

Enhancing Early Prediction of Gestational Diabetes Mellitus through Data Augmentation and Feature Guidance

Xiekun Chen, Zhifa Jiang, Dong Su, Xiaoping Chen, Aiping Chen, Zhen Zhang, Huabin Wang

Submitted to: JMIR Medical Informatics
on: October 06, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Enhancing Early Prediction of Gestational Diabetes Mellitus through Data Augmentation and Feature Guidance

Xiekun Chen^{1*}; Zhifa Jiang^{2*}; Dong Su³; Xiaoping Chen⁴; Aiping Chen⁴; Zhen Zhang¹; Huabin Wang⁵

¹ School of Computer Science and Engineering Huizhou University Huizhou CN

² Obstetrics and Gynaecology Huizhou First Maternal and Child Health Care Hospital Huizhou CN

³ School of Mathematics and Statistics Huizhou University Huizhou CN

⁴ Huizhou First Maternal and Child Health Care Hospital Huizhou CN

⁵ Huizhou University Huizhou CN

*these authors contributed equally

Corresponding Author:

Zhen Zhang

School of Computer Science and Engineering
Huizhou University
No. 46 Yanda Avenue
Huizhou
CN

Abstract

Background: Early prediction of gestational diabetes mellitus (GDM) is critical for improving maternal health outcomes. However, predictive models are often challenged by limited early-pregnancy samples, severe class imbalance in datasets, and complex interrelationships among clinical features.

Objective: This study aimed to develop and evaluate a unified dual-dimensional enhancement framework that integrates data augmentation and feature engineering to improve early GDM prediction performance by addressing data imbalance and leveraging medical prior knowledge.

Methods: We proposed a framework combining Generative Adversarial Network (GAN)-based data augmentation with Large Language Model (LLM)-inspired feature engineering. GAN sampling was used to generate clinically plausible synthetic minority class samples to mitigate data imbalance. The LLM was guided to organize features into domains (e.g., basic demographics, metabolic syndrome, core liver biomarkers) and generate higher-order composite features, integrating medical prior knowledge. Machine learning models were subsequently developed, and interpretability analyses were performed using SHAP to identify key predictors.

Results: The Random Forest model enhanced by TVAE-based feature augmentation demonstrated the best performance. On the test dataset, it achieved a recall of 0.7559, an accuracy of 0.8444, and an area under the receiver operating characteristic curve (AUROC) of 0.8873. SHAP analysis identified the following five features as the most influential predictors: fasting blood glucose, the composite feature (fasting blood glucose + triglycerides) $\times$ pre-pregnancy BMI, activated partial thromboplastin time, leukocyte count, and neutrophil count.

Conclusions: The proposed dual-dimensional enhancement framework effectively alleviates data limitations and captures complex feature interactions in early GDM prediction. This strategy not only improves model performance, particularly in recall, but also provides interpretable biological evidence to support rapid clinical screening, stratified management, and early intervention in pregnancy.

(JMIR Preprints 06/10/2025:85335)

DOI: <https://doi.org/10.2196/preprints.85335>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/>



Original Manuscript

Enhancing Early Prediction of Gestational Diabetes Mellitus through Data Augmentation and Feature Guidance

Xiekun Chen¹, Zhifa Jiang^{2#}, Dong Su³, Xiaoping Chen², Aiping Chen², Zhen Zhang^{*1} and Huabin Wang^{*1}

1. School of Computer Science and Engineering, Huizhou University, Huizhou, Guangdong, China.
2. Obstetrics and Gynaecology, Huizhou First Maternal and Child Health Care Hospital, Huizhou, Guangdong, China
3. School of Mathematics and Statistics, Huizhou University, Huizhou, Guangdong, China.

Contributing authors: xminck@163.com; jiangzf@hzu.edu.cn; dsu186666@outlook.com; chaipivy@qq.com; Vera_421@163.com; zzsjbme@sjtu.edu.cn; whuab@163.com.

* Corresponding author(s). E-mail(s): zzsjbme@sjtu.edu.cn.

Co-first author, * Co-corresponding author

Abstract

Background: Early prediction of gestational diabetes mellitus (GDM) is critical for improving maternal health outcomes. However, predictive models are often challenged by limited early-pregnancy samples, severe class imbalance in datasets, and complex interrelationships among clinical features.

Objective: This study aimed to develop and evaluate a unified dual-dimensional enhancement framework that integrates data augmentation and feature engineering to improve early GDM prediction performance by addressing data imbalance and leveraging medical prior knowledge.

Methods: We proposed a framework combining Generative Adversarial Network (GAN)-based data augmentation with Large Language Model (LLM)-inspired feature engineering. GAN sampling was used to generate clinically plausible synthetic minority class samples to mitigate data imbalance. The LLM was guided to organize features into domains (e.g., basic demographics, metabolic syndrome, core liver biomarkers) and generate higher-order composite features, integrating medical prior knowledge. Machine learning models were subsequently developed, and interpretability analyses were performed using SHAP to identify key predictors.

Results: The Random Forest model enhanced by TVAE-based feature augmentation demonstrated the best performance. On the test dataset, it achieved a recall of 0.7559, an accuracy of 0.8444, and an area under the receiver operating characteristic curve (AUROC) of 0.8873. SHAP analysis identified the following five features as the most influential predictors: fasting blood glucose, the composite feature (fasting blood glucose + triglycerides) $\times$ pre-pregnancy BMI, activated partial thromboplastin time, leukocyte count, and neutrophil count.

Conclusion: The proposed dual-dimensional enhancement framework effectively alleviates data limitations and captures complex feature interactions in early GDM prediction. This strategy not only

improves model performance, particularly in recall, but also provides interpretable biological evidence to support rapid clinical screening, stratified management, and early intervention in pregnancy.

Keywords

Gestational Diabetes Mellitus (GDM); Data Augmentation; Feature Enhancement; Machine Learning; SHAP analysis

1. Introduction

Gestational diabetes mellitus (GDM), defined as glucose intolerance first detected during pregnancy, has a persistently high incidence and typically occurs in the second or third trimester, leading to adverse maternal and fetal outcomes [1]. Women with GDM face increased risks of gestational hypertension, infections, postpartum hemorrhage, preterm birth, and later progression to type 2 diabetes mellitus (T2DM) [2]. However, efficient identification and management remain challenging due to predictive uncertainty [3]. In China, the prevalence of GDM has been reported to reach 24.24%, emphasizing the critical clinical importance of early and precise prediction to enable timely and effective intervention [4]. Although the oral glucose tolerance test (OGTT) at 24–28 weeks is the standard diagnostic method, maternal hyperglycemia may already harm the fetus before this stage, and mid-to-late pregnancy interventions often miss the optimal window to prevent irreversible risks [5–7]. Evidence suggests that early assessment between 12–14 weeks, and no later than 20 weeks, is optimal, and that lifestyle interventions in early pregnancy can reduce GDM risk [8].

The increasing maturity of machine learning (ML) has made the identification of GDM before 20 weeks' gestation a feasible new approach [9]. ML models can enable precise screening of high-risk populations and provide in-depth analyses of early-pregnancy risk factors to support timely interventions. Since 2010, related research has grown rapidly, with developed predictive models emphasizing different aspects [10–12]. Yingting Wu et al. employed machine learning algorithms to develop a GDM risk stratification model for Chinese pregnant women, assessing GDM risk before 16 weeks' gestation [13]. Shamil D et al. used multivariable regression to integrate multiple indicators and constructed the PeRSONal (Prediction for Risk-Stratified care for women with GDM) model, achieving favorable results [14]. Yitayeh Belsti et al. compared several ML algorithms — including logistic regression (logit model), K-Nearest Neighbors (KNN), and Gaussian Naïve Bayes (GNB) — to develop GDM risk prediction models and identified the optimal model for GDM prediction [15]. Bo Zhu et al., in a multi-center longitudinal cohort study, validated that steroid hormone (SH) indicators used in early- and mid-trimester Down syndrome screening can be employed to construct GDM prediction models [16]. Jinlang Lyu et al. applied logistic regression to select predictive factors; the final model incorporated nine clinical and biochemical features and demonstrated good predictive performance in the training cohort [17].

The early prediction of GDM faces three major challenges. First, the scarcity of early-pregnancy data is a core bottleneck limiting model performance, as primary healthcare institutions often lack complete prenatal testing protocols before 14 weeks' gestation, thereby restricting the acquisition of sufficient high-quality clinical data [18]. Second, severe class imbalance, manifested as a disproportionately low proportion of positive cases, undermines model recall and markedly increases the risk of missed diagnoses among high-risk pregnant women [19]. Third, the complex interrelationships among clinical indicators, spanning glucose-lipid metabolism, coagulation function, and other physiological systems, limit the effectiveness of traditional feature engineering approaches that largely rely on expert-driven selection, thereby weakening model generalizability. Collectively, these challenges impede the development of robust, clinically relevant, and

generalizable early prediction models for GDM.

To address the aforementioned challenges, this study proposes a unified dual-dimensional enhancement framework, with specific strategies encompassing the following approaches:

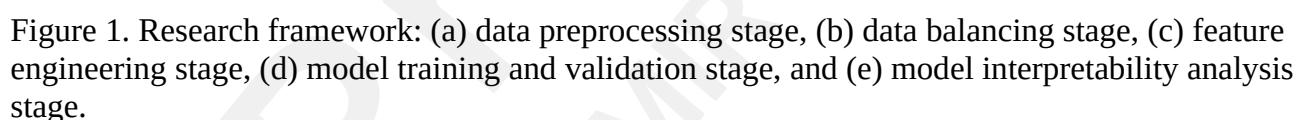
1) We introduce a GAN-based data augmentation method to address class imbalance in GDM prediction. Unlike conventional over-sampling or under-sampling approaches that often sacrifice either accuracy or recall, the proposed method substantially improves recall while maintaining high accuracy.

2) Inspired by Large Language Models (LLMs), we propose a feature-enhancement framework that deeply integrates medical prior knowledge with data distribution patterns, thereby further improving predictive performance.

3) We employ interpretable analyses grounded in information theory and game theory to investigate model outputs, elucidate and evaluate the decision process of the GDM model, and validate the effectiveness of the feature-enhancement framework. This work links model-derived findings with clinical mechanisms underlying GDM.

Our experimental results demonstrate that the model established using the dual-dimensional enhancement framework, particularly the Random Forest model enhanced by TVAE-based feature augmentation (TFRFM), achieved the best performance in predicting GDM. The model attained a recall of 0.7559, an accuracy of 0.8444, and an AUROC of 0.8873, indicating that TFRFM substantially enhances recall without compromising accuracy. This effectively reduces the risk of missed diagnoses among high-risk pregnant women and significantly outperforms conventional sampling methods.

The research workflow is illustrated in Figure 1. The remainder of this paper is organized as follows: Section 2 describes the data sources, data preprocessing methods, and the construction and validation of the machine learning models; Section 3 presents the evaluation of the machine learning models and the results of interpretability analyses; Section 4 discusses the clinical significance of early-pregnancy laboratory indicators for GDM prediction and the impact of improved model recall on GDM management; finally, Section 5 concludes the study.



2.1. Data Source and Preprocessing

[unpublished, non-peer-reviewed preprint]

No.: 20240328A14), and all pregnant women signed informed consent forms.

The study variables include maternal demographic characteristics such as maternal age, pre-pregnancy body mass index, parity, and gestational weight gain; blood test results during pregnancy and the puerperium, including total cholesterol (TC), triglycerides (TG), oral glucose tolerance test results at 24-28 weeks of gestation, blood routine, and coagulation function test results.

The preprocessing was conducted in three steps:

(1) Integration and cleaning, which involved removing irrelevant indicators and invalid OGTT records;

(2) Deletion of abnormal zeros and missing values;

(3) Deduplication of rows and columns, with row-level deduplication based on IDs and column-level removal of redundant features using Pearson correlation coefficients.

2.2. Data augmentation strategy based on GAN sampling

Clinical data on GDM exhibit a marked class imbalance. Let the original datasets be denoted as D_A and D_B , and define the imbalance ratio ρ as the ratio of the number of minority-class samples to that of majority-class samples. In this study, ρ_A and ρ_B were 0.23 and 0.25, respectively, indicating a severely imbalanced dataset. Directly using the original datasets for modeling would lead to insufficient learning of minority-class samples, typically manifested as a low recall rate.

To address this issue, five data balancing strategies were compared [22]: no resampling (Baseline), undersampling, oversampling, Conditional Tabular Generative Adversarial Network (CTGAN), and Tabular Variational Autoencoder (TVAE). An optimized GAN-based minority sample generation approach was proposed to ensure both distributional consistency and clinical plausibility of the generated samples.

2.2.1. Traditional Sampling

In handling imbalanced datasets, traditional sampling methods primarily alleviate the impact of class imbalance on model performance by adjusting the number of samples in the majority and minority classes. The non-sampling method directly uses the original imbalanced dataset as a baseline control to quantify the impact of imbalance. Undersampling randomly reduces samples from the majority class, which is simple and efficient but may lose key information in the majority class, leading to a decline in the model's generalization ability. The Synthetic Minority Over-sampling Technique (SMOTE) algorithm is a classic oversampling method [23]. It generates new samples through linear interpolation within the neighborhood of minority class samples, which is simple and effective. However, its linear assumption may deviate from the true data distribution.

2.2.2. Conditional Table GAN (CTGAN)

CTGAN is a GAN-based method used to model the distribution of tabular data and extract sample rows from this distribution [24]. As shown in Figure 2, CTGAN introduces a mode-specific normalization method to overcome the problems of non-Gaussian and multimodal distributions, and then trains the generator and discriminator in combination with conditional vectors. The generator takes random noise and class labels as input and outputs synthetic samples; the discriminator distinguishes between real and fake samples under the PacGAN framework. The training adopts WGAN-GP loss and incorporates gradient penalty to alleviate mode collapse, ultimately obtaining minority class samples that are highly consistent with the real distribution.

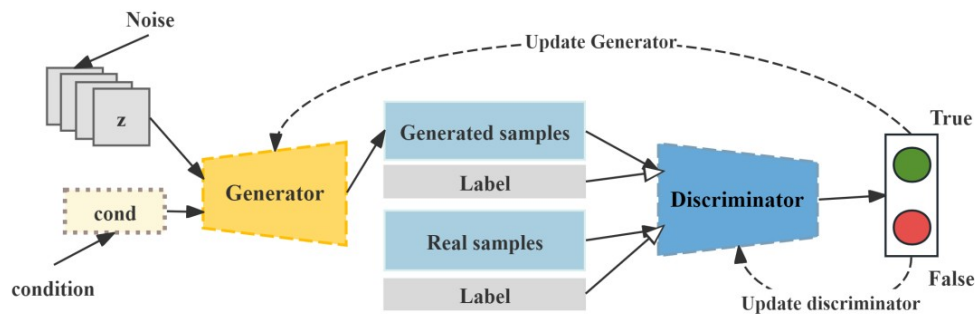


Figure 2. CTGAN network architecture

2.2.3. Table-based Variational Auto-Encoder (TVAE)

Figure 3 illustrates the network structure of TVAE. Based on the Variational Autoencoder (VAE) framework, TVAE captures the complex distribution of tabular data by introducing a probabilistic encoding-decoding structure [25]. The encoder maps raw data to a latent Gaussian distribution, and the decoder reconstructs samples from latent variables while modeling continuous and discrete variables separately. The objective function maximizes the Evidence Lower Bound (ELBO), balancing reconstruction accuracy and distribution approximation to generate minority class samples that are both realistic and diverse.

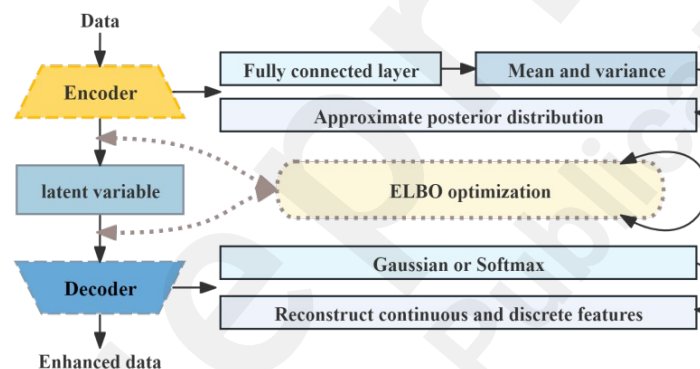


Figure 3. TVAE network structure

2.3. LLM-inspired feature enhancement method

A single laboratory indicator contains limited information, while large language models (LLMs) can automatically extract unstructured pathological knowledge from literature into structured features, breaking through the reliance of traditional feature engineering on expert experience. In this study, information value (IV) is used to measure the predictive contribution of each variable to the outcome, and Pearson correlation coefficient is employed to quantify the linear relationship between continuous variables. Both serve as inputs to the LLM, providing empirical basis for feature selection.

This study employs a systems biology approach, leveraging the knowledge distillation capabilities of large language models (LLMs) and statistical methods to generate descriptions of pathological mechanisms and recommend associated metrics. Biomarkers are initially grouped based on related organ systems, followed by expert review, optimization, and adjustment of the groupings, thereby constructing feature aggregations guided by medical knowledge. Some features are divided into three distinct subgroups:

- (1) Basic Demographic Characteristics Group (BDCG), encompassing maternal age, gravidity, and parity;
- (2) Metabolic Syndrome Group (MSG), including fasting blood glucose, pre-pregnancy body mass index (BMI), triglycerides, and total cholesterol;

(3) Core Liver Biomarkers Group (CLBG), comprising γ -glutamyl transferase, cholinesterase, and total bilirubin.

Each grouping will calculate corresponding statistics, including mean, variance, maximum value, and range. The mean reflects the overall level of the system, variance assesses the coordination of indicators within the group, the maximum value identifies extreme deviations within the group, and the range quantifies the degree of heterogeneity.

Feature combinations create new higher-order features by integrating multiple original or derived features, revealing complex relationships hidden within the data. This study designed four composite features to quantify various metabolic risk factors and their synergistic effects before and during early pregnancy. These features are: TyG index (Triglycerides Glucose Index, TyG), Age $\times$ Pre-pregnancy BMI (Age_BMI), (Fasting Blood Glucose+Triglycerides) $\times$ Pre-pregnancy BMI ((FBG+TG)_BMI), Fasting Blood Glucose $\times$ Pre-pregnancy BMI (BMI_FBG). The TyG index is calculated as follows:

$$TyG = \ln [TG (mg / dl) \times FBG (mg / dl) / 2] . \quad (1)$$

2.4. Predictive model construction

This study will focus on tree-based models for predicting GDM. We selected three machine learning algorithms—decision tree, random forest, and extreme gradient boosting (XGBoost)—to develop GDM prediction models.

The decision tree recursively splits data based on the principle of minimizing mean squared error. After reaching preset termination conditions, it outputs predictions using the mean value of leaf nodes [26]. The random forest constructs multiple trees through Bootstrap resampling and randomly selects feature subsets at each node for splitting; it improves stability by averaging the outputs of all trees and estimates generalization error using out-of-bag samples [27]. XGBoost incorporates regularization into the gradient boosting framework, iteratively fits residuals from previous rounds, uses a greedy algorithm to find optimal split points, and gradually accumulates outputs from each tree, thereby reducing overfitting risks and improving accuracy [28]. These models were implemented in Python using the Scikit-Learn and XGBoost libraries.

The models were fed with specified features, and hyperparameter tuning was performed using the RandomizedSearchCV method [29]. This method randomly samples 20 hyperparameter combinations from a predefined parameter distribution space, with F1_Score as the optimization target. It employs 10-fold cross-validation to ensure the model's generalization performance and selects the parameter combination with the best evaluation results to train the final model. Unlike GridSearchCV, which attempts all possible parameter combinations, this approach randomly samples a subset of combinations from the parameter space. It significantly reduces search time and computational resource consumption without compromising performance.

2.5. Performance Evaluation Metrics

This study aims to predict the occurrence of GDM using machine learning models and evaluate the models with multiple assessment metrics. The model performance is evaluated through the following metrics: Accuracy (ACC), Recall, Precision (positive predictive value, PPV), and F1_Score. We also use the area under the Receiver Operating Characteristic curve (auROC) as one of the evaluation metrics [30]. auROC is used to assess the model's performance across different thresholds. The closer the auROC value is to 1, the better the predictive model is at distinguishing between individuals with and without GDM.

2.6. Interpretable analysis

The purpose of model interpretability analysis methods is to visualize the reasoning or prediction

process of machine learning models through visual images and other means, helping people intuitively understand why a model makes a certain decision. Especially in the task of predicting GDM, interpreting and verifying the model's prediction results is conducive to clinical diagnosis and decision-making.

SHAP (SHapley Additive exPlanations) analysis, as a model interpretation method based on the game theory framework, enhances the interpretability of machine learning models systematically by decomposing the model's output results and assigning feature importance scores for specific predictions to each input feature [31]. A positive score indicates that the feature plays a supporting role in the model's prediction results, while a negative score means the feature has an inhibitory effect on the prediction results. With the SHAP method, people can clearly analyze the model's output mechanism, identify the core features that contribute the most to the prediction results, and quantify the direction and extent of the features' influence, ultimately achieving a transparent interpretation of the model's prediction logic [32-33].

3. Results

3.1. Experimental Setup

Baseline datasets A (5,251 cases, with a positive rate of 18.3%) and datasets B (2,963 cases, with a positive rate of 20.1%) underwent the same preprocessing and were used for the training of CTGAN, TVAE, and GDM modeling. To alleviate class imbalance, the study compared five strategies: non-sampling, undersampling, SMOTE, CTGAN, and TVAE, with a focus on optimizing the GAN workflow. Specifically, GAN networks were used to model the training dataset and generate new data at a certain proportion; positive samples from the generated new data were merged into the original training dataset for the first data balancing; subsequently, undersampling was applied for the second balancing to ensure complete balance in both quantity and distribution of the dataset.

Both datasets retained 36 original features, including demographic, physiological, and laboratory indicators. Notably, the feature enhancement method was divided into two stages: in the first stage, 4 statistical metrics were generated for each of the three groups (BDCG, MSG, and CLBG), resulting in a total of 12 new features; in the second stage, 4 combined features were further constructed to explore non-linear correlations.

3.2. Model Results

To ensure the stability and reliability of the prediction models, the evaluation was conducted in two phases, with each phase comprising three independent assessments: training using only the original features, training with the addition of organ statistical features, and training with the integration of medical knowledge features. In Phase I, dataset A was used as the training set with 10-fold cross-validation to build the prediction models, while dataset B served as the test set for performance evaluation. In Phase II, the roles of the datasets were reversed, with dataset B as the training set and dataset A as the test set.

Table 1 presents the evaluation results of models using different features in the first stage. In terms of data augmentation, except for the CTGAN method that only used the original features, models applying augmentation methods all achieved significant improvements in recall compared with the baseline. Models using the undersampling method showed particularly notable gains in recall. In Evaluation II, the recall of the Random Forest model reached as high as 0.8344. However, this method tended to show a decline in F1_score and accuracy compared with the baseline. This conclusion also applies to all evaluations of XGBoost and Decision Tree models. Models using the TVAE method also achieved remarkable improvements in recall, second only to the undersampling

method. In particular, for all evaluations of the XGBoost model, the increase in recall brought by TVAE exceeded that of the SMOTE method. At the same time, TVAE had a weaker negative impact on F1_score and accuracy than undersampling and could even improve F1_score in some cases.

Table 2 presents the evaluation results of models using different features in the second phase, aimed at verifying the models' generalization capability. It can be observed that the conclusions drawn from Table 1 also apply here. Notably, when comparing the evaluation results of Phase I and Phase II, we found that for the Random Forest model using the TVAE method, the recall rate in Phase I increased progressively from 0.7157 in Assessment I to 0.7508 in Assessment II and then to 0.7559 in Assessment III. The same trend in recall rate improvement was also observed for the model in Phase II after the introduction of new features.

Table 1 Evaluation results of models with different feature enhancement schemes (Phase I)

Models	Methods	Train Samples	Test Samples	Accuracy	Recall	Precision	F1_Score	auROC
Assessment I Using only original features								
XGBoost	Baseline	5251	2963	0.8765	0.5184	0.7990	0.6288	0.8555
	Undersampled	1932	2963	0.7398	0.8278	0.4256	0.5622	0.8459
	SMOTE	8570	2963	0.8994	0.6355	0.8261	0.7183	0.9112
	CTGAN	8566	2963	0.8113	0.4281	0.5412	0.4781	0.7810
	TVAE	7754	2963	0.8761	0.7140	0.6854	0.6994	0.9015
Randomforest	Baseline	5251	2963	0.8616	0.3261	0.9653	0.4875	0.8826
	Undersampled	1932	2963	0.6905	0.7174	0.3645	0.4834	0.7733
	SMOTE	8570	2963	0.8653	0.7308	0.6474	0.6866	0.8931
	CTGAN	8566	2963	0.8556	0.5953	0.6568	0.6246	0.8517
	TVAE	7754	2963	0.8201	0.7157	0.5411	0.6163	0.8559
Decision Tree	Baseline	5251	2963	0.8171	0.3980	0.5667	0.4676	0.7784
	Undersampled	1932	2963	0.7172	0.4465	0.3450	0.3892	0.6666
	SMOTE	8570	2963	0.7688	0.6271	0.4480	0.5226	0.7716
	CTGAN	8566	2963	0.7617	0.3880	0.4056	0.3966	0.6807
	TVAE	7754	2963	0.6895	0.5151	0.3284	0.4010	0.6627
Assessment II Introducing organ statistical features								
XGBoost	Baseline	5251	2963	0.9015	0.5769	0.8984	0.7026	0.8836
	Undersampled	1932	2963	0.7351	0.8211	0.4200	0.5557	0.8371
	SMOTE	8570	2963	0.9035	0.6488	0.8362	0.7307	0.9178
	CTGAN	8566	2963	0.8201	0.4448	0.5696	0.4995	0.7882
	TVAE	7754	2963	0.8147	0.6957	0.5313	0.6025	0.8428
Randomforest	Baseline	5251	2963	0.8630	0.3462	0.9324	0.5049	0.8854
	Undersampled	1932	2963	0.7243	0.8344	0.4100	0.5499	0.8491
	SMOTE	8570	2963	0.8596	0.7191	0.6342	0.6740	0.8951
	CTGAN	8566	2963	0.8569	0.6187	0.6537	0.6357	0.8548
	TVAE	7754	2963	0.8458	0.7508	0.5931	0.6627	0.8833
Decision Tree	Baseline	5251	2963	0.8167	0.4448	0.5577	0.4949	0.7605
	Undersampled	1932	2963	0.5089	0.7458	0.2550	0.3801	0.6391
	SMOTE	8570	2963	0.7685	0.6722	0.4507	0.5396	0.7722
	CTGAN	8566	2963	0.7665	0.5234	0.4347	0.4750	0.7601
	TVAE	7754	2963	0.4857	0.7876	0.2521	0.3820	0.6276
Assessment III Integrating medical knowledge features								
XGBoost	Baseline	5251	2963	0.8788	0.5552	0.7812	0.6491	0.8762
	Undersampled	1932	2963	0.7324	0.8311	0.4180	0.5562	0.8462
	SMOTE	8570	2963	0.9038	0.6522	0.8351	0.7324	0.9136
	CTGAN	8566	2963	0.8873	0.7124	0.7245	0.7184	0.9018
	TVAE	7754	2963	0.8758	0.7274	0.6797	0.7027	0.9084
Randomforest	Baseline	5251	2963	0.8687	0.3645	0.9604	0.5285	0.8792
	Undersampled	1932	2963	0.6841	0.7391	0.3617	0.4857	0.7832
	SMOTE	8570	2963	0.8498	0.7358	0.6052	0.6642	0.8923
	CTGAN	8566	2963	0.8404	0.6237	0.6006	0.6120	0.8411
	TVAE	7754	2963	0.8444	0.7559	0.5893	0.6623	0.8873

Decision Tree	Baseline	5251	2963	0.8137	0.4783	0.5437	0.5089	0.7661
	Undersampled	1932	2963	0.6473	0.6003	0.3082	0.4073	0.678
	SMOTE	8570	2963	0.7617	0.6505	0.4391	0.5243	0.7738
	CTGAN	8566	2963	0.7648	0.5518	0.4348	0.4864	0.7485
	TVAE	7754	2963	0.5886	0.6271	0.2735	0.3809	0.6449

Table 2 Evaluation results of models with different feature enhancement schemes (Phase II)

Models	Methods	Train Samples	Test Samples	Accuracy	Recall	Precision	F1_Score	auROC
Assessment I Using only original features								
XGBoost	Baseline	2963	5251	0.8608	0.4576	0.6810	0.5474	0.7893
	Undersampled	1196	5251	0.7149	0.7246	0.3625	0.4833	0.7850
	SMOTE	4730	5251	0.8463	0.5321	0.5915	0.5602	0.8209
	CTGAN	8666	5251	0.8469	0.5342	0.5931	0.5621	0.8135
	TVAE	6328	5251	0.8324	0.5994	0.5401	0.5682	0.8384
Randomforest	Baseline	2963	5251	0.8581	0.2567	0.9018	0.3997	0.8203
	Undersampled	1196	5251	0.7020	0.7091	0.3479	0.4668	0.7744
	SMOTE	4730	5251	0.8096	0.6077	0.4859	0.5400	0.8156
	CTGAN	8666	5251	0.8094	0.5507	0.4841	0.5153	0.7746
	TVAE	6328	5251	0.7793	0.6812	0.4361	0.5317	0.8193
Decision Tree	Baseline	2963	5251	0.7976	0.3602	0.4388	0.3957	0.7097
	Undersampled	1196	5251	0.6109	0.6035	0.2599	0.3634	0.6590
	SMOTE	4730	5251	0.6203	0.6739	0.2794	0.3950	0.6828
	CTGAN	8666	5251	0.7001	0.5166	0.3105	0.3879	0.6784
	TVAE	6328	5251	0.6376	0.6108	0.2787	0.3827	0.6745
Assessment II Introducing organ statistical features								
XGBoost	Baseline	2963	5251	0.8541	0.2277	0.9167	0.3648	0.8185
	Undersampled	1196	5251	0.7119	0.7246	0.3595	0.4806	0.7907
	SMOTE	4730	5251	0.8459	0.5362	0.5893	0.5615	0.8210
	CTGAN	8666	5251	0.8259	0.5145	0.5276	0.5210	0.7869
	TVAE	6328	5251	0.7616	0.6656	0.4090	0.5067	0.8023
Randomforest	Baseline	2963	5251	0.8530	0.2277	0.8943	0.3630	0.8270
	Undersampled	1196	5251	0.7052	0.7308	0.3541	0.4770	0.7903
	SMOTE	4730	5251	0.8065	0.6242	0.4801	0.5428	0.8101
	CTGAN	8666	5251	0.8077	0.5466	0.4800	0.5111	0.7836
	TVAE	6328	5251	0.7701	0.7039	0.4247	0.5298	0.8195
Decision Tree	Baseline	2963	5251	0.7924	0.3406	0.4207	0.3764	0.6880
	Undersampled	1196	5251	0.6757	0.4565	0.2724	0.3412	0.6390
	SMOTE	4730	5251	0.6165	0.6646	0.2753	0.3893	0.6627

	CTGAN	8666	5251	0.7220	0.5228	0.3358	0.4089	0.6937
	TVAE	6328	5251	0.6905	0.6460	0.3272	0.4344	0.7181
Assessment III Integrating medical knowledge features								
XGBoost	Baseline	2963	5251	0.8659	0.4172	0.7408	0.5338	0.7926
	Undersample d	1196	5251	0.7041	0.7236	0.3520	0.4736	0.7908
	SMOTE	4730	5251	0.8448	0.5507	0.5827	0.5663	0.8246
	CTGAN	8666	5251	0.8448	0.5466	0.5834	0.5644	0.8162
	TVAE	6328	5251	0.8284	0.6315	0.5281	0.5752	0.8371
Randomforest	Baseline	2963	5251	0.8619	0.2992	0.8576	0.4436	0.8199
	Undersample d	1196	5251	0.6999	0.6988	0.3444	0.4614	0.7724
	SMOTE	4730	5251	0.7987	0.6325	0.4653	0.5362	0.8114
	CTGAN	8666	5251	0.8056	0.5642	0.4760	0.5163	0.7798
	TVAE	6328	5251	0.7734	0.7143	0.4302	0.5370	0.8234
Decision Tree	Baseline	2963	5251	0.7993	0.3675	0.4449	0.4025	0.7101
	Undersample d	1196	5251	0.5309	0.7464	0.2453	0.3693	0.6314
	SMOTE	4730	5251	0.7250	0.5559	0.3460	0.4265	0.7058
	CTGAN	8666	5251	0.7313	0.5580	0.3539	0.4331	0.6984
	TVAE	6328	5251	0.6254	0.6532	0.2788	0.3908	0.6869

To validate the effectiveness of the proposed framework combining GAN-based data augmentation with LLM-inspired feature enhancement, this study systematically evaluated the impact of different strategies on model performance using statistical methods. First, a non-parametric Kruskal-Wallis test was conducted to determine whether data augmentation methods and feature enhancement methods produced statistically significant differences in model performance metrics such as accuracy and recall [34]. When the Kruskal-Wallis test indicated significant differences, pairwise comparisons were further performed using the Games-Howell post hoc test to identify which specific augmentation strategies led to superior performance. Meanwhile, Cohen's d effect size was employed to quantify the practical significance of the differences: an effect size of 0.2-0.5 indicated a small difference, 0.5-0.8 a medium difference, and greater than 0.8 a large difference [35], thereby distinguishing whether statistically significant differences also had practical value.

Table 3 presents the results of the Kruskal-Wallis test. Data augmentation methods showed significant effects on Accuracy ($p = 0.0004$), Recall ($p = 2.496 \times 10^{-5}$), Precision ($p = 0.0002$), and F1_Score ($p = 0.0428$), with no significant effect observed on AUROC ($p = 0.0880$). In contrast, feature enhancement methods did not have a significant impact on any of the evaluated metrics ($p > 0.05$).

Table 3 Kruskal-Wallis Test Results

Metric	Factor	Statistic	P-value
Accuracy	feature enhancement	0.055	0.9730
	data augmentation	20.257	0.0004***
Recall	feature enhancement	2.429	0.2968
	data augmentation	26.510	0.0000****
Precision	feature enhancement	0.033	0.9835
	data augmentation	22.285	0.0002***
F1_Score	feature enhancement	0.842	0.6565
	data augmentation	9.864	0.0428*
auROC	feature enhancement	0.281	0.8691
	data augmentation	8.099	0.0880

In the table, * denotes $P \leq 0.05$, ** denotes $P \leq 0.01$, *** denotes $P \leq 0.001$, and **** denotes $P \leq 0.0001$. The same notation applies elsewhere in the manuscript.

For the metrics that showed significance in the Kruskal-Wallis test, group differences were further analyzed using the Games-Howell post hoc test. Figure 4 presents the results as a forest plot. Recall is a core metric for GDM prediction, reflecting the model's ability to identify positive cases and avoid missed diagnoses. The recall of TVAE was significantly higher than that of the baseline (0.4454, $p < 0.001$) and CTGAN (0.5429, $p = 0.0241$), with large effect sizes (Cohen's d = 2.894 and 1.637, respectively), indicating that TVAE-generated samples effectively enhance the model's recognition of positive cases. For Accuracy, SMOTE (0.8423) and TVAE (0.7601) did not differ significantly ($p = 0.5186$), but both were significantly higher than undersampling (0.6866, $p < 0.01$), indicating that generated samples did not compromise overall classification ability.

In summary, we consider that the Random Forest model enhanced by TVAE-based feature augmentation (TFRFM) represents an effective approach for early-pregnancy prediction of GDM, providing strong support for clinical decision-making. This model demonstrated stable and superior performance across different data augmentation and feature enhancement strategies, showing high accuracy and auROC. Notably, it effectively improved the identification of positive cases while maintaining a balance between recall and precision, which is critical for the early prediction and precise management of GDM.

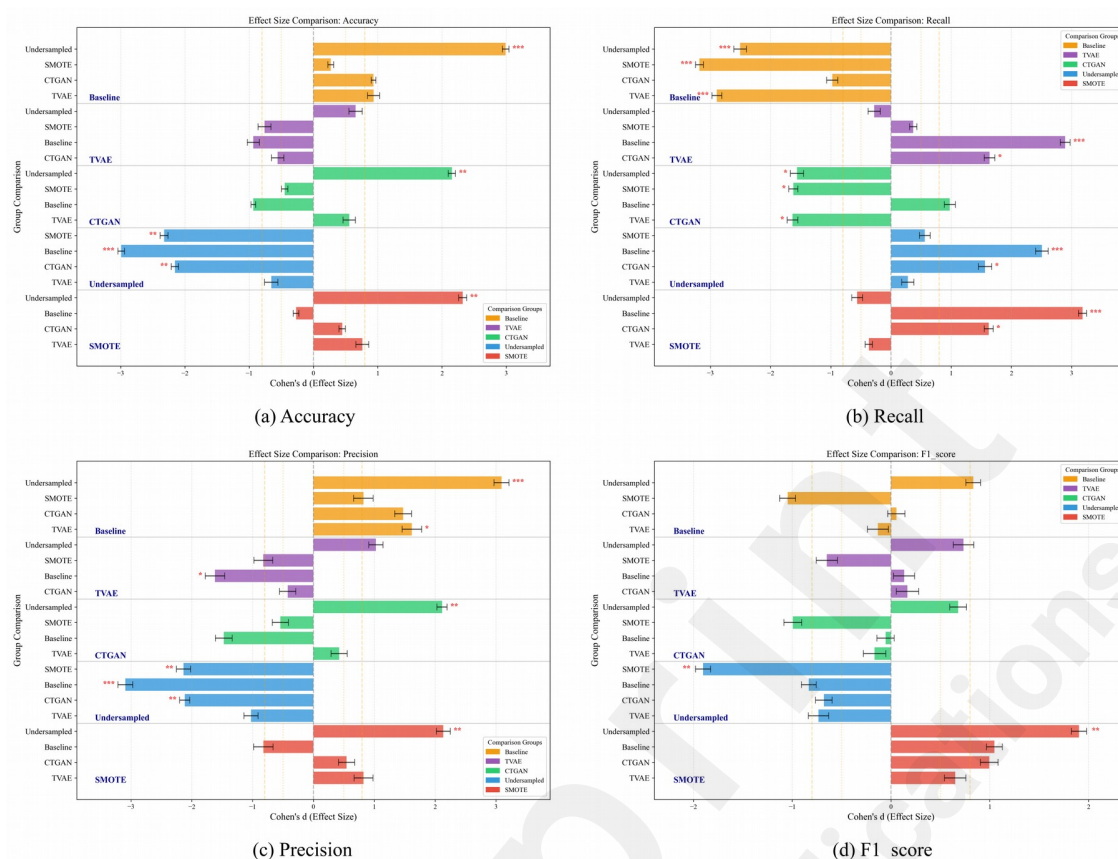


Figure 4 Forest Plot

3.3 Interpretability Analysis Results

3.2.1. Feature Importance and Correlation Analysis

To investigate the role of features in the model, we conducted feature importance and correlation analyses using the TFRFM. By calculating each feature's contribution to the model's decision-making process, we identified the top 20 features that play a critical role in predicting early-pregnancy GDM. Additionally, correlation analysis was performed on these 20 features to reveal potential relationships and interactions among them.

Figure 5 presents the feature importance analysis of the TFRFM. Fasting Blood Glucose (FBG) achieved the highest importance score, followed by Leukocyte Count (WBC) and Activated Partial Thromboplastin Time (APTT), indicating that these features contribute substantially to the model's decision-making process and can be regarded as the most critical predictors of GDM. In addition, the four composite features investigated in this study, along with the mean of MSG, were also ranked among the top 20 features, demonstrating their important role in prediction and further validating the effectiveness of integrating medically informed features. Notably, although some traditional indicators such as Age, TG, and BMI may not appear prominent when analyzed independently, their contributions become substantial when combined with other features or when potential inter-feature correlations are considered. This finding highlights that interactions and relationships among features are also crucial for accurately predicting early-pregnancy GDM.

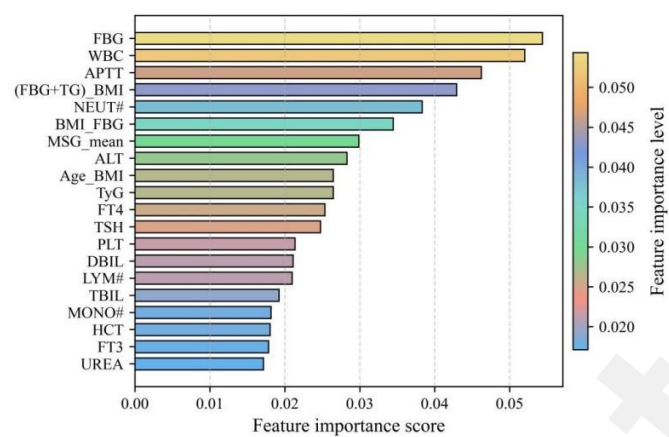


Figure 5 Feature Importance Analysis of the TFRFM

Figure 6 also presents the correlation analysis among these 20 features. The composite feature (FBG+TG)_BMI, ranked fourth in feature importance, exhibited the highest Pearson correlation with GDM, followed by BMI_FBG, Age_BMI, TyG, MSG_mean, FBG, APTT, and WBC, with APTT showing a negative correlation with GDM. These correlation results suggest that the relationships among combined features and their association with GDM may be more complex and informative than analyses of individual features alone. This underscores the importance of considering inter-feature interactions and associations when predicting early-pregnancy GDM rather than relying solely on the significance of individual features.

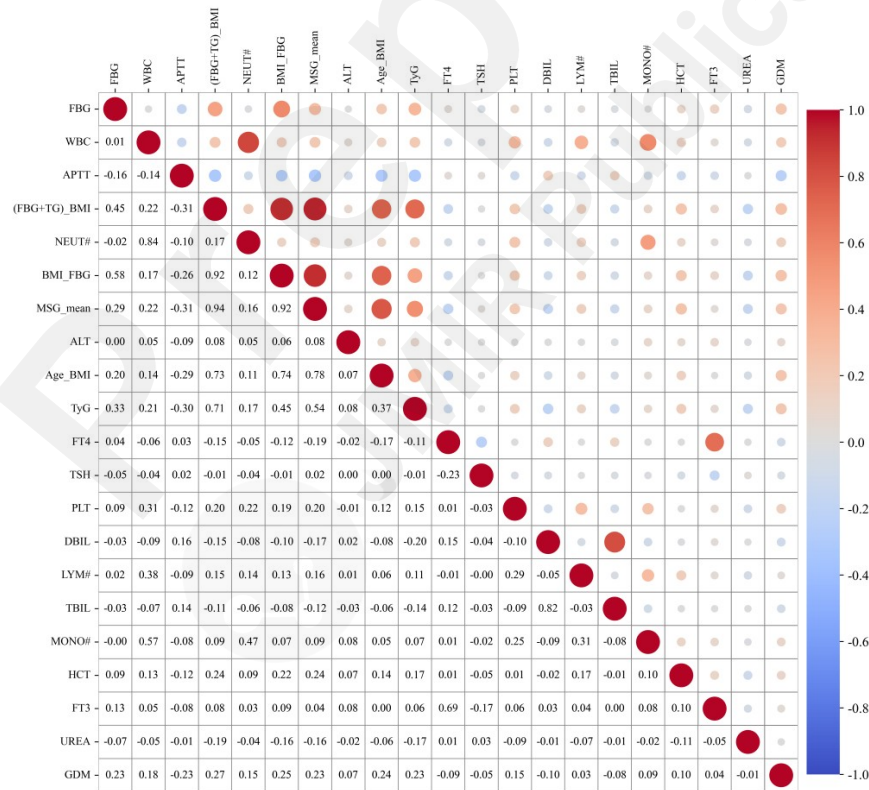


Figure 6 Correlation Analysis of the Top 20 Important Features in the TFRFM

3.2.2. SHAP Analysis Results

SHAP analysis quantifies the contribution of each feature to the model’s predictions, providing an intuitive view of feature importance rankings and influence patterns under different methods. In this section,

we focus on comparing the SHAP analysis results of the TFRFM, the Random Forest model enhanced by the TVAE method (TVRFM), and the Random Forest model trained on baseline data (BLRFM), highlighting the advantages of integrating the TVAE approach with feature enhancement in capturing key predictive factors, stabilizing feature effects, and characterizing feature relationships.

As shown in Figure 7, the top five features in terms of average SHAP values for the TFRFM are FBG, APTT, WBC, (FBG+TG)_BMI, and Neutrophil (NEUT#), indicating that these are the key drivers for predicting blood glucose levels in the model. In terms of the magnitude of feature impact, FBG has a SHAP range of $[-0.10, 0.125]$ with a span of 0.225, and WBC has a SHAP range of $[-0.125, 0.10]$ with the same span of 0.225, showing the greatest fluctuations in prediction influence. This aligns with clinical understanding, where blood glucose (e.g., FBG) and inflammatory indicators (e.g., WBC) are directly associated with pregnancy risk. Figure 7 clearly shows that the average SHAP values of the top 20 features are all greater than 0, indicating that these features contribute positively to the prediction, and the feature importance ranking aligns well with medical knowledge.

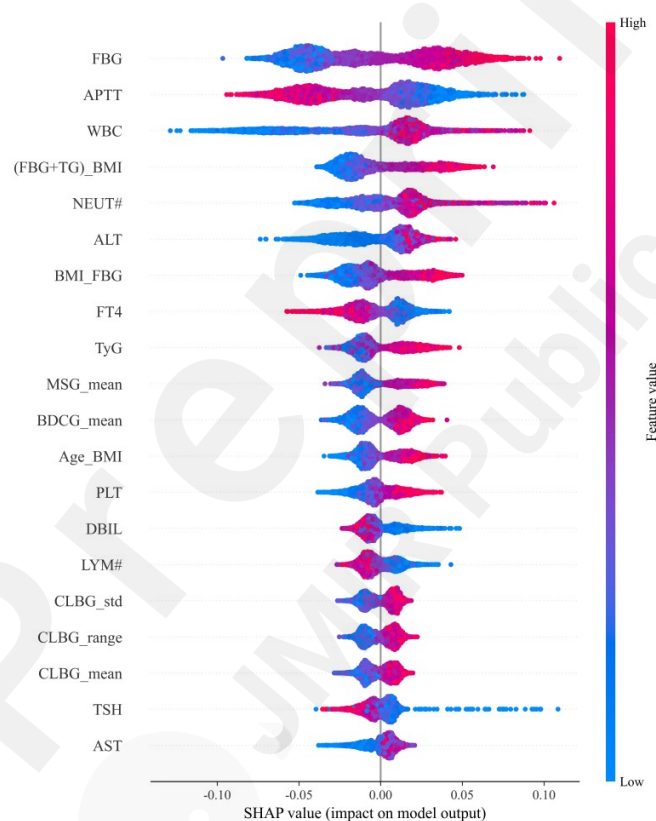


Figure 7 SHAP Beeswarm Plot of the TFRFM

Figure 8 shows the SHAP beeswarm plots for TVRFM and BLRFM. The top five features by average SHAP value in TVRFM are FBG, APTT, WBC, ALT, and NEUT#, with only 45% overlap (9 out of 20) with the top 20 features in TFRFM. In TVRFM, the SHAP value range for FBG is 0.3 $[-0.125, 0.175]$ and for WBC is 0.275 $[-0.15, 0.125]$, both larger than in TFRFM, indicating greater variability of key feature influence and lower model stability without feature enhancement. In BLRFM, the top five features by average SHAP value are FBG, TG, BMI, Age, and APTT, with only 35% overlap (7 out of 20) with the top 20 features in TFRFM. Notably, inflammation-related indicators such as WBC and NEUT# are absent, while Age, which has weaker direct clinical relevance, is included. In the baseline model, the SHAP value range for FBG reaches 0.325 $[-0.05, 0.275]$, exhibiting the largest fluctuation. The scatter points are elongated on the right side of the zero line and concentrated on the left, indicating substantial prediction bias for high-risk

pregnant women.

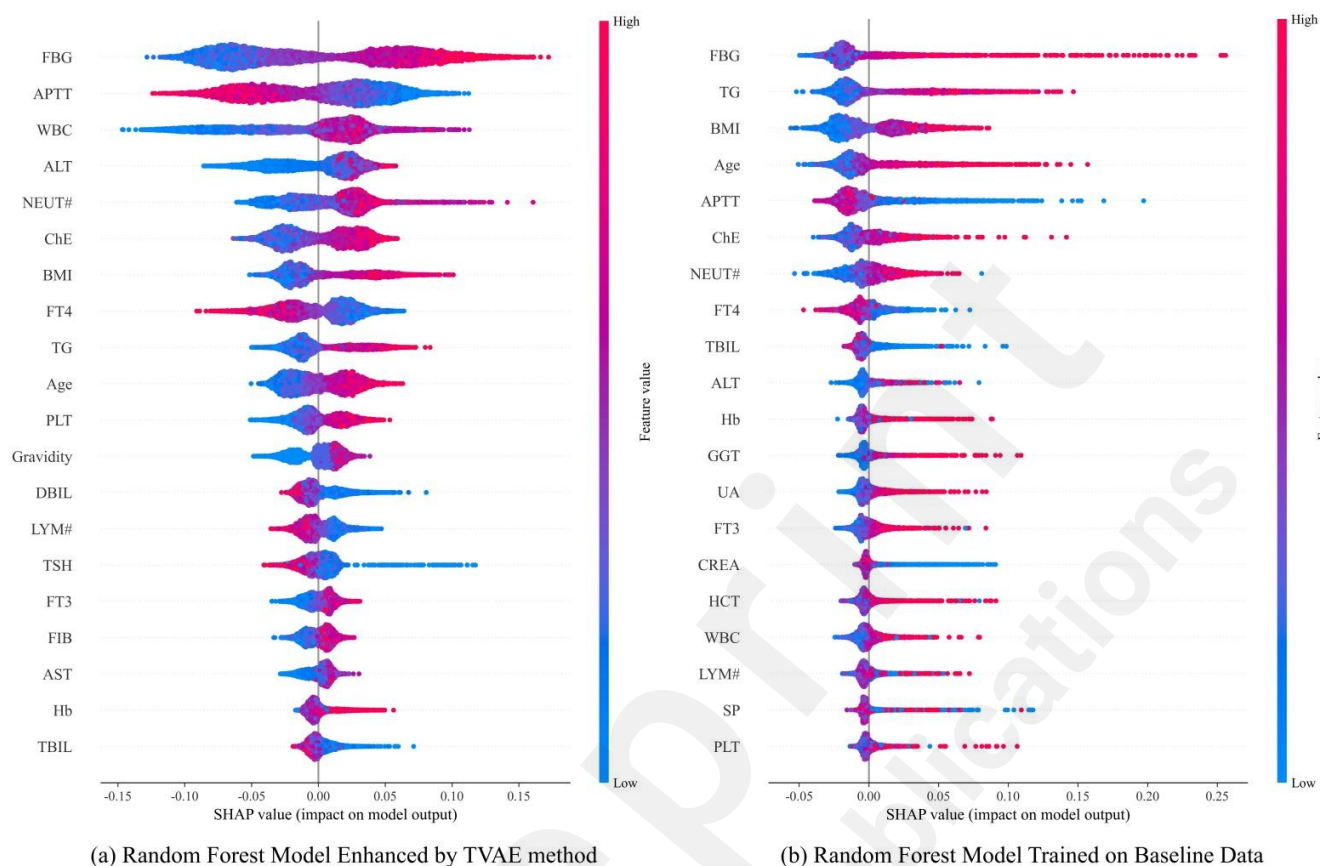


Figure 8 SHAP beeswarm plots of TVRFM (a) and BLRFM (b)

The above findings indicate that the introduction of enhanced features such as (FBG+TG)_BMI strengthens the impact of interactions among physiological indicators, aligning more closely with clinical logic. The SHAP value fluctuations of key features, including FBG and WBC, are reduced, improving the robustness of model predictions. The model also decreases the presence of irregular feature–target relationships, captures linear associations more clearly, and retains the ability to indicate potential nonlinear relationships. In summary, TFRFM, through the combination of feature enhancement and generative modeling, significantly improves the identification of key predictive factors and enhances model stability, providing a more reliable reference for assessing pregnancy-related risks.

4. DISCUSSION

Early screening and risk-stratified management of GDM are crucial for reducing the risk of pregnancy-related complications. Effective prediction of GDM in early pregnancy allows for the identification of high-risk pregnant women, enabling clinicians to promptly implement evidence-based interventions. This approach also facilitates the development of precise and individualized management plans tailored to high-risk groups, thereby improving maternal and neonatal health outcomes. Given the profound impact of GDM-related complications on maternal and fetal health, early screening combined with risk-stratified management has become a core component of optimizing GDM interventions [36]. Consequently, exploring earlier and more accurate methods for predicting GDM to enable timely intervention has emerged as an urgent and critical challenge.

The pathogenesis of GDM involves pancreatic β -cell dysfunction and tissue insulin resistance. The combined indicator of FBG, TG, and BMI, namely the (FBG+TG)_BMI proposed in this study, exhibits a close biological association with insulin resistance. Its underlying mechanism primarily reflects the interplay between dysregulated glucose-lipid metabolism and obesity-related pathophysiological processes. FBG serves as a direct marker of insulin sensitivity, and its elevation indicates increased hepatic glucose output due to hepatic insulin resistance as well as impaired peripheral glucose utilization. Available studies suggest a biologically plausible link between hyperglycaemia and dyslipidaemia [37]. Hyperglycaemia can trigger reactive oxygen species production in pancreatic β -cells, leading to oxidative stress and cellular dysfunction, thereby exacerbating insulin resistance. On this pathological basis, elevated TG levels directly impair glucose metabolism in muscle tissue and inhibit insulin signaling pathways, further aggravating insulin resistance. Additionally, free fatty acids generated from TG hydrolysis can exert lipotoxic effects, further damaging β -cell function. Consequently, FBG and TG jointly drive the progression of metabolic imbalance, forming a vicious cycle of overlapping glucotoxicity and lipotoxicity. BMI, as a core indicator of obesity, amplifies insulin resistance through ectopic fat deposition, pro-inflammatory factor release, and endoplasmic reticulum stress. Therefore, the combined effect of FBG, TG, and BMI more comprehensively reflects the metabolic burden by superimposing glucotoxicity, lipotoxicity, and obesity-related states, and its synergistic effect can more accurately capture the complex pathophysiology of insulin resistance than any single indicator. Accordingly, the composite index (FBG+TG)_BMI may serve as a simple surrogate tool for assessing insulin resistance, particularly suitable for early-pregnancy screening to rapidly identify high-risk individuals, thereby providing a biological basis for risk-stratified management and early intervention.

Clinical studies have shown that pregnant women with GDM often exhibit shortened APTT from early to mid-pregnancy, which is consistent with a mild hypercoagulable state. For example, in GDM pregnancies, prothrombin time (PT) and APTT in early pregnancy are significantly reduced, and multivariate analyses indicate that a shorter APTT is an independent predictor of GDM [38]. Studies examining laboratory data from women at 24–28 weeks of gestation also report that the GDM group shows markedly shorter PT and APTT and higher fibrinogen levels compared with controls [39]. Research on type 2 diabetes has similarly found that overt diabetes is associated with shortened APTT and elevated fibrinogen, suggesting a general thrombotic tendency in diabetes [40].

GDM is characterized by key pathophysiological features, including insulin resistance, pancreatic β -cell dysfunction, and low-grade inflammation, all of which contribute to a procoagulant state. Under diabetic conditions, hyperglycemia and insulin resistance increase the expression of coagulation factors in the intrinsic pathway, such as FXII, FXI, and FVIII, resulting in APTT shortening [41]. In GDM, FXII levels are significantly elevated, and its critical role in activating the intrinsic coagulation pathway helps explain the observed APTT reduction [42].

WBC and NEUT#, as markers of inflammation, are involved in the development of GDM by mediating chronic low-grade inflammatory responses. Pregnant women with GDM often exhibit insulin resistance and dysregulated glucose and lipid metabolism, and elevated blood glucose and free fatty acids can activate inflammatory signaling pathways, stimulating immune cells such as neutrophils to release proinflammatory factors, thereby inducing a systemic low-grade inflammatory state. Neutrophils, being the most abundant leukocyte subset in peripheral blood, can further exacerbate inhibition of insulin signaling and β -cell dysfunction when activated or functionally impaired. Mild elevations of WBC and NEUT# are observed in early pregnancy among GDM patients, and their levels correlate with the degree of insulin resistance, poor glycemic control, and adverse pregnancy outcomes. These findings suggest that inflammatory markers may not only be concomitant phenomena in GDM pathophysiology but also actively contribute to disease development by amplifying the inflammatory cascade, providing potential biological evidence for early identification of high-risk populations and the development of anti-inflammatory targeted interventions.

The risk probability of developing GDM in pregnant women can be quantitatively estimated using

machine learning models [43]. Given the notable advantage of machine learning in handling complex, high-dimensional data, several machine learning-based predictive models have demonstrated superior performance compared with traditional regression models [44]. For instance, Yuzhen Ye and colleagues constructed GDM prediction models using more than 30 machine learning algorithms, including decision trees and random forests, and compared them with logistic regression models. Their results confirmed the stronger predictive capability of machine learning approaches and identified triglycerides, fasting blood glucose, BMI, and HbA1c as among the most important predictors [45]. Similarly, Yitayeh Belsti et al. developed new models based on the internationally validated Monash GDM model using Extreme Gradient Boosting and CatBoost classifiers, achieving optimal discrimination performance with AUCs of 92% and 93%, respectively, while decision curve analysis confirmed the highest clinical utility of these models [15]. Furthermore, our findings indicate that the TFRFM model, enhanced through data augmentation and feature augmentation, maximizes early prediction of GDM, demonstrating superior robustness and generalization ability, with an auROC of 0.8873 and a recall of 0.7559.

Recall is an indicator that measures a model's ability to correctly identify positive cases. An improvement in recall means that the model reduces the risk of missed diagnoses, allowing more pregnant women who truly have GDM to be identified in a timely manner, although this may be accompanied by an increase in false positives. Reducing missed diagnoses directly benefits maternal and neonatal health outcomes, as early prevention or diagnosis of GDM enables pregnant women to receive targeted interventions such as personalized nutritional guidance or blood glucose monitoring, thereby lowering the risks of macrosomia and preterm birth. At the same time, although the false-positive rate may increase, non-pharmacological interventions such as dietary management and exercise guidance following a misdiagnosis still provide positive health effects. Their potential adverse impacts can be effectively mitigated through standardized OGTT reassessment in mid-pregnancy. Non-pharmacological interventions, being the foundation of GDM management, can help pregnant women develop healthy lifestyles and reduce metabolic burden even in cases of misdiagnosis [46]. Furthermore, OGTT reassessment in early pregnancy serves as the final diagnostic reference, ensuring that high-risk individuals receive appropriate scientific interventions while minimizing unnecessary medical treatment and reducing the potential impact of misdiagnosis on maternal and neonatal health. Therefore, an increase in recall promotes the transition of GDM stratified management from a reactive to a proactive approach and allows primary healthcare resources to be optimally allocated to high-risk populations.

It must be acknowledged that the screening and management of GDM largely depend on the available medical resources and the clinical judgment of healthcare providers. However, primary hospitals often exhibit considerable variability in these aspects, which limits their capacity to prevent and manage GDM and reduces their ability to provide timely and reliable personalized interventions for pregnant women. Furthermore, given that GDM may share underlying pathophysiological mechanisms with other pregnancy-related disorders, exploring the relationships among these conditions and applying the proposed framework combining GAN-based data augmentation with LLM-inspired feature enhancement to develop integrated prediction and management tools for multiple gestational diseases holds promise for further improving maternal and neonatal health outcomes.

5. Conclusion

This study proposes a framework that combines GAN-based data augmentation with LLM-inspired feature enhancement, effectively addressing data imbalance in GDM datasets and significantly improving model recall. In our experiments, we observed that a random forest model incorporating TVAE and feature enhancement not only increased recall but also maintained high precision, accuracy, and auROC. The model

demonstrated superior overall performance and stability across multiple independent evaluations, highlighting its potential as an effective tool for early-pregnancy GDM prediction. Furthermore, interpretability analysis revealed that FBG, (FBG+TG)_BMI, APTT, WBC, and NEUT# contributed most significantly to GDM prediction. These findings provide important biological evidence to support rapid screening, stratified management, and early intervention for GDM.

The proposed method not only enhances predictive accuracy but also offers new insights into the pathophysiological mechanisms of GDM. By emphasizing interactions among physiological indicators, the model better captures complex relationships within clinical data, which has important implications for guiding clinical practice. In the future, we aim to further optimize this framework and explore its applicability to predicting other pregnancy-related diseases, advancing the prediction and management of gestational disorders and contributing to improved maternal and neonatal health outcomes.

Funding

This work was supported by [the Guangdong Basic and Applied Research Foundation] grant number [2023A1515140146; 2022A1515110138; 2024A1515140193]; [Huizhou University College Students' Innovation and Entrepreneurship Training Program Funding] grant number [S202410577052].

Credit authorship contribution statement

Xiekun Chen and **Zhifa Jiang** had equal contributions and were jointly responsible for the original draft, Visualization, Validation, Software, Resources, Methodology, and Conceptualization. **Dong Su**: Resources, Methodology, Investigation, Data curation. **Xiaoping Chen**: Writing – review & editing, Writing – original draft, Visualization. **Aiping Chen**: Writing – review & editing, Writing – original draft, Formal analysis. **Huabin Wang** and **Zhen Zhang**: Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We sincerely appreciate the collaboration and assistance of the physicians from the Department of Obstetrics and Gynecology and the Information Department at Huizhou First Maternal and Child Health Care Hospital, Guangdong, China, in collecting maternal medical records for this study.

Data availability statement

Data will be made available on request.

References

- [1] Sweeting A, Wong J, Murphy HR, et al. (2022). A clinical update on gestational diabetes mellitus. *Endocr. Rev.*, 43(5), 763–793. doi:10.1210/edrev/bnac003
- [2] Moon JH, Jang HC (2022). Gestational diabetes mellitus: diagnostic approaches and maternal-offspring complications. *Diabetes Metab. J.*, 46(1), 3–14. doi:10.4093/dmj.2021.0335
- [3] Kim EK, Kwak SH, Jung HS, et al. (2019). The effect of a smartphone-based, patient-centered diabetes care system in patients with type 2 diabetes: A randomized, controlled trial for 24 weeks. *Diabetes Care*, 42(1), 3–9. doi:10.2337/dc17-2197
- [4] Ellahham S (2020). Artificial intelligence: The future for diabetes care. *Am. J. Med.*, 133(8), 895–900.

doi:10.1016/j.amjmed.2020.03.033

[5] Schmidt MI, Duncan BB, Reichelt AJ, et al. (2001). Gestational diabetes mellitus diagnosed with a 2-h 75-g oral glucose tolerance test and adverse pregnancy outcomes. *Diabetes Care*, 24(7), 1151–1155. doi:10.2337/diacare.24.7.1151

[6] Thériault S, Forest J-C, Massé J, et al. (2014). Validation of early risk-prediction models for gestational diabetes based on clinical characteristics. *Diabetes Res. Clin. Pract.*, 103(3), 419–425. doi:10.1016/j.diabres.2013.12.009

[7] Yogev CH, et al. (2010). Hyperglycemia and adverse pregnancy outcome (HAPO) study: Preeclampsia. *Am. J. Obstet. Gynecol.*, 202(3), 255.e1–255.e7. doi:10.1016/j.ajog.2010.01.018

[8] Sovio U, Murphy HR, Smith GC (2016). Accelerated fetal growth prior to diagnosis of gestational diabetes mellitus: A prospective cohort study of nulliparous women. *Diabetes Care*, 39(6), 982–987. doi:10.2337/dc16-0160

[9] Jiang Z, Chen X, Lai Y, et al. (2025). Early prediction of postpartum dyslipidemia in gestational diabetes using machine learning models. *Sci. Rep.*, 15, 8028. doi:10.1038/s41598-025-92299-9

[10] Artzi NS, Shilo S, Hadar E, et al. (2020). Prediction of gestational diabetes based on nationwide electronic health records. *Nat. Med.*, 26(1), 71–76. doi:10.1038/s41591-019-0724-8

[11] Kennelly MA, McAuliffe FM (2016). Prediction and prevention of gestational diabetes: An update of recent literature. *Eur. J. Obstet. Gynecol. Reprod. Biol.*, 202, 92–98. doi:10.1016/j.ejogrb.2016.03.032

[12] Snyder BM, Baer RJ, Oltman SP, et al. (2020). Early pregnancy prediction of gestational diabetes mellitus risk using prenatal screening biomarkers in nulliparous women. *Diabetes Res. Clin. Pract.*, 163, 108139. doi:10.1016/j.diabres.2020.108139

[13] Wu Y, Ma S, Wang Y, et al. (2021). A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women. *Diabetes Res. Clin. Pract.*, 179, 108963. doi:10.1016/j.diabres.2021.109001

[14] Cooray SD, Boyle JA, Soldatos G, et al. (2022). Development, validation and clinical utility of a risk prediction model for adverse pregnancy outcomes in women with gestational diabetes: The PeRSONal GDM model. *eClinicalMedicine*, 52, 101334. doi:10.1016/j.eclinm.2022.101637

[15] Belsti Y, Moran L, Du L, et al. (2023). Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. *Int. J. Med. Inform.*, 179, 105292. doi:10.1016/j.ijmedinf.2023.105228

[16] Zhu B, Yin B, Li H, et al. (2024). A prediction model for gestational diabetes mellitus based on steroid hormonal changes in early and mid-down syndrome screening: A multicenter longitudinal study. *Diabetes Res. Clin. Pract.*, 217, 111865. doi:10.1016/j.diabres.2024.111865

[17] Lyu J, Peng Y, Yang L, et al. (2025). Development and validation of a prediction model for gestational diabetes mellitus based on clinical characteristics and laboratory biomarkers among Chinese women. *Nutr. Metab. Cardiovasc. Dis.*, 35, 104065. doi:10.1186/s12884-025-07442-1

[18] Wang S, Hang R, Wei Y, et al. (2024). Dilemma in primary care facility development: A qualitative study based on grounded theory in Guilin city. *Chin. Gen. Pract. J.*, 1(1), 11–17. doi:10.1016/j.cgpj.2024.03.008

[19] Ghorbani M, Kazi A, Soleymani Baghshah M, et al. (2022). RA-GCN: Graph convolutional network for disease prediction problems with imbalanced data. *Med. Image Anal.*, 75, 102272. doi:10.1016/j.media.2021.102272

[20] Lapolla A, Dalfrà MG, Ragazzi E, et al. (2011). New International Association of the Diabetes and Pregnancy Study Groups (IADPSG) recommendations for diagnosing gestational diabetes compared with former criteria: A retrospective study on pregnancy outcome. *Diabet. Med.*, 28(9), 1074–1077. doi:10.1111/j.1464-5491.2011.03351

[21] Abd Aleankushiu YF, Radhi MM (2025). Digital Addiction and Its Association with Nutrition-Related

Health Behaviors Among Female Secondary School Students. *Natl. J. Community Med.*, 16(3), -. doi:10.55489/njcm.160320255049

[22] Rdn LTB, Shrestha A, Shrestha A, et al. (2024). Development and Testing of a Mobile App for Management of Gestational Diabetes in Nepal: Protocol for a User-Centered Design Study and Exploratory Randomized Controlled Trial. *JMIR Res. Protoc.*, 13(1), e43210. doi:10.2196/59423

[23] Li, X., Wang, P., Zhu, Y. et al. (2024). Interpretable machine learning model for predicting acute kidney injury in critically ill patients. *BMC Med Inform Decis Mak* 24, 148. doi:10.1186/s12911-024-02537-9

[24] Xu L, Skoularidou M, Cuesta-Infante A, et al. (2019). Modeling tabular data using conditional GAN. In: *Adv. Neural Inf. Process. Syst.* (Vol. 32, Article 659, pp. 7335–7345). Curran Associates, Inc. doi:10.48550/arXiv.1907.00503. (Conference held in Vancouver, Canada, December 8–14, 2019)

[25] Lopez, R., Regier, J., Jordan, M. I., & Yosef, N. (2018). *Adv. Neural Inf. Process. Syst.* (Vol. 31, pp. 6114–6125). Curran Associates, Inc. (Conference held in Montréal, Canada, December 3–8, 2018). doi:10.48550/arXiv.1805.08672

[26] Bigdeli SK, Ghazisaedi M, Ayyoubzadeh SM, et al. (2025). 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 Med. Inform. Decis. Mak.*, 25(1). doi:10.1186/s12911-024-02799-3

[27] He L, Lu L, Zhang X, et al. (2025). Cross-talk between mitophagy pathways in pre-eclampsia and gestational diabetes mellitus: A systematic analysis of shared molecular mechanisms. *Eur. J. Med. Res.*, 30(1). doi:10.1186/s40001-025-02823-w

[28] Wang X, Wang Y, Zhang S, et al. (2022). Analysis and prediction of gestational diabetes mellitus by the ensemble learning method. *Int. J. Comput. Intell. Syst.*, 15(1), 1–20. doi:10.1007/s44196-022-00110-8

[29] Tripathi KD, Rajput MVS (2020). Fraudulent Credit Card Transactions Classification using Randomized Search CV with XGB Classifier. *Int. J. Eng.*, 6(2), 123–130. doi:10.21203/rs.3.rs-1722294/v1

[30] Karen IM (2016). Area under the receiver operating characteristic (AUROC). In: *Encyclopedia of Biostatistics* (pp. 1–6). doi:10.1001/jama.2016.0288

[31] Lundberg SM, Lee SI (2017). A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.*, 30, 4765–4774. doi:10.48550/arXiv.1705.07874

[32] Jiang Z, Chen X, Lai Y, et al. (2025). Early prediction of postpartum dyslipidemia in gestational diabetes using machine learning models. *Sci. Rep.*, 15, 8028. doi:10.1038/s41598-025-92299-9

[33] Zhou F, Ran X, Song F, et al. (2024). A stepwise prediction and interpretation of gestational diabetes mellitus: Foster the practical application of machine learning in clinical decision. *Heliyon*, 10(12), e22345. doi:10.1016/j.heliyon.2024.e32709

[34] Werner K, Lin TK, Risko N, et al. (2021). The costs of delivering emergency care at regional referral hospitals in Uganda: A micro-costing study. *BMC Health Serv. Res.*, 21(1), 789. doi:10.1186/s12913-021-06197-7

[35] Janevic T, Zeitlin J, Egorova N, et al. (2018). The role of obesity in the risk of gestational diabetes among immigrant and US-born women in New York City. *Ann. Epidemiol.*, 28(1), 1–7.e1. doi:10.1016/j.annepidem.2017.12.005

[36] Will JS, Crellin H (2023). Gestational Diabetes Mellitus: Update on Screening, Diagnosis, and Management. *Am. Fam. Physician*, 108(3), 249–258. PMID: 37725457

[37] Hu J, Gillies CL, Lin S, et al. (2021). Association of maternal lipid profile and gestational diabetes mellitus: A systematic review and metaanalysis of 292 studies and 97,880 women. *eClinicalMedicine*, 34, 100830. doi:10.1016/j.eclinm.2021.100830

[38] Zheng Y, Hou W, Xiao J, et al. (2022). Application Value of Predictive Model Based on Maternal Coagulation Function and Glycolipid Metabolism Indicators in Early Diagnosis of Gestational Diabetes Mellitus. *Front. Public Health*, 10, 850191. doi:10.3389/fpubh.2022.850191

[39] Dong C, Gu X, Chen F, et al. (2020). The variation degree of coagulation function is not responsible for

extra risk of hemorrhage in gestational diabetes mellitus. *J. Clin. Lab. Anal.*, 34(4), e23129. doi:10.1002/jcla.23129

[40] Zhao Y, Zhang J, Zhang J, et al. (2011). Diabetes mellitus is associated with shortened activated partial thromboplastin time and increased fibrinogen values. *PLoS One*, 6(1), e16470. doi:10.1371/journal.pone.0016470

[41] Li X, Weber NC, Cohn DM, et al. (2021). Effects of Hyperglycemia and Diabetes Mellitus on Coagulation and Hemostasis. *J. Clin. Med.*, 10(11), 2419. doi:10.3390/jcm10112419

[42] Ozbasli E, Takmaz O, Karabuk E, Gungor M (2020). Comparison of factor XII levels in gestational diabetes, fetal macrosomia, and healthy pregnancies. *BMC Pregnancy Childbirth*, 20(1), 752. doi:10.1186/s12884-020-03455-0

[43] Cubillos G, Monckeberg M, Plaza A, et al. (2023). Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. *BMC Pregnancy Childbirth*, 23(1), 469. doi:10.1186/s12884-023-05766-4

[44] Sikora A, Zhang T, Murphy DJ, et al. (2023). Machine learning vs. traditional regression analysis for fluid overload prediction in the ICU. *Sci. Rep.*, 13, 19654. doi:10.1038/s41598-023-46735-3

[45] Ye Y, Xiong Y, Zhou Q, et al. (2020). Comparison of machine learning methods and conventional logistic regressions for predicting gestational diabetes using routine clinical data: A retrospective cohort study. *J. Diabetes Res.*, 2020, 8894592. doi:10.1155/2020/8894592

[46] Lu HY, Ding X, Hirst JE, et al. (2024). Digital Health and Machine Learning Technologies for Blood Glucose Monitoring and Management of Gestational Diabetes. *IEEE Rev. Biomed. Eng.*, 17, 98–117. doi:10.1109/RBME.2023.3242261