractical, as it would lead to unnecessary interventions and diagnostic testing in real-world clinical practice.

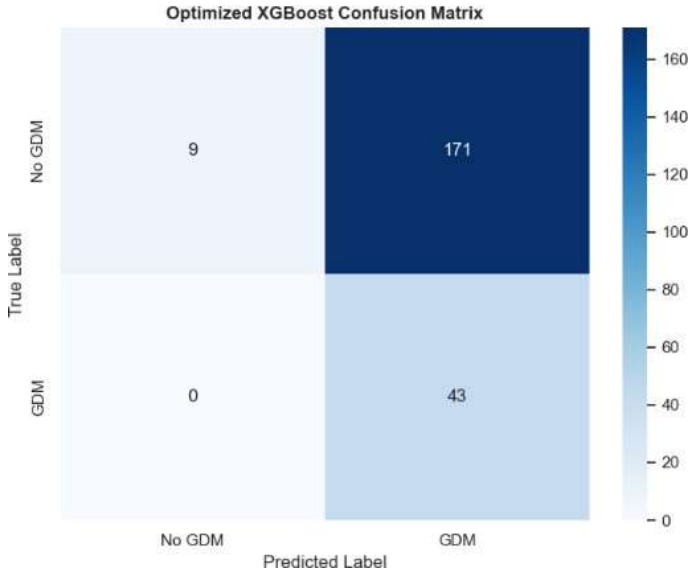

Figure 10: Confusion Matrix for XGBoost (Optimised)

The optimised models demonstrated diverse strengths and weaknesses: Logistic Regression offered the best balance, with strong recall and reasonable accuracy, though precision declined; Random Forest delivered moderate improvements in recall while maintaining good stability; and XGBoost achieved perfect recall but suffered from critically low accuracy and precision, leading to an impractical number of false positives.

Table 3: Comparative Performance of Optimised Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	86.55	60.32	88.37	71.70
Random Forest	83.41	56.25	62.79	59.34
XGBoost	23.32	20.09	100.00	33.46

4.3. Ensemble Model Performance

To further enhance predictive performance, ensemble learning techniques were applied by combining Logistic Regression, Random Forest, and XGBoost, using both Hard Voting and Soft Voting strategies to explore whether integrating the strengths of multiple models could yield more robust results. The Hard Voting Ensemble achieved an accuracy of 82.96%, with a precision of 53.42%, recall of 90.70%, and F1-score of 67.24%, correctly classifying 146 non-GDM cases and 39 GDM cases while misclassifying 34 non-GDM cases as GDM and 4 GDM cases as non-GDM. Compared to the optimised individual models, the ensemble demonstrated a clear improvement in recall, making it highly sensitive to GDM cases, though precision remained relatively low, reflecting a higher tendency to misclassify non-GDM patients as GDM.

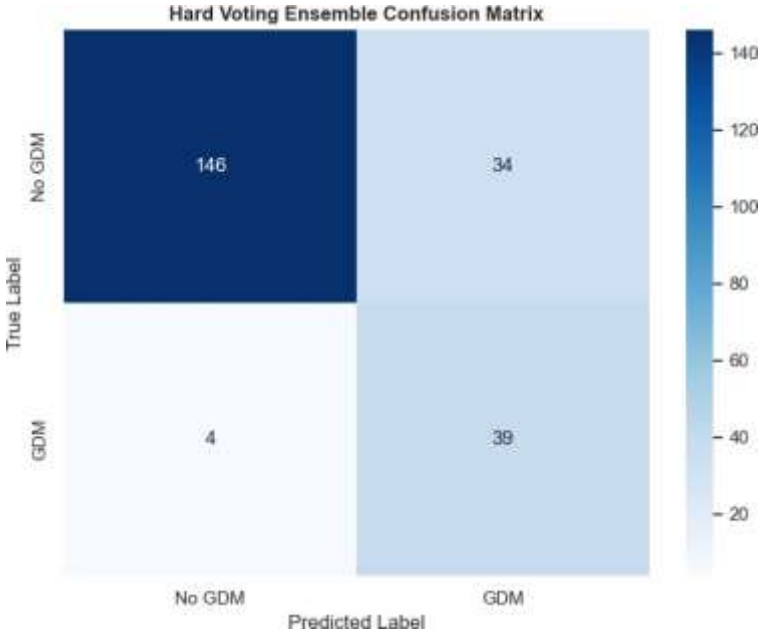

Figure 11: Confusion Matrix for Hard Voting Ensemble

The Soft Voting Ensemble achieved an accuracy of 82.96%, precision of 53.33%, recall of 93.02%, and F1-score of 67.80%, correctly classifying 145 non-GDM cases and 40 GDM cases while misclassifying 35 non-GDM cases as GDM and only 3 GDM cases as non-GDM. It slightly outperformed the Hard Voting Ensemble in terms of recall and F1-score, suggesting that probability averaging provided a better balance between sensitivity and stability, and achieved the highest recall among all approaches, ensuring that nearly all true GDM cases were detected.

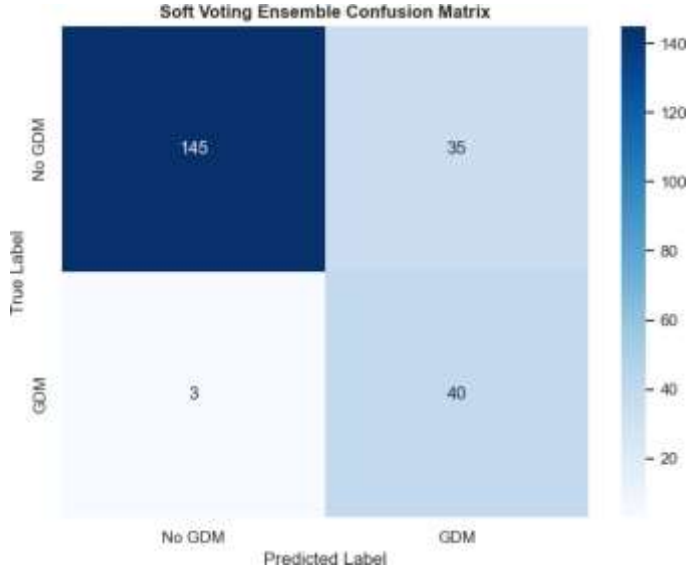

Figure 12: Confusion Matrix for Soft Voting Ensemble

ROC curve analysis further confirmed the strength of the models: Logistic Regression and the Soft Voting Ensemble both achieved an AUC of 0.933, indicating excellent discriminatory power, while Random Forest and XGBoost achieved slightly lower but still strong AUC values of 0.901 and 0.909, respectively.

Figure 13: ROC Curves for LR, RF, XGBoost, and Soft Voting Ensemble

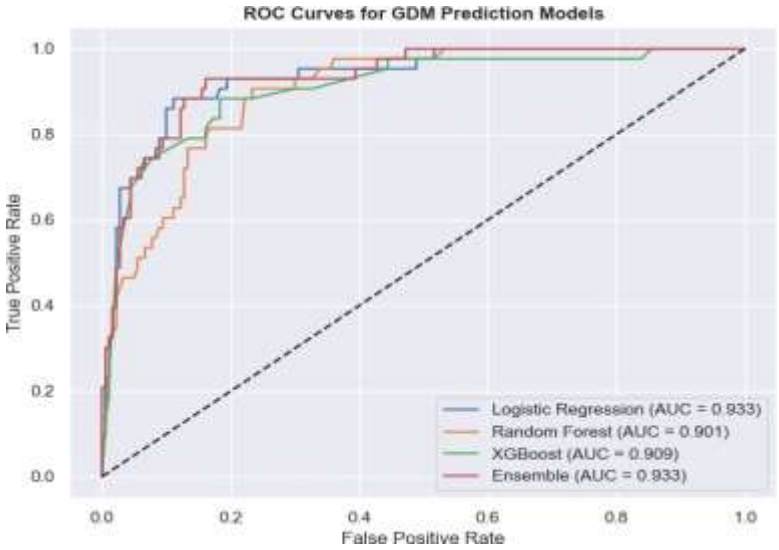

Table 4: Performance of Ensemble Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Hard Voting Ensemble	82.96	53.42	90.70	67.24	–
Soft Voting Ensemble	82.96	53.33	93.02	67.80	0.93

Both Hard and Soft Voting Ensembles demonstrated notable improvements in recall compared to the individual models, confirming the effectiveness of combining classifiers in sensitive clinical applications. While precision remained relatively low in both approaches, this trade-off is clinically acceptable, as false negatives (missed GDM cases) pose a far greater risk than false positives. Among the two approaches, the Soft Voting Ensemble emerged as the most effective, achieving the highest recall (93.02%) and the strongest balance of performance metrics, highlighting its potential as a clinically useful predictive tool for early GDM detection.

4.4. Final Model Selection and Discussion

The principal objective of this study is early detection of GDM. In clinical screening tasks, the cost of false negatives (missed GDM cases) is typically higher than the cost of false positives; therefore, model selection prioritised sensitivity (recall) and overall discriminative ability (AUC), while still accounting for accuracy and precision. Models were assessed on the held-out test set using accuracy, precision, recall, F1-score, and AUC where available, with the final selection emphasising clinical usefulness: the model must capture as many true GDM cases as possible while remaining reasonably specific.

Table 5: Overview of Results

Model (Stage)	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUC**
Logistic Regression – baseline	91.03	87.10	62.79	72.97	–
Random Forest – baseline	87.00	73.33	51.16	60.27	–
XGBoost – baseline	86.10	65.79	58.14	61.73	–
Logistic Regression – optimised (FE + SMOTE)	86.55	60.32	88.37	71.70	0.93

Model (Stage)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
Random Forest – optimised (tuned)	83.41	56.25	62.79	59.34	0.90
XGBoost – optimised (tuned)	23.32	20.09	100.00	33.46	0.91
Hard Voting Ensemble	82.96	53.42	90.70	67.24	–
Soft Voting Ensemble	82.96	53.33	93.02	67.80	0.93

The recall comparison demonstrates a clear improvement in sensitivity as models progressed from baseline to optimised forms. Baseline models achieved recall scores between 51% and 63%, indicating poor detection of GDM cases. After optimisation, Logistic Regression and Random Forest improved considerably, while XGBoost reached a perfect recall of 100%; however, this came at the cost of very low precision and overall accuracy, suggesting overfitting and limited clinical reliability. In contrast, the Soft Voting Ensemble achieved a recall of 93.02% alongside balanced performance across other metrics, making it the most clinically dependable option.

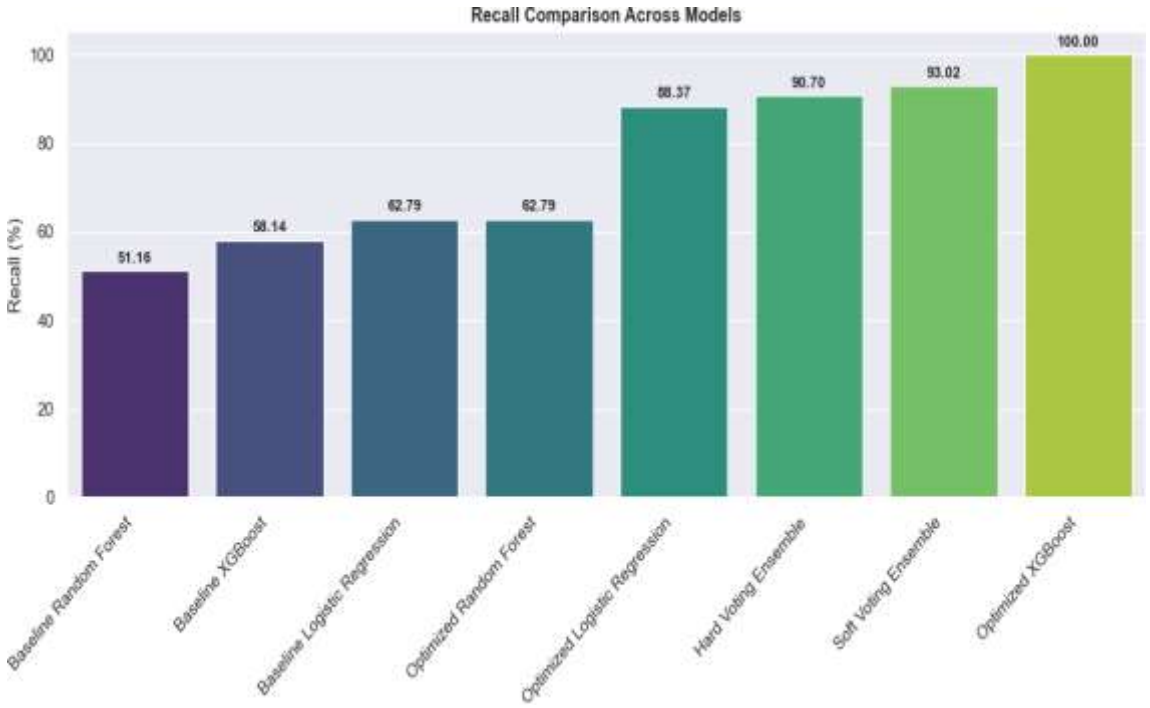

Figure 14: Recall Comparison across Baseline, Optimised, and Ensemble Models

4.5.1. Clinical Significance of Model Results

The results underscore the critical trade-off between accuracy and recall in medical prediction tasks. Among baseline models, Logistic Regression showed the highest accuracy (91.03%), but its recall (62.79%) was unsatisfactory for clinical use, as many GDM cases would be missed; Random Forest and XGBoost also struggled with recall despite reasonable accuracy. Among optimised models, Logistic Regression demonstrated a significant recall improvement (88.37%), confirming the importance of SMOTE and feature engineering, while optimised XGBoost achieved perfect recall (100%) but at the cost of very low precision (20.09%) and poor accuracy (23.32%), making it clinically unreliable. Among ensemble models, both Hard and Soft Voting Ensembles balanced recall and precision better than

individual models, with the Soft Voting Ensemble achieving the highest recall (93.02%) while maintaining acceptable F1 (67.80%) and AUC (0.933), making it the most clinically relevant model for GDM detection, where minimising false negatives is crucial.

4.5.2. Comparison with Related Works

The findings of this project are consistent with, and in some respects exceed, those reported in prior studies on GDM prediction. [22] achieved an AUC greater than 0.90 using large datasets, diverse clinical features, and engineered feature interactions; despite a smaller dataset, the Soft Voting Ensemble in this study reached a comparable AUC (0.933) through a similar strategy of feature transformation (Age × BMI, BMI^2), underscoring the effectiveness of thoughtful pre-processing and feature engineering. Ref [21] and [28] highlighted the importance of prioritising recall (sensitivity) and interpretability for clinical decision-making, where missing true GDM cases carries significant risk; the Soft Voting Ensemble developed here directly supports this need by achieving a recall of 93.02% while remaining interpretable through feature importance analysis. The meta-analysis in [23] reported typical AUC values of 0.85–0.92, a benchmark this project's ensemble model slightly exceeded (AUC = 0.933) even without biochemical markers. [24], using CatBoost and a large dataset, achieved an AUC of 0.93; this project achieved a nearly identical AUC with fewer features and a smaller dataset, suggesting that well-designed feature engineering and balanced resampling can partly substitute for dataset size or complexity.

5.0. Conclusion

This project successfully developed and evaluated machine learning models for the prediction of Gestational Diabetes Mellitus (GDM) using only demographic and anthropometric features. Through systematic pre-processing, feature engineering, and model optimisation, the study demonstrated that robust predictive performance can be achieved without reliance on biochemical data. Among the evaluated models, the Soft Voting Ensemble emerged as the most effective, achieving 93.02% recall and an AUC of 0.933, indicating a strong ability to correctly identify women at risk of GDM, with discriminative power comparable to, and in some cases exceeding, those reported in the literature that employed larger datasets or biochemical features. The focus on recall was intentional, as minimising false negatives is critical in clinical settings where failing to detect GDM can lead to severe complications for both mother and child. By achieving high sensitivity alongside competitive overall performance, the final model supports the broader goal of early, accessible, and reliable GDM risk prediction in resource-constrained healthcare environments. The study therefore concludes that Soft Voting strategies offer a practical and effective solution for GDM prediction when supported by careful pre-processing and feature engineering.

REFERENCES

Adefuwa, O. H., Akinrolabu, O. D., & Ayeni, A. O. (2026). An inclusive diabetes detection system using machine learning and explainable artificial intelligence. *Communications on Applied Electronics, 8*(2), 9–21. https://doi.org/10.5120/cae1e52adc932f2

American Diabetes Association. (2021). Standards of medical care in diabetes 2021. *Diabetes Care, 44*(Supplement 1), S1–S232. https://doi.org/10.2337/dc21-SINT

American Journal of Obstetrics and Gynaecology. (2011). Prior macrosomic birth as a risk factor for gestational diabetes mellitus.

Belsti, Y., Lake, E. A., Kassa, G. M., Abate, M. W., & Haile, T. G. (2023). Predicting gestational diabetes mellitus using CatBoost classifier on Australian population data.

Breiman, L. (2001). Random forests. *Machine Learning, 45*. https://doi.org/10.1023/A:1010933404324

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms.

Ceriani, L., & Verme, P. (2012). The origins of the Gini index.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chen, T., et al. (2018). Random forest algorithm for predicting gestational diabetes mellitus. *Journal of Clinical Epidemiology*.

Dabelea, D., et al. (2000). Increasing prevalence of gestational diabetes mellitus. *Diabetes Care*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Du, J., Sun, L., Wang, Y., & Li, X. (2022). Development of an explainable ML-based clinical decision support system for GDM.

Gao, X., et al. (2020). Risk prediction model for gestational diabetes mellitus.

García-Gil, D., Luengo, J., García, S., & Herrera, F. (2018). Enabling smart data: Noise filtering in Big Data classification. *Information Sciences, 479*, 135–152. https://doi.org/10.1016/j.ins.2018.02.060

Gupta, A., Sharma, R., Verma, P., & Singh, M. (2023). A review of machine learning approaches in GDM prediction: Accuracy, integration, and interpretability.

Hopkins Medicine. (2025). *Gestational diabetes*. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/gestational-diabetes

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

Kampmann, U., Madsen, L. R., Skajaa, G. O., Iversen, D. S., Moeller, N., & Ovesen, P. (2015). The impact of insulin resistance and placental hormones on gestational diabetes mellitus.

Kang, B. S., Lee, J. H., Park, S. Y., Kim, H. J., & Choi, Y. J. (2023). Machine learning models for GDM prediction in Korean women using XGBoost and LGBM.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal, 13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Liu, Y., et al. (2020). Neural network model for predicting gestational diabetes mellitus. *IEEE Journal of Biomedical and Health Informatics*.

Li, X., Zhou, Y., Wang, H., Chen, J., & Liu, Z. (2023). LSTM-based deep learning model for predicting gestational diabetes using EHRs in China.

Menze, B. H., et al. (2009). A comparison of random forest and its Gini importance with standard chemometric methods.

McIntyre, H. D., et al. (2019). Gestational diabetes mellitus. *Nature Reviews Disease Primers*.

Polikar, R. (2006). Ensemble-based systems in decision making. *IEEE Circuits and Systems Magazine, 6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Rao, P., Mehta, S., Iyer, R., & Deshpande, A. (2023). Explainable AI for GDM prediction in low-resource settings using XGBoost and SHAP.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*(1–2), 1–39. https://doi.org/10.1007/s10462-009-9124-7

Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review. *Frontiers in Ageing Neuroscience, 9*.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures.

Wang, X., Wang, Y., Zhang, S., Yao, L., & Xu, S. (2022). Analysis and prediction of gestational diabetes mellitus by the ensemble learning method. *International Journal of Computational Intelligence Systems, 15*(1), Article 72. https://doi.org/10.1007/s44196-022-00110-8

Wang, Y., et al. (2019). Support vector machine algorithm for predicting gestational diabetes mellitus.

Wang, X., et al. (2012). Family history and risk of gestational diabetes mellitus.

Yuen, L., Wong, V. W., & McElduff, A. (2018). Early detection of gestational diabetes using machine learning-based risk models.

Zhang, Y., Chen, L., Wu, Q., & Zhao, H. (2022). Meta-analysis of ML models for GDM: A comparison of XGBoost, RF, and deep learning techniques.

Zheqing, Z., Luqian, Y., Wentao, H., Yaoyu, W., Linhui, Z., Kui, J., Liu, Y., & Wu, H. (2022). Machine learning prediction models for gestational diabetes mellitus: Meta-analysis. *Journal of Medical Internet Research, 24*(3), e26634. https://pubmed.ncbi.nlm.nih.gov/35294369/

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. Chapman and Hall/CRC.