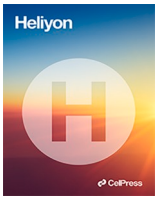
Heliyon 10 (2024) e32709

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A stepwise prediction and interpretation of gestational diabetes mellitus: Foster the practical application of machine learning in clinical decision

Fang Zhou [b,1], Xiao Ran [d,e,1], Fangliang Song [d], Qinglan Wu [b], Yuan Jia [b], Ying Liang [d], Suichen Chen [b], Guojun Zhang [b], Jie Dong [a,*], Yukun Wang [b,c,**]

[a] Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, 410083, PR China
[b] Department of Pharmacy, Southern University of Science and Technology Hospital, Shenzhen, Guangdong, 518055, PR China
[c] Department of Pharmacology, School of Medicine, Southern University of Science and Technology, Shenzhen, Guangdong, 518055, PR China
[d] School of Food Science and Engineering, Central South University of Forestry and Technology, Changsha, 410004, PR China
[e] SINOCARE Inc., Changsha, 410004, PR China

ARTICLE INFO

Keywords:
Gestational diabetes mellitus
Machine learning
Risk prediction
Model explanation
Computational medicine

ABSTRACT

Background: Machine learning has shown to be an effective method for early prediction and intervention of Gestational diabetes mellitus (GDM), which greatly decreases GDM incidence, reduces maternal and infant complications and improves the prognosis. However, there is still much room for improvement in data quality, feature dimension, and accuracy. The contributions and mechanism explanations of clinical data at different pregnancy stages to the prediction accuracy are still lacking. More importantly, current models still face notable obstacles in practical applications due to the complex and diverse input features and difficulties in redeployment. As a result, a simple, practical but accurate enough model is urgently needed.

Design and methods: In this study, 2309 samples from two public hospitals in Shenzhen, China were collected for analysis. Different algorithms were systematically compared to build a robust and stepwise prediction system (level A to C) based on advanced machine learning, and models under different levels were interpreted.

Results: XGBoost reported the best performance with ACC of 0.922, 0.859 and 0.850, AUC of 0.974, 0.924 and 0.913 for the selected level A to C models in the test set, respectively. Tree-based feature importance and SHAP method successfully identified the commonly recognized risk factors, while indicated new inconsistent impact trends for GDM in different stages of pregnancy.

Conclusion: A stepwise prediction system was successfully established. A practical tool that enables a quick prediction of GDM was released at https://github.com/ifyoungnet/MedGDM.This study is expected to provide a more detailed profiling of GDM risk and lay the foundation for the application of the model in practice.

* Corresponding author. Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, 410083, PR China.
** Corresponding author. Department of Pharmacy, Southern University of Science and Technology Hospital, Shenzhen, Guangdong, 518055, PR China.
E-mail addresses: jiedong@csu.edu.cn (J. Dong), wangyk@sustech.edu.cn (Y. Wang).
[1] The first two authors contributed equally to this article.

https://doi.org/10.1016/j.heliyon.2024.e32709
Received 6 June 2023; Received in revised form 22 April 2024; Accepted 7 June 2024
Available online 8 June 2024

1. Introduction

Gestational diabetes mellitus (GDM) refers to abnormal glucose metabolism with first recognition during pregnancy, which is one of the most common complications in pregnant women [1]. GDM will seriously endanger the health of mothers and infants. It may increase the probability of adverse pregnancy outcomes in women in the short term, and increase the risk of metabolic diseases and adverse cardiovascular diseases in GDM women and their offspring in the long term [1–4]. In recent years, with the development of society and the change in people's living and dietary habits, the proportion of obese women before pregnancy and elderly pregnant women is increasing year by year [5]. Moreover, GDM shows an ever-increasing incidence rate among pregnant women [6,7]. As shown by the diabetes map (the 9th edition) published by the International Diabetes Federation (IDF) in 2019, nearly 223 million women suffer from diabetes, and this number is expected to increase to 343 million by 2045, of which 1/6 pregnancies are affected by GDM [8]. Therefore, as an increasingly serious public health problem involving a large number of patients, GDM needs to be urgently solved [9–11]. At present, the routine screening of GDM among pregnant women is mostly performed at 24–28 weeks of pregnancy, with the oral 75 g glucose tolerance test (OGTT) during the third trimester of pregnancy. However, accumulating studies have demonstrated that the hyperglycemia environment in pregnant women may have adverse effects on the fetus before GDM diagnosis [12,13]. It has been confirmed that early pregnancy intervention helps to reduce GDM incidence, which will notably decrease maternal and infant complications and improve the prognosis [14,15]. Therefore, early prediction and intervention of GDM are of great importance.

At present, statistical analysis and machine learning methods for constructing early GDM prediction models are gaining attention worldwide. These methods can be used to analyze the risk factors of GDM, conduct GDM screening in early pregnancy, and identify high-risk pregnant women, thereby providing references for early GDM intervention. Recently, multiple studies have reported the application of the machine learning method for constructing the GDM prediction model. From a time perspective, in 2003, HCJ et al. identified GDM and abnormal glucose tolerance (IGT) among Korean women through some prenatal factors using the logistic regression (LR) method [16]. KVS et al. tried to assess the occurrence of GDM in early pregnancy using serum biomarkers in 2007 [17]. After 2010, more studies were reported to construct GDM prediction models with different emphases using machine learning methods [18–20]. From the perspective of the included factors or so-called features, growing studies have shown that the construction of early GDM risk prediction models is mainly based on a single or a class of biomarkers [16,17,21]. Later, multiple relevant factors (such as biomarkers, clinical electronic medical records, and patient's personal information) were integrated into the models [22,23]. In terms of model performance, some earlier models often exhibit poor performance due to limited data or features. Most of these models show effectiveness via evaluating the significance of factors, and only a few models have reported general indicators. For example, Jo et al. have reported a model using the pre-pregnancy weight for GDM prediction, with a sensitivity (SE) of 47.8 % and a specificity (SP) of 65.9 % [21]. In later models, the model performance has been improved by using more advanced machine learning algorithms combined with multi-level features. For example, APN et al. have reported a multivariate logistic model with an area under the receiver operating characteristic (ROC) curve (AUC) value of 0.64 [22]. The AUC of the random forest (RF) model and LR model reported by Wang J et al. is 0.777 ± 0.034 and 0.755 ± 0.032, respectively [24]. Wang X et al. have reported an AUC value of 0.748 in the multivariate LR model [25]. Recently, Wu Y et al. reported a risk prediction model for GDM among Chinese pregnant women before 16 weeks of gestation, with an AUC value of 0.746 [26]. These studies and models try to analyze and predict GDM risks from different perspectives. In order to make readers understand these models more clearly, we summarized their features in Table S1.

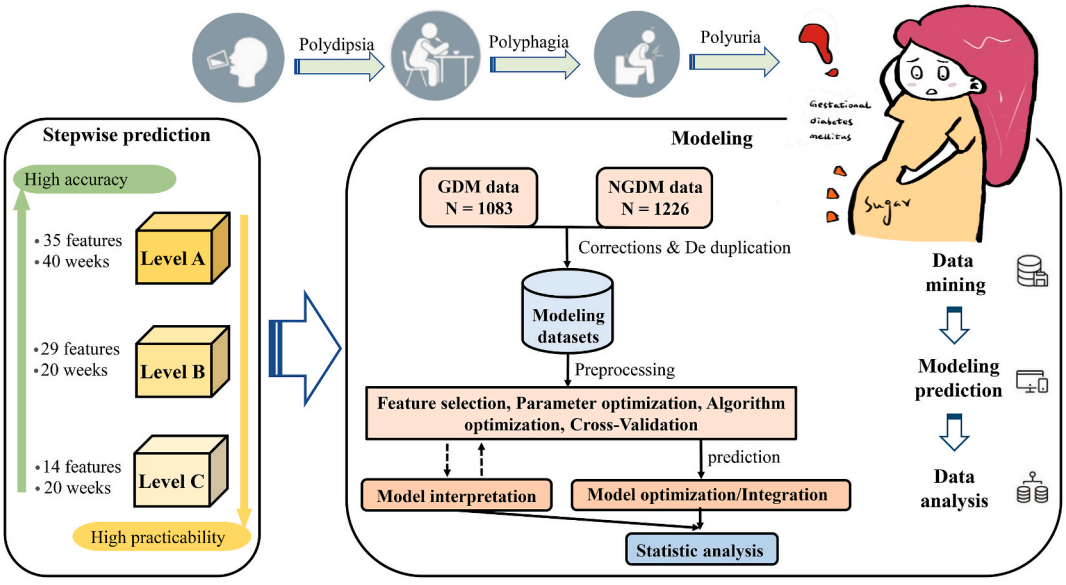


Fig. 1. The overview of our methodology.

Taken together, according to the above results, the model may show poor performance in the first trimester; although incorporating more complex factors into the model may increase model accuracy, meanwhile, it will reduce model practicability. Additionally, the performance of the current model is not high, with a primary AUC value of about 0.7. More importantly, although the most important significance and goal of GDM prediction are to provide clinicians with available assistance in decision-making, there is still a lack of tools that clinicians can use directly. Hence, at present, how to construct a robust prediction model with high performance and high practicability is an important problem that needs to be solved urgently. This study attempted to construct a robust and accurate prediction model by building a high-quality GDM dataset and systematically comparing various advanced machine learning algorithms; the feature space was broadened by incorporating the multi-level features (including demographic, clinical, blood routine indicators, and patient social attributes). Based on the theory of interpretable machine learning, the importance of the included indicators was assessed and explained to evaluate the performance of the model and the practicability of the indicators. Finally, a model that can be directly used by clinicians was obtained, which was conducive to GDM diagnosis. The expectation was not limited to better model performance, more detailed profiling of GDM risk and interpretation than existing reports but rather explores practical application value. An overview of our methodology was presented in Fig. 1.

2. Materials and methods

2.1. GDM patients and diagnostic criteria

The data of patients in two public hospitals in Shenzhen, China from June 2019 to July 2021 were collected through the prescription automatic screening system (PASS), with "GDM" as the screening diagnosis. According to the diagnostic criteria of GDM revealed by the "*ADA Diabetes Medical diagnosis and treatment Standard* (2018)" and the 9th edition textbook of "*Internal Medicine*", after OGTT examination at 24–28 weeks of pregnancy, blood glucose exceeds 5.1 (before taking sugar), 10.0 (1 h after taking sugar), or 8.5 mmol/L (2 h after taking sugar) will be diagnosed as GDM. Pregestational diabetes mellitus (PGDM) was excluded. Pregnant women with no GDM in the same period were randomly selected as the no GDM (NGDM) group. The features that we collected for each patient included demographic characteristics, clinical characteristics (scale) and indicators from blood routine examination. This study was approved by the institutional review committee of the hospital (Ethics Committee of Southern University of Science and Technology Hospital, Number: 2022–09). As this is a retrospective study and the data are desensitized, the informed consent is exempted.

2.2. Data preprocessing

Data quality is the key determinant of a machine learning model. Based on the above criteria and information, 1130 samples with 48 features were obtained and considered as the GDM group; 1226 samples with 67 features were considered as the NGDM group. After rough deduplication and inspection, we obtained the first version of the dataset: 1083 samples for GDM and 1226 samples for NGDM, each with 40 features (Table S2).

Subsequently, the source and value of each sample were manually checked in detail. Firstly, 202 data were found quite different from the overall distribution, which was then manually checked. Among 101 entries in the GDM dataset, 4 data were abnormal with neutrophil ratio (NEUT) $>$ white blood cells (WBC), 21 data deviated from the normal range, and 76 data were abnormal data with words or symbols at the time of entry. For 95 entries in the NGDM dataset, 6 data were abnormal for pregnant women with NEUT $\geq$ WBC, 15 data were abnormal for the weight difference between the Pre_preg_weight and Birth_weight of pregnant women exceeding 50 kg, 66 data were Pre_preg_BMI marked as 0, 1 data was out of the normal range, and 7 data were mistaken by inputs. These data were corrected after checking the original records through the steps described above. Specifically, "Occupation" and "Pay" were classified into 6 types according to their actual meanings (Table S2). The continuous feature "anti_TPOAb" was set as 19 bins according to reasonable distribution intervals, and then we imputed missing values according to the specific meaning of each column. For the category features (such as "1st_T2DM", "Hypertension", "Smoking", and "Drinking"), we used mode to fill the missing values; for the continuous features (such as "Age", "Weight", and "Height"), we used the average value. Finally, 1083 GDM samples and 1226 NGDM samples were used for model building. Prior to modeling the data, we preprocessed the data, while removing features with low variance and high correlation, following best practice recommendations. Then the recursive feature elimination (RFE) algorithm was used for feature selection. Recursive feature elimination is an efficient and commonly used method. Cross-validated recursive feature elimination (RFECV) iteratively selects subsets of features to identify optimal sets.

2.3. Machine learning algorithms

To obtain the best model, we compared multiple machine learning algorithms. In this study, we not only tried simple and/or interpretive algorithms [such as LR, decision trees (DT), and k-Nearest Neighbors (KNN)], but also used advanced algorithms [such as RF, support vector machine (SVM), and XGBoost] to construct models with better performance. LR is a simple linear classifier, which constructs models that are interpretable for simple problems; the impact on the final result can be reflected by the weight of each feature [27]. DT algorithm is a method to approximate the value of the discrete function, which can construct a model of the tree structure [28]. When used for classification tasks, DT is naturally interpretable and has no need for complex parameter tuning, but it is very sensitive to outliers and has poor generalization ability. RF is an algorithm that integrates multiple trees through ensemble learning, with a decision tree as the basic unit; it combines the independent decisions from multiple decision trees to improve the overall performance [29]. RF can process a large number of data and has good generalization ability and strong anti-over-fitting

ability. Additionally, the RF model can explain the model while ensuring the model's accuracy by outputting the importance of each feature. SVM attempts to fit the margin maximization between two categories and to find the optimal hyperplane separating two different classes of data for decision-making [30]. SVM is a novel and applicable few-shot learning method, and the computational complexity depends on the number of support vectors rather than the dimension of the sample space, thereby avoiding the "curse of dimensionality" in a sense and showing better robustness and generalization performance. XGBoost, as an implementation of the boosting algorithm, can perform classification or regression tasks, which is also known as extreme gradient boosting trees [31]. XGBoost has good performance in many tabular data tasks because it not only has a good modeling effect and fast speed but also can avoid overfitting by tuning parameters. In the present study, the relevant algorithm functions were used through the *scikit-learn* library (https://scikit-learn.org/stable/) in the Python environment. The optimal model was selected through lattice search and manual fine-tuning. The parameters that needed to be tuned were summarized in Table S3.

2.4. Model evaluation

In the process of model construction, the data were split into the training set and test set at random (8:2). Moreover, a 10-fold cross-validation (CV) was performed on the training set to further evaluate the stability and actual prediction ability of the model. The evaluation metrics for classification models are usually the accuracy (ACC), SP, SE, and AUC of the model. ACC represents the overall evaluation accuracy of the model; SE and SP can be used to evaluate the classification ability of positive and negative samples, respectively; and AUC can be used to comprehensively evaluate the overall performance of the model without being affected by sample imbalance. The threshold of probability for classification was set as 0.5. The descriptions for each evaluation metric are as follows:

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

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

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

While the TP, TN, FP and FN represent the "true positive", "true negative", "false positive" and "false negative" in the "Confusion Matrix" yielded by the classification modeling. When applied to a population, the accuracy will be as follows: Accuracy (population) = Prevalence × Sensitivity + (1 - Prevalence) × Specificity, where the "Prevalence" represents the occurrence rate of GDM in the target application environment.

2.5. Model explanation

The interpretation of the model is to explain the reason for model performance through certain mathematical methods and to indicate the contribution of various factors to the model. In this research, the currently advanced tree-based feature importance and additive feature attribution method SHAP (https://pypi.org/project/shap/) were adopted to explain the size and trend of the influence of each feature on different models, which determines the importance of the feature via evaluating the prediction error changes before and after adding noise and calculating the individual contribution to the population based on the idea of game theory. The biggest advantage of the SHAP method is that it can not only evaluate the importance of features but also report the trend (positive and negative) of the impact of features on the prediction results.

3. Results

In this study, three models at different levels were constructed. First of all, all features were used to construct the model (level A) to find the features showing the greatest impact on GDM and the model with the highest accuracy. Secondly, the features from the former 20 weeks in pregnancy were selected to construct the model (level B) to enable an early prediction of GDM. Finally, the features on the level-B models were further simplified while ensuring a relatively good performance to obtain a model (level C) with good operability and accuracy that enables less examining items as input in clinical decision-making scenarios.

3.1. Model performance

In this study, we divided our data set into independent training and test sets. Training set was used to build the model, we selected the optimal model parameters and realized model selection based on the test dataset. The test dataset was employed to evaluate the generalization capability of the final model. The data of 2309 GDMs were randomly divided into the training and test sets containing 1847 and 462 data. For the level-A models, 40 features were fed into the RFECV process in this study. Two representative base evaluators (RF and SVM) were chosen to evaluate the changes in model performance during feature selection.

As revealed by the results, the value of ACC reached a maximum when the number of features was 35 with RF as the base evaluator and 34 with SVM as the base evaluator (Fig. S1). Reducing the number of features from the optimal number of features did not increase the predictive accuracy of the model. For level-A models, based on the best-selected features, different algorithms were used to

construct a series of machine learning models, and the results were summarized in Table 1.

As shown in Table 1 and Table S4, the models had good overall performance; the 10-fold CV set showed similar results to the test set, which indicated the reasonably split dataset and the credibility of the constructed model. The results obtained using different algorithms were of certain difference. The XGBoost model achieved the best results both in the 10-fold CV (ACC = 0.891; AUC = 0.955) and test set (ACC = 0.922; AUC = 0.974) (Fig. 2A). The KNN model for the test set showed the comparatively worst performance (ACC = 0.803, AUC = 0.860).

The level-B models incorporated 33 features from the first 20 weeks of pregnancy. Firstly, the recursive feature deletion method was used for feature selection. As shown by the results, the ACC value was the largest when the number of features was 29 (RF as the base evaluator) or 28 (SVM as the base evaluator) (Fig. S1). The prediction accuracy of the model was not increased by decreasing the number of features from the optimal number. Finally, 29 features were selected for model construction combining with multiple machine learning algorithms. The results were summarized in Table 1 and Table S5.

The results in Table 1 and Table S5 indicated that the models exhibited good overall performance; the CV and test set results were consistent. Among these models, XGBoost performed the best both in the CV (ACC = 0.842; AUC = 0.912) and test set (ACC = 0.859; AUC = 0.924) (Fig. 2B). The SE and SP were more close than other models. KNN still performed the worst for the test set (ACC = 0.741; AUC = 0.808).

Based on level-B models, we found that if we used the RF method for recursive feature deletion; when the number of features reached 14, a further increase in the number of features would not significantly increase the ACC value. Therefore, 14 features from the first 20 weeks of pregnancy were screened and used to construct level-C models combined with multiple machine learning algorithms. The results were summarized in Table 1 and Table S6. All details about the selected features of level-A, level-B and level-C models were listed in Table S7.

The results in Table 1 and Table S6 showed that the models had good overall performance; CV and test sets showed consistent results. Among these models, XGBoost still performed the best both for the CV (ACC = 0.861; AUC = 0.922) and test set (ACC = 0.850; AUC = 0.913) (Fig. 2C). KNN still performed the worst performing for the test set (ACC = 0.716; AUC = 0.796). Compared with level-B models, level-C models showed slightly fluctuated performance of the same modeling, as shown by a small improvement in the CV set of the XGBoost model, and slightly weakened data of other models. In summary, the model with 29 features (level B) was more accurate, but the model with 14 features (level C) achieved its intended purpose and thus can also be used to drive an operational application for clinical use.

During the level A-C modeling process, it was found that the accuracy of the models from high to low was XGBoost > RF > DT > SVM > LR > KNN. According to their prediction scores on the test set, the tree-based ensemble model showed the most prominent performance; the RF model showed only slightly lower prediction accuracy than the XGBoost model, because different algorithms can compose the optimal modeling feature subset, and selecting different features may produce similar results. Additionally, the precision of other models for positive and negative samples was relatively balanced except for the obvious deviation of SE and SP of the KNN model; the deviation may be attributed to the systematic error generated during data collection. The level-A prediction models showed relatively higher accuracy than level-B and level-C models because appropriately increasing the number of features is conducive to improving the learning ability of the machine model. The score of the test set of level-A models was higher than that of the CV set; however, for level-B and level-C models, most of the model test sets showed lower scores than CV sets; for XGBoost with the highest model accuracy, the test set score was higher than the CV set score when 29 features were selected, but the Test score was lower than the CV score when 14 features were selected. It was noticed that although level-B models had slightly improved overall model accuracy relative to the level-C models, the CV set score of the XGBoost model of level C was higher than that of level B, which may be attributed to data splitting of different groups.

3.2. Risk factor evaluation

Fig. 3 showed the feature importance of the three models with better performance at level A and level B, respectively. The definitions of these features can be found in Table S2. It should be noted that GDM in the plots reports the "History of Gestational Diabetes Mellitus". The first three features with the greatest impact were high-density lipoprotein (HDL), Pay, and total cholesterol (TC) for level-A models (Fig. 3A) and Pay, GDM, and FBG for level-B models (Fig. 3B). First of all, it can be seen that features [Pay, FBG, GDM,

Table 1
The performance of the top 3 best-selected models after comparing different algorithms.

	Model	Test				10-CV			
		SE	SP	ACC	AUC	SE	SP	ACC	AUC
Level A	**XGBoost**	**0.894**	**0.945**	**0.922**	**0.974**	**0.858**	**0.921**	**0.891**	**0.955**
	RF	0.773	0.859	0.820	0.894	0.790	0.891	0.843	0.906
	LR	0.754	0.894	0.831	0.884	0.759	0.894	0.831	0.887
Level B	**XGBoost**	**0.821**	**0.890**	**0.859**	**0.924**	**0.821**	**0.862**	**0.842**	**0.912**
	RF	0.773	0.859	0.820	0.894	0.790	0.891	0.843	0.906
	SVM	0.720	0.804	0.766	0.837	0.724	0.832	0.781	0.856
Level C	**XGBoost**	**0.802**	**0.890**	**0.850**	**0.913**	**0.815**	**0.903**	**0.861**	**0.922**
	RF	0.778	0.859	0.823	0.887	0.784	0.884	0.837	0.899
	DT	0.667	0.906	0.799	0.842	0.715	0.846	0.783	0.846

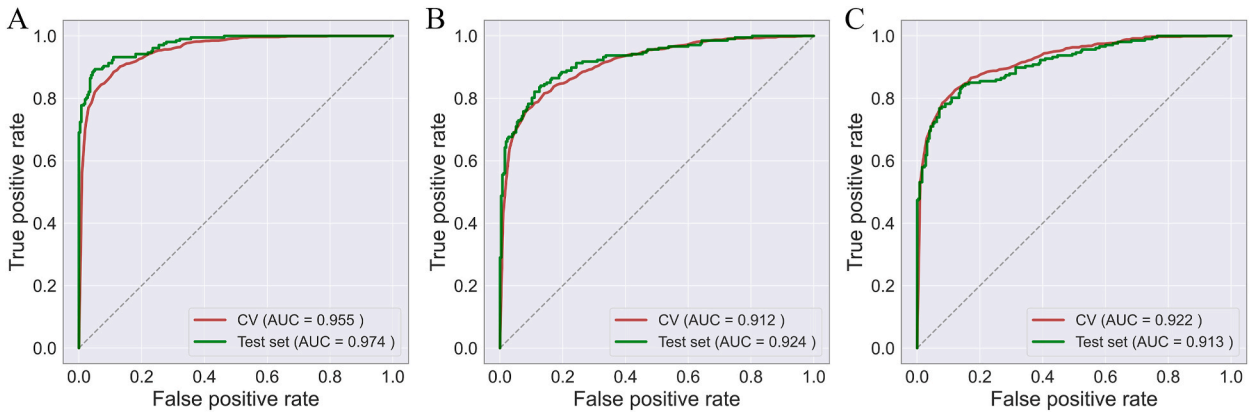


Fig. 2. The performance of the selected best models of different model levels. (A), The AUC of the selected XGBoost model of level A. (B), The AUC of the selected XGBoost model of level B. (C), The AUC of the selected XGBoost model of level C.

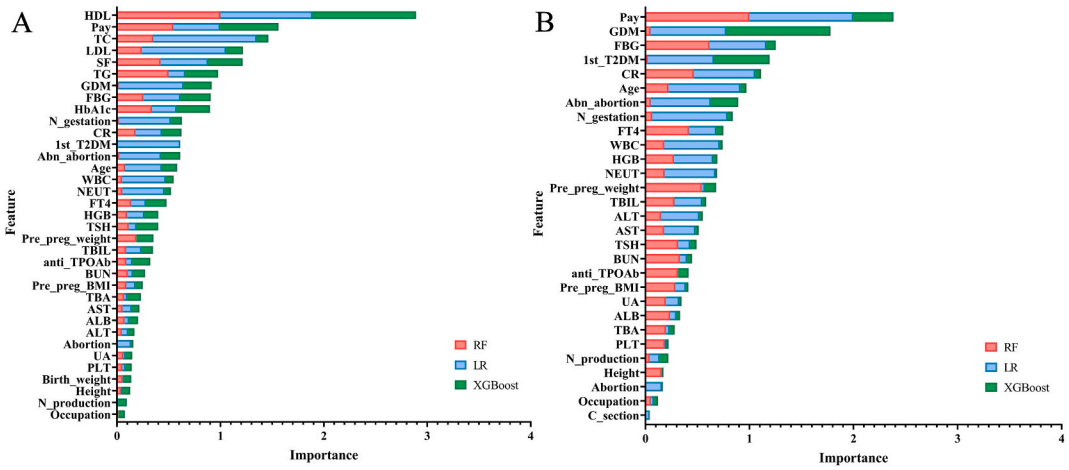


Fig. 3. The feature importance of models of level A (A) and level B (B) using different algorithms.

N_gestation, Abn_abortion, WBC, thyroid hormone (FT4), and creatinine (CR)] showed consistent importance in the models of level A and level B, indicating that these features were of great importance to GDM evaluation. Secondly, although blood lipid-related indexes [such as triglyceride (TG) and TC], lipoprotein-related indicators [such as HDL and low-density lipoprotein (LDL)], and indicators [such as serum ferritin (SF), HbA1c, and Birth_weight] were not included in level B, the accuracy was not reduced, which indicated that other feature subsets can also construct accurate models, and GDM cannot be simply affected by regular features. Moreover, the importance of "Pay" as a special indicator was observed. As indicated by the results, the patient's economic condition and social environment exhibited quite important impacts on GDM. "Pay" is complex and integrates factors concerning the economic and living environment, which certainly affect the physical function in a concealed and long-term manner, so it is worth further exploration and research. According to the ratio of the importance of different colors, the important features displayed in different algorithms were more consistent; however, there were also differences in the importance of features (such as 1st_T2DM and N_gestation), which indicated that the model accuracy can be ensured based on multiple feature subsets.

Fig. 4 showed the feature importance calculated by SHAP values for level-A and level-B models, respectively. Firstly, it was observed that the feature importances based on trees and regression coefficients in Fig. 3 were more consistent than those based on SHAP. In the first 10 important features, some features (HDL, Pay, SF, LDL, TG, and HbA1c) were common for level-A models, and some features (Pay, FBG, Age, N_gestation, and FT4) were common for level-B models. However, features (1st_T2DM, GDM, and TC) in level-A models and features (1st_T2DM, GDM, and Abn_abortion) in level-B models showed a small ratio of importance in the SHAP graph but a high ratio in Fig. 3, which may be attributed to different importance scoring mechanisms; however, different scoring mechanisms can still rank the most important features as expected. Secondly, as revealed by Figs. 3 and 4, HDL, Pay, SF, and TG were the features showing the greatest impact on level-A models, and Pay, FBG, Age, and FT4 were the features showing the greatest impact on level-B models [32]. It can be seen that the effects of FBG, Pre-pregnancy weight, and FT4 in early pregnancy were more closely related to GDM than other eigenvalues. Consistently, previous research has also demonstrated that FBG in the first trimester can be used as a screening test to identify pregnancies with GDM risk [33]. Finally, based on distinguishing the color of the feature

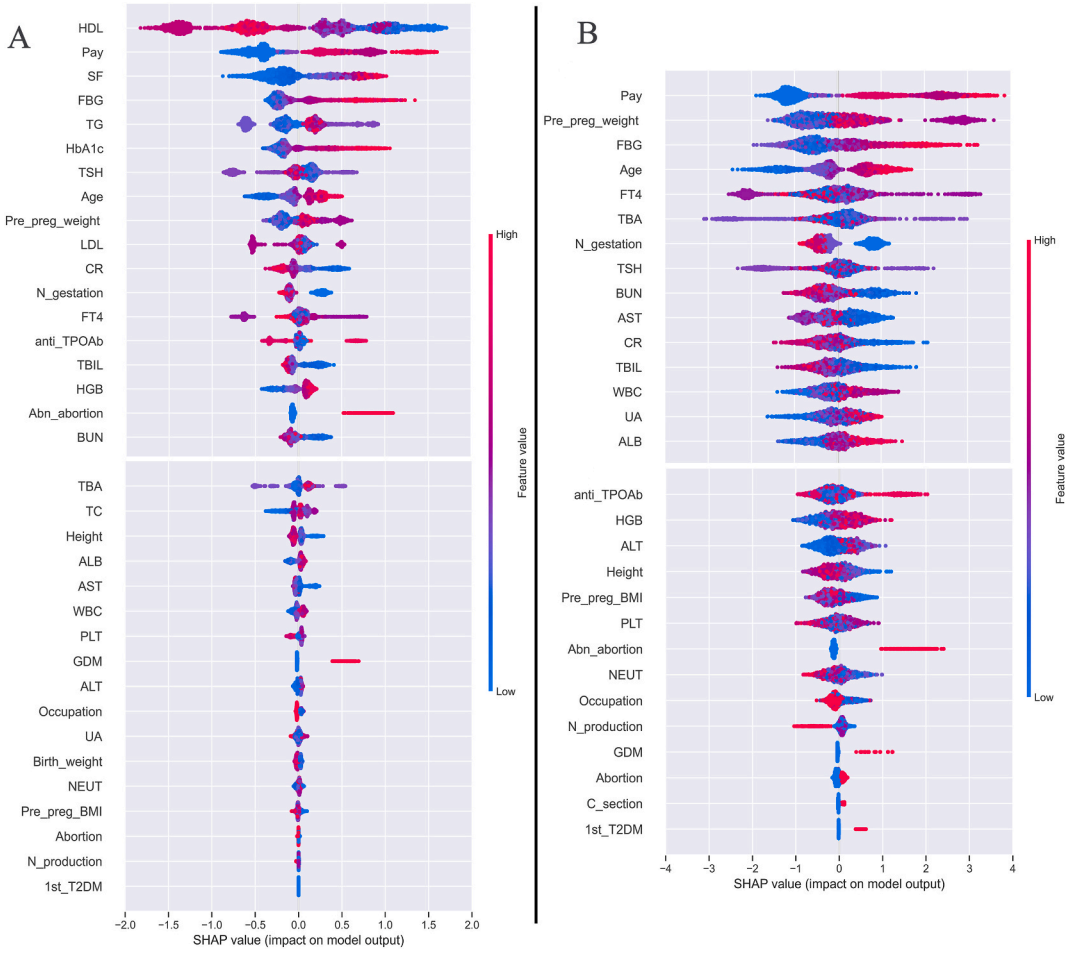


Fig. 4. The plots of SHAP values for the selected models. (A), the 35 features for the level A model. (B), the 29 features for the level B model.

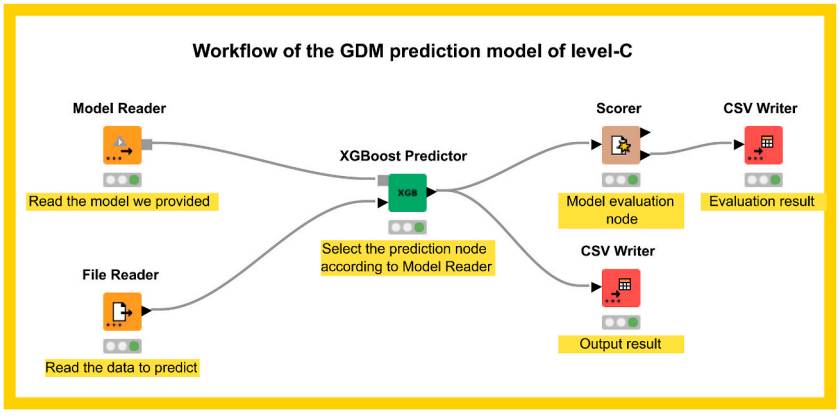


Fig. 5. The snapshot of the workflow for the model of level-C. This pipeline enables the prediction by uploading a data file containing multiple samples (a ".csv" file is recommended) to enable prediction and statistics. The detailed explanations of the 14 features and how to read the results were listed in the workflow file in the repository.

importance, the SHAP plot was used to show the positive and negative correlation between the feature and the model prediction results, which can more intuitively judge the influence trend of these features on the model.

3.3. The prediction tool based on KNIME software

In order to enable researchers or clinicians to quickly and conveniently predict GDM risk, we have developed a simple tool based on KNIME software (https://github.com/ifyoungnet/MedGDM). This tool implemented the level-C model (Fig. 5). This model considered both practicality and accuracy. The ACC of 0.850 and AUC of 0.913 can be reached in the test set only by using 14 indicators that are very easy to obtain. In this tool, we provided three pipelines for different usage scenarios. It allows users to directly input the corresponding features in the table for prediction, and also allows users to upload a general data file containing multiple samples for simultaneous prediction of multiple samples. In addition, it also supports the input of data with known labels for prediction and statistics of the accuracy of new predictions.

4. Discussion

The phenomenon and trends in the results were further explored and discussed. First of all, since HDL plays a role not only in the level of pancreas and insulin secretion but also in glucose uptake levels by skeletal muscle, increasing HDL can reduce the risk of diabetes to a certain extent; Yi Wang and Wu et al. have reported consistent results [34,35]. Moreover, as demonstrated by Wu Y et al., high maternal TG level and low HDL level are significantly implicated in the increased risk of GDM, while a high HDL level acts as a protective factor [26]. Our results showed that TG and TC levels were positively correlated with the occurrence of GDM, and TG with a high value was a crucial risk factor for GDM. Accordingly, similar results have been demonstrated by Wang, Yi Wang, and Wu et al. [25, 34,35] Secondly, as revealed by the results of level-A models with more features in the middle and late stage of pregnancy, blood lipid indicators showed the greatest correlation with GDM; the blood lipid level during pregnancy increased gradually with the increase of gestational age and reached its peak in the third trimester. It's widely accepted that elevating blood lipid levels within a certain range is conducive to providing energy for the normal development of the fetus and reserving energy for pregnancy, childbirth, and postpartum lactation, which is a normal physiological phenomenon. Inflammatory factors and adipokines in the body contribute to vascular endothelial dysfunction in GDM people, which causes weakened effects of insulin on fatty acids, thereby resulting in abnormal glucose metabolism. Additionally, lipid metabolism in these people also changed to varying degrees; despite body mass index (BMI) showing greater impact than TG in early pregnancy, TG showed notably increased importance with the change of time. Consistently, studies have also confirmed that abnormal lipid metabolism is tightly implicated in patients with GDM. For example, Klop B et al. have shown that pregnant women with GDM present aberrantly increased levels of lipid metabolism indexes (including TC, TG, and LDL) and decreased HDL level in the first trimester. Recent studies have also provided evidence for the relationship between blood lipids and GDM [36]. Furthermore, our model results revealed that age was significantly positively correlated with GDM. Similarly, accumulating studies have evidenced these results [24–26,37]. We also proposed the complex relationship between BMI and GDM. As has been pointed out previously, maternal age and BMI before pregnancy are remarkably positively correlated with GDM risk [37]. Kautzky-Willer et al. have reported that the mechanisms of insulin resistance and defective insulin secretion are intrinsically associated with high BMI and low BMI in individuals with GDM, respectively [38]. Both elevated BMI and low BMI ($\leq$17) are risk factors for GDM [34]. In the present study, the pre-pregnancy BMI was found to be positively associated with the risk of GDM. Additionally, Zhang CJ et al. have proposed that blood lipid-related indicators are more accurate than BMI for prediction, which also explains the low importance of BMI in this study. In this project, pre-pregnancy weight, pre-pregnancy height, pre-pregnancy BMI, and birth weight on physical examination were considered when we collected the features. Multiple previous studies have evidenced the close relationship between pre-pregnancy BMI and GDM occurrence; the pre-pregnancy body weight represents the visceral fat content in pregnant women, with a heavier body weight indicative of more visceral fat; the number of islet receptors per unit area of fat cells is relatively reduced, and the sensitivity to insulin decreases. Failure to compensate during pregnancy leads to elevated blood sugar and disordered glucose metabolism. Our results revealed that pre-pregnancy weight was positively associated with GDM. HbA1c is formed by the combination of glucose and hemoglobin (HGB). As a slow, continuous, and irreversible non-enzymatic reaction, HbA1c formation is not only not controlled by exercise or things, showing good stability, but also reflects the average blood sugar level in the past 3 months. Our model results also indicated that HbA1c exhibited a positive correlation with GDM [39–41]. Besides, FT4 was a significant risk factor for GDM in early pregnancy and was negatively correlated with GDM occurrence, which was consistent with previous studies. As has been indicated by a previous study on the Chinese population, increased FT4 level may enhance the protective mechanism of GDM, as evidenced by the findings that higher FT4 levels are associated with a lower incidence of GDM [42]. There was also a certain correlation between TSH and anti-TPOAb in thyroid function indicators. Luo J et al. have confirmed that thyroid dysfunction and positive thyroid antibodies are closely related to the risk of GDM [43]. Low FT4 levels are tightly implicated in GDM occurrence in the first and second trimesters. Our study showed that SF in the later stage was positively correlated with GDM occurrence. Durrani L et al. have demonstrated that high levels of maternal SF play an important role in GDM development, showing a positive correlation between high levels of heme iron intake and GDM occurrence [44]. This may be because women with high SF levels have increased insulin resistance and increased pancreatic secretion, which leads to pancreatic beta cell failure; heme iron increases the body's iron reserves and may cause oxidative damage to pancreatic cells. Xiong et al. have shown that platelets (PLT) were significantly elevated, together with higher levels of liver and kidney function variables [glutamyl transpeptidase (r-GT), FBG, and fibrinogen (Fg)] in GDM patients [45]. The levels of total bilirubin (TBIL) and direct bilirubin (DBIL) showed opposite trends. Higher values of these parameters indicated greater risks of GDM. Our results also indicated that FBG was positively correlated with

GDM risk. Consistent results have been reported by Wu and Tong et al. [33,34] Notably, a positive correlation between Pay and GDM risk was observed; Pay from 0 to 5 represented maternity hospitalization, non-local/labor worker medical insurance, hospital medical insurance, comprehensive medical insurance, self-pay, and maternity insurance, respectively. Since these categories do not reflect the absolute economic gap, it can only be seen in this study that maternity insurance was more likely to make positive contributions to GDM than labor medical insurance and out-of-town medical insurance. In addition to the above-mentioned features closely associated with GDM, CR, blood urea nitrogen (BUN), TBIL, age, and HGB are also found to show relatively important effects on GDM. Consistently, multiple studies have recognized age as an independent factor of GDM. However, TBIL, CR, and BUN have not been confirmed to be related to GDM. Furthermore, this study also indicated a certain relationship between liver and kidney function and late GDM.

It is worth noting that this project showed inconsistent SHAP-based impact trends with few publications, mainly involving AST, NEUT, BUN, and CR. Wang Y et al. have previously reported a positive correlation between AST and GDM [37]. However, our study showed that AST was negatively correlated with GDM. The impact of AST on GDM is still controversial. Multiple previous studies have reported that high ALT is a risk factor for GDM, and AST may not be related to GDM [46–48]. BUN and CR are commonly used indicators for renal function evaluation. Feng et al. have proposed that these early renal function indicators are positively correlated with GDM [49]. Different results have been reported by WU Y et al., they have found that there is no significant difference between Asian women with GDM and those without GDM in terms of renal insufficiency [26]. Interestingly, our results showed a negative correlation. NEUT has been identified to be positively correlated with GDM development by SUN T et al. [50] However, a previous study has proposed that there is no difference between the GDM group and the NGDM group in leukocyte count [51]. As an inflammatory factor, NEUT plays a crucial role in regulating various processes and can be affected by many other factors. Our results reflected a weak negative correlation of NEUT with GDM. In future versions, we would improve the research by further clarifying the population and geographical differences, and take some measures (such as increasing the sample size and unifying the detection methods of the instruments) to explain the results of different influences more clearly.

Our feature data were collected on a large scale, not only including routine biomarkers but also other features unrelated to blood biochemical examination. The machine learning algorithms were highly effective and yielded a simple, convenient, and effective screening method for the clinic. However, there are some limitations in this study. Despite the good performance of the models, this study was limited by restricted patient profiles and differences in regions, populations, and collection standards. In particular, these models constructed can only work properly within a specific application domain. This reminds us that there is a trend to build a system that allows models to cover different regions, populations, and application scenarios in different scenarios. It can not only overcome the problems of different feature weights and influence trends but also realize efficient data and model sharing. In this way, computational methods can serve the clinic.

5. Conclusion

GDM seriously threatens the health of pregnant women and fetuses. Therefore, it is of great value to assess and prevent the occurrence of GDM at an early stage, and to provide auxiliary information for clinicians to make GDM diagnosis decisions. This study applied advanced machine learning algorithms to construct three accurate prediction models by incorporating features at different levels and presetting different application scenarios. Meanwhile, advanced interpretability algorithms were used to analyze the feature impact mechanism of different models. The level-A model and level-B models in this work can explain the risk factors of GDM in the whole span and approximate early pregnancy, respectively, while the level-C models that take into account both practicability and accuracy provide models enabling few examining items input in clinical decision scenarios. In future works, we will try to transform this model system and strive to build an integrated multi-center model, which is an essential step for precision medicine.

Data availability

The data that has been used is confidential. The prediction tool based on KNIME software was available at: https://github.com/ifyoungnet/MedGDM.

Ethics declarations

This study was reviewed and approved by the Ethics Committee of Southern University of Science and Technology Hospital, with the approval number: 2022–09. Informed consent was not required for this study because this is a retrospective study and the data were desensitized.

CRediT authorship contribution statement

Fang Zhou: Writing – original draft, Investigation, Funding acquisition, Data curation, Conceptualization. **Xiao Ran:** Writing – original draft, Investigation, Formal analysis. **Fangliang Song:** Visualization, Software. **Qinglan Wu:** Validation, Data curation. **Yuan Jia:** Resources. **Ying Liang:** Visualization, Resources. **Suichen Chen:** Resources. **Guojun Zhang:** Resources. **Jie Dong:** Writing – review & editing, Project administration, Methodology, Conceptualization. **Yukun Wang:** Writing – review & editing, Supervision, Resources.

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.

Acknowledgments

This work was supported by the Shenzhen Science and Technology R&D Fund (20200925160201001); Guangdong Provincial Hospital Pharmaceutical Research Foundation (2023A03); Research Fund of Southern University of Science and Technology Hospital (2020D8, 2020-A2). Hunan Provincial Natural Science Foundation of China (2023JJ70055). We thank all the participants and staff in Huazhong University of Science and Techology Union ShenZhen Hospital for their support in this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e32709.

References

[1] H.D. McIntyre, P. Catalano, C. Zhang, et al., Gestational diabetes mellitus, Nat. Rev. Dis. Prim. 5 (1) (2019) 47.
[2] S.C. Tinker, S.M. Gilboa, C.A. Moore, et al., Specific birth defects in pregnancies of women with diabetes: national birth defects prevention study, 1997–2011, Obstet. Gynecol. Surv. 75 (7) (2020) 385–387.
[3] M. Hod, A. Kapur, D.A. Sacks, et al., The International Federation of Gynecology and Obstetrics (FIGO) Initiative on gestational diabetes mellitus: a pragmatic guide for diagnosis, management, and care, Int. J. Gynecol. Obstet. 131 (2015) S173.
[4] M. Vääräsmäki, A. Pouta, P. Elliot, et al., Adolescent manifestations of metabolic syndrome among children born to women with gestational diabetes in a general-population birth cohort, Am. J. Epidemiol. 169 (10) (2009) 1209–1215.
[5] S.S. Casagrande, B. Linder, C.C. Cowie, Prevalence of gestational diabetes and subsequent Type 2 diabetes among U.S. women, Diabetes Res. Clin. Pract. 141 (2018) 200–208.
[6] L. Bellamy, J.-P. Casas, A.D. Hingorani, et al., Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis, Lancet 373 (9677) (2009) 1773–1779.
[7] E. Vounzoulaki, K. Khunti, S.C. Abner, et al., Progression to type 2 diabetes in women with a known history of gestational diabetes: systematic review and meta-analysis, BMJ 369 (2020) m1361.
[8] Diagnostic criteria and classification of hyperglycaemia first detected in pregnancy: a World Health Organization Guideline, Diabetes Res. Clin. Pract. 103 (3) (2014) 341–363.
[9] A. Basu, Feng Du, P. Planinic, et al., Dietary blueberry and soluble fiber supplementation reduces risk of gestational diabetes in women with obesity in a randomized controlled trial, J. Nutr. 151 (5) (2021) 1128–1138.
[10] G. Putoto, E. Somigliana, F. Olivo, et al., A simplified diagnostic work-up for the detection of gestational diabetes mellitus in low resources settings: achievements and challenges, Arch. Gynecol. Obstet. 302 (5) (2020) 1127–1134.
[11] E.A. Reece, G. Leguizamón, A. Wiznitzer, Gestational diabetes: the need for a common ground, Lancet 373 (9677) (2009) 1789–1797.
[12] S. Thériault, J.-C. Forest, J. Massé, et al., Validation of early risk-prediction models for gestational diabetes based on clinical characteristics, Diabetes Res. Clin. Pract. 103 (3) (2014) 419–425.
[13] ChenHod Yogev, et al., Hyperglycemia and adverse pregnancy outcome (HAPO) study: preeclampsia, Am. J. Obstet. Gynecol. 202 (3) (2010) 255.e1–255.e7.
[14] S.B. Koivusalo, K. Rönö, M.M. Klemetti, et al., Gestational diabetes mellitus can be prevented by lifestyle intervention: the Finnish gestational diabetes prevention study (RADIEL): a randomized controlled trial, Diabetes Care 39 (1) (2016) 24–30.
[15] V. Seshiah, A. Cynthia, V. Balaji, et al., Detection and care of women with gestational diabetes mellitus from early weeks of pregnancy results in birth weight of newborn babies appropriate for gestational age, Diabetes Res. Clin. Pract. 80 (2) (2008) 199–202.
[16] H.C. Jang, C.-H. Yim, K.O. Han, et al., Gestational diabetes mellitus in Korea: prevalence and prediction of glucose intolerance at early postpartum, Diabetes Res. Clin. Pract. 61 (2) (2003) 117–124.
[17] K.V. Smirnakis, A. Plati, M. Wolf, et al., Predicting gestational diabetes: choosing the optimal early serum marker, Am. J. Obstet. Gynecol. 196 (4) (2007) 410.e1–410.e6. ; discussion 410.e6-7.
[18] N.S. Artzi, S. Shilo, E. Hadar, et al., Prediction of gestational diabetes based on nationwide electronic health records, Nat. Med. 26 (1) (2020) 71–76.
[19] M.A. Kennelly, F.M. McAuliffe, Prediction and prevention of Gestational Diabetes: an update of recent literature, Eur. J. Obstet. Gynecol. Reprod. Biol. 202 (2016) 92–98.
[20] B.M. Snyder, R.J. Baer, S.P. Oltman, et al., Early pregnancy prediction of gestational diabetes mellitus risk using prenatal screening biomarkers in nulliparous women, Diabetes Res. Clin. Pract. 163 (2020) 108139.
[21] J. Ogonowski, T. Miazgowski, M. Kuczyńska, et al., Pregravid body mass index as a predictor of gestational diabetes mellitus, Diabet. Med. 26 (4) (2009) 334–338.
[22] A.P. Nombo, A.W. Mwanri, E.M. Brouwer-Brolsma, et al., Gestational diabetes mellitus risk score: a practical tool to predict gestational diabetes mellitus risk in Tanzania, Diabetes Res. Clin. Pract. 145 (2018) 130–137.
[23] M. van Leeuwen, B.C. Opmeer, E.J.K. Zweers, et al., Estimating the risk of gestational diabetes mellitus: a clinical prediction model based on patient characteristics and medical history, BJOG 117 (1) (2010) 69–75.
[24] J. Wang, B. Lv, X. Chen, et al., An early model to predict the risk of gestational diabetes mellitus in the absence of blood examination indexes: application in primary health care centres, BMC Pregnancy Childbirth 21 (1) (2021) 814.
[25] X. Wang, X. Zheng, J. Yan, et al., The clinical values of Afamin, triglyceride and PLR in predicting risk of gestational diabetes during early pregnancy, Front. Endocrinol. 12 (2021) 723650.
[26] Y. Wu, S. Ma, Y. Wang, et al., A risk prediction model of gestational diabetes mellitus before 16 gestational weeks in Chinese pregnant women, Diabetes Res. Clin. Pract. 179 (2021) 109001.
[27] J.D. Conklin, Applied logistic regression, Technometrics 44 (1) (2002) 81–82.
[28] J.R. Quinlan, Mach. Learn. 1 (1) (1986) 81–106.
[29] L. Breiman, Mach. Learn. 45 (1) (2001) 5–32.
[30] M. Pontil, A. Verri, Properties of support vector machines, Neural Comput. 10 (4) (1998) 955–974.

[31] T. Chen, C. Guestrin, XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16: the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 13 08 2016 17 08, New York, NY, USA: ACM, San Francisco California USA, 2016 8132016, 785–794.
[32] H. Liu, J. Li, J. Leng, et al., Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China, Diabetes Metab Res Rev 37 (5) (2021) e3397.
[33] J.-N. Tong, L.-L. Wu, Y.-X. Chen, et al., Fasting plasma glucose in the first trimester is related to gestational diabetes mellitus and adverse pregnancy outcomes, Endocrine 75 (1) (2022) 70–81.
[34] Y.-T. Wu, C.-J. Zhang, B.W. Mol, et al., Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning, J. Clin. Endocrinol. Metab. 106 (3) (2021) e1191–e1205.
[35] Yi Wang, Yichao Huang, Ping Wu, et al. Plasma Lipidomics in Early Pregnancy and Risk of Gestational Diabetes Mellitus: a Prospective Nested Case–Control Study in Chinese Women.
[36] B. Klop, J.W.F. Elte, M.C. Cabezas, Dyslipidemia in obesity: mechanisms and potential targets, Nutrients 5 (4) (2013) 1218–1240.
[37] Y. Wang, Z. Ge, L. Chen, et al., Risk prediction model of gestational diabetes mellitus in a Chinese population based on a risk scoring system, Diabetes Ther 12 (6) (2021) 1721–1734.
[38] A. Kautzky-Willer, R. Prager, W. Waldhausl, et al., Pronounced insulin resistance and inadequate beta-cell secretion characterize lean gestational diabetes during and after pregnancy, Diabetes Care 20 (11) (1997) 1717–1723.
[39] L. Bozkurt, C.S. Göbl, K. Leitner, et al., HbA1c during early pregnancy reflects beta-cell dysfunction in women developing GDM, BMJ Open Diabetes Res Care 8 (2) (2020).
[40] P.B. Renz, G. Cavagnolli, L.S. Weinert, et al., HbA1c test as a tool in the diagnosis of gestational diabetes mellitus, PLoS One 10 (8) (2015) e0135989.
[41] M. Valadan, Z. Bahramnezhad, F. Golshahi, et al., The role of first-trimester HbA1c in the early detection of gestational diabetes, BMC Pregnancy Childbirth 22 (1) (2022) 71.
[42] S. Yang, F.-T. Shi, P.C.K. Leung, et al., Low thyroid hormone in early pregnancy is associated with an increased risk of gestational diabetes mellitus, J. Clin. Endocrinol. Metab. 101 (11) (2016) 4237–4243.
[43] J. Luo, X. Wang, L. Yuan, et al., Association of thyroid disorders with gestational diabetes mellitus: a meta-analysis, Endocrine 73 (3) (2021) 550–560.
[44] L. Durrani, S. Ejaz, L.B. Tavares, et al., Correlation between high serum ferritin level and gestational diabetes: a systematic review, Cureus 13 (10) (2021) e18990.
[45] Y. Xiong, L. Lin, Y. Chen, et al., Prediction of gestational diabetes mellitus in the first 19 weeks of pregnancy using machine learning techniques, J. Matern. Fetal Neonatal Med. 35 (13) (2022) 2457–2463.
[46] P.C. Tan, A.Z. Aziz, I.S. Ismail, et al., Gamma-glutamyltransferase, alanine transaminase and aspartate transaminase levels and the diagnosis of gestational diabetes mellitus, Clin. Biochem. 45 (15) (2012) 1192–1196.
[47] J. Zhang, N. Cheng, Y. Ma, et al., Liver enzymes, fatty liver and type 2 diabetes mellitus in a jinchang cohort: a prospective study in adults, Can. J. Diabetes 42 (6) (2018) 652–658.
[48] L. Zhao, W. Li, F. Ping, et al., Associations of white blood cell count,alanine aminotransferase,and aspartate aminotransferase in the first trimester withgestational diabetes mellitus, Zhongguo Yi Xue Ke Xue Yuan Xue Bao 38 (3) (2016) 283–287.
[49] P. Feng, G. Wang, Q. Yu, et al., First-trimester blood urea nitrogen and risk of gestational diabetes mellitus, J. Cell Mol. Med. 24 (4) (2020) 2416–2422.
[50] T. Sun, F. Meng, H. Zhao, et al., Elevated first-trimester neutrophil count is closely associated with the development of maternal gestational diabetes mellitus and adverse pregnancy outcomes, Diabetes 69 (7) (2020) 1401–1410.
[51] S. Gorar, G.B. Abanonu, A. Uysal, et al., Comparison of thyroid function tests and blood count in pregnant women with versus without gestational diabetes mellitus, J. Obstet. Gynaecol. Res. 43 (5) (2017) 848–854.