Journal of Pharmaceutical and Medical Research
Print ISSN: 2663-1954
Online ISSN: 2663-1962
DOI: https://doi.org/10.61784/jpmr3088

TIMING SELECTION OF NIPT BASED ON A GENERALIZED LINEAR MIXED MODEL WITH A BINARY OUTCOME VARIABLE

YuanBo Dai[1], QingHua Liu[2*], Yu Qiu[3], XinXi Zhou[4]
[1]*School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China.*
[2]*School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China.*
[3]*School of Computer Science and Information Security, Guilin University of Electronic, Technology, Guilin 541004, Guangxi, China.*
[4]*School of Business, Guilin University of Electronic, Technology, Guilin 541004, Guangxi, China.*
[*]*Corresponding Author: QingHua Liu*

Abstract: Non-invasive prenatal testing (NIPT) mitigates the uncertainties associated with childbirth and enhances neonatal quality, thereby facilitating policy implementation. Stratification by body mass index (BMI) can enhance the accuracy and safety of these tests. The study initially processed data and employed models to ascertain that maternal age, gestational age, BMI, and parity influence Y chromosome concentration. Subsequently, the K-means++ algorithm categorized pregnant women with male fetuses based on their BMI. Through survival analysis, the optimal NIPT timing for each group was predicted to be 12.9, 13.3, 13.9, 13.9, and 16.1 weeks. Monte Carlo simulations indicate that elevated testing errors disproportionately affect women with higher BMI. A novel model evaluates whether the Y chromosome meets established standards and assesses testing errors, yielding improved NIPT timings of 16.0, 17.3, 19.2, 21.1, and 12.0 weeks, thereby illustrating the impact of errors on timing. To detect abnormalities in female fetuses, a model utilizing random forest and light gradient boosting machine was developed. It achieved an area under the curve (AUC) of 0.9903 and an accuracy of 0.9889, delivering efficient and precise outcomes.
Keywords: K-means++ clustering; Two result variables; Combined stacking method; Generalized linear mixed model

1 INTRODUCTION

Non-invasive prenatal testing (NIPT) is a pivotal biotechnological innovation in prenatal care. To further optimize NIPT performance—specifically to improve its detection efficiency and safety—this study focuses on addressing core technical challenges through model construction and algorithm application. Key technical efforts include: constructing relational models to quantify correlations between fetal Y chromosome concentration (a critical factor for NIPT accuracy) and maternal physiological indicators; applying K-means++ clustering to stratify pregnant women with male fetuses based on body mass index (BMI)[1-2]; integrating survival analysis and Monte Carlo simulation to predict optimal testing timing and assess the impact of measurement errors; developing a bivariate Generalized Linear Mixed Model (GLMM) to optimize NIPT timing by considering dual outcome variables ("concentration meeting standards" and "error-free detection"); and establishing a stacking ensemble learning framework (combining Random Forest, LightGBM, and Logistic Regression) to achieve high-precision detection of chromosomal abnormalities in female fetuses. These model-driven and algorithm-based approaches form the core technical path for advancing NIPT optimization in this study.

In NIPT, the concentration of the Y chromosome in male fetuses is a core factor determining the testing accuracy. Existing studies indicate that Y chromosome concentration correlates with maternal physiological features, such as gestational age and body mass index (BMI). This suggests that refining groupingings of pregnant women based on individual characteristics (particularly BMI) and recommending personalized optimal testing windows is an effective strategy to improve NIPT efficiency and reduce potential risks from testing errors for expectant mothers. This study utilized NIPT result data from pregnant women in a specific region, with a sample notable for high BMI, providing unique data support for an in-depth exploration of BMI's impact of BMI on NIPT accuracy.

This study aimed to address the following core scientific questions: First, to construct relational models and quantitatively analyze the correlations and significance between fetal Y chromosome concentration and various maternal physiological indicators, such as maternal age, gestational age, BMI, and parity. Second, we aimed to scientifically group pregnant women carrying male fetuses based on BMI, determine the optimal NIPT testing time for each group to minimize potential risks, and evaluate the quantitative impact of testing errors on the results. Next, building upon the above research, more complex multidimensional factors (such as height, weight, etc.) and detection errors should be incorporated to construct a more sophisticated model for identifying more reasonable groupings and optimal testing points. Finally, for female fetuses lacking the Y chromosome as a direct reference indicator, we explored an intelligent classification method that integrates multi-source data (such as X chromosome Z-value, GC content, sequencing read count, and maternal individual factors) to accurately identify aneuploidy abnormalities on chromosomes 21, 18, and 13.

2 METHODS

2.1 Data Preprocessing and Descriptive Statistics

First, the raw data were cleaned to remove outliers and missing values. Next, descriptive statistical analyses were conducted on key variables such as age, BMI, gestational weeks, and Y chromosome concentration, and distribution histograms were plotted to preliminarily observe the overall distribution patterns and characteristics of the data.

2.2 Correlation Analysis Model for Y Chromosome Concentration

Linear Mixed-Effects Model (LMM)[3]: Considering that the data have longitudinal characteristics (the same pregnant woman may have multiple test records), meaning that data within a group (the same woman) are correlated while data between groups (different women) are independent, we used a linear mixed-effects model to analyze the relationship between Y chromosome concentration and multiple covariates. This model can simultaneously estimate fixed effects (the average trend across all samples) and random effects (differences among individuals). For the i-th cluster, the model is expressed as follows:

$$Y_i=X_i\beta+Z_ib_i+\epsilon_i \quad i=1,\cdots,m \tag{1}$$

Where, Y_i is the dependent variable vector (Y chromosome concentration), X_i and Z_i are the covariate matrices, β is the fixed effect coefficient vector, bi is the random effect coefficient vector, $b_i \sim N(0, G)$, ϵi is the random error vector, and $\epsilon_i \sim N(0,\sigma^2 I_{n_i})$. Model parameters are estimated via a two-step iterative procedure.

Locally Weighted Regression: Since preliminary scatter plot analysis in Figure 1 indicated potential non-linear relationships among variables, we incorporated locally weighted regression to test the significance of the LMM fitting results. LOESS is a non-parametric regression method that fits weighted polynomials to local subsets of the data, allowing for flexible detection of local trends in the dataset and thereby validating the goodness-of-fit of the model.

2.3 Optimal NIPT Timing Selection Based on BMI Grouping

K-means++ Clustering[4]: The K-means++ clustering algorithm was employed to categorize pregnant women carrying male fetuses according to their BMI. This algorithm enhances the stability and accuracy of clustering by optimizing the initial cluster center selection strategy, ensuring that the centers are maximally distant from each other. In this study, the number of clusters was set to K=5.

Survival Analysis[5-6]: To ascertain the optimal timing for NIPT within each BMI group, a survival analysis model was introduced. The event of interest was defined as "the first instance when the fetal Y chromosome concentration reaches or exceeds 4%," and the survival time T represented the gestational weeks required to reach this event. The optimal testing window for each group can be predicted by calculating the survival function $S(t)=P(T>t)$, which denotes the probability that the concentration has not yet reached the threshold at time t, and the hazard function $\lambda(t)$, which represents the instantaneous probability of reaching the threshold per unit time after time t. The curve of the probability of "not meeting the standard" as gestational age progresses in Figure 2 intuitively reflects the differences among various BMI groups.

Monte Carlo Simulation[7]: To evaluate the impact of measurement errors on the results, the Monte Carlo simulation method was utilized. Various levels of relative error (0%, 1%, 5%, 10%, and 15%) were applied to the actual Y chromosome concentration values to generate simulated detection values. By comparing these simulated values with the threshold (4%), the misclassification rates under different error levels were calculated, thereby assessing the impact of errors on the different BMI groups. The results are presented in the form of a heat map in Figure 2, which clearly shows the correlation between error rate, error level, and BMI level.

2.4 Optimization Model for NIPT Timing Considering Dual-Outcome Variables

$$S(t)=P(T>t)=\int_t^{\infty} f(u)du=1-F(t) \tag{2}$$

$$h(t)=\frac{f(t)}{S(t)}=\lim_{\Delta t\to 0}\frac{P(t<T<t+\Delta t)}{P(T>t)\Delta t} \tag{3}$$

$$g(\pi_i)=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\eta=X\beta+Zu \tag{4}$$

$$\begin{cases} R=f_{gw}\cdot f_{bmi}\cdot \begin{pmatrix} 0.5|Z_{21}|+0.25|Z_X|+0.15|Z_{18}| \\ +0.1|Z_{13}| \end{pmatrix} \\ f_{gw}\in \begin{cases} 1.3, & X_2<12 \\ 1.05, & 12\le X_2<16 \\ 1.0, & 16\le X_2 \end{cases} \\ f_{bmi}\in \begin{cases} 1.25, & 24<X_3 \\ 1.1, & 22\le X_3\le 24 \\ 1.0, & X_3\in \text{others} \end{cases} \end{cases} \tag{5}$$

In continuation of the analysis presented, we developed a bivariate Generalized Linear Mixed Model (GLMM) to optimize the timing of noninvasive prenatal testing (NIPT).

2.5 Maintaining the Integrity of the Specifications

Stacking Ensemble Learning Framework[8-9]: To address the challenge posed by the absence of a Y chromosome reference in female fetuses, we propose a method utilizing the stacking ensemble learning framework. This methodology combines the strengths of multiple base learners to enhance predictive performance.
Base Learners: Random Forest (RF) and LightGBM have been selected as the base learners for the initial layer[10-11]. RF is noted for its robust nonlinear modeling capabilities and resistance to noise, whereas LightGBM is recognized for its high efficiency and accuracy in processing large-scale datasets.
Meta-Learner: The cross-validated predictions from the two base learners on the training set are employed as new features (meta-features), which are subsequently input into a Logistic Regression model for second-layer training. As the meta-learner, Logistic Regression is tasked with optimally combining the outputs of the base learners.
Label Transformation and Risk Prediction: Initially, a prior risk scoring formula was established based on Z-score, gestational week, and BMI. Subsequently, clinical labels from the raw data were transformed into binary 0/1 labels. Ultimately, the trained two-layer stacking model produced a probability value between 0 and 1, indicating the risk of chromosomal abnormalities in female fetuses.
Model Evaluation: The efficacy of the female fetal abnormality model was thoroughly assessed using five metrics: Area Under the Receiver Operating Characteristic Curve (AUC), Accuracy, Precision, Recall, and F1-Score. Feature importance analysis in Figure 3 shows key prediction features, and the ROC curve in Figure 4 intuitively reflects their performance comparison.

3 RESULTS AND ANALYSIS

3.1 Basic Data Characteristics and Correlation Analysis

The descriptive statistics histogram shows that pregnant women's age and BMI are approximately normally distributed, with most samples concentrated in the intermediate range in Figure 1. The concentration of Y chromosome exhibits a relatively stable distribution between 0.04 and 0.09. The distribution of gestational weeks displays a multimodal pattern, reflecting the diversity in the selection of gestational weeks for NIPT testing.
The scatter plot of Y chromosome concentration, BMI and gestational age in Figure 2 reveals their potential non-linear relationships, laying the foundation for using locally weighted regression to verify model fitting.
Using a linear mixed-effects model analysis, the relationship model between Y chromosome concentration and each indicator was derived as follows:

$$Y=-0.01338X_1+0.398879X_2-0.14718X_3 \quad (2)$$

The model parameter results (see Table 1) show that maternal age (C), gestational age (J), BMI (K), and number of pregnancies (AC) all have a statistically significant impact on Y chromosome concentration ($P<0.05$). Among these, only gestational age has a positive effect—that is, the higher the gestational age, the higher the Y chromosome concentration; while age, BMI, and number of pregnancies all show a negative effect.
To validate the model's fit, we plotted a graph comparing the actual and model-predicted Y chromosome concentrations. in Figure 3 The trend line fitted using locally weighted regression (LOWESS) is quite close to the ideal fit line, and the model's coefficient of determination R^2 is 0.69951, indicating that the model explains the variability in the data well overall, and there is a strong correlation between the actual and predicted values. This suggests that although the linear relationship between any single variable and Y chromosome concentration is not obvious, the combined effect of multiple variables can be effectively captured by the LMM.

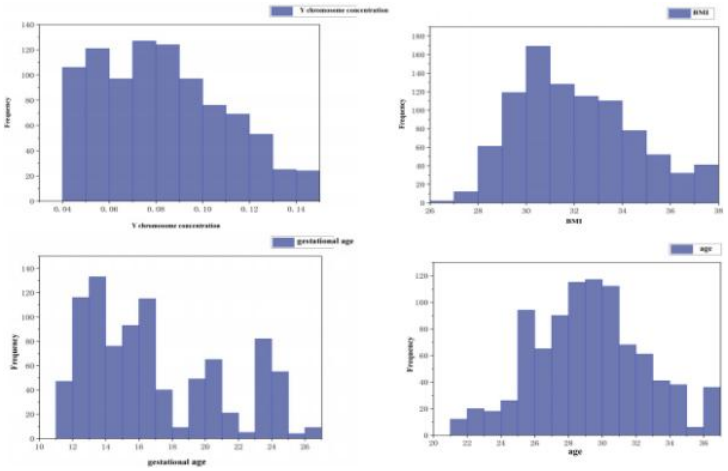

Figure 1 Histogram of the Distribution of Certain NIPT Test Indicators

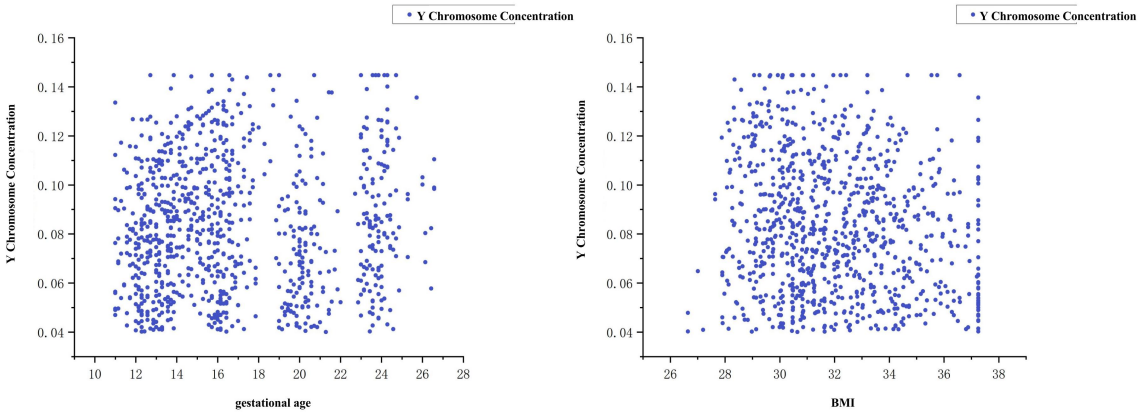

Figure 2 Scatter Plot of Y Chromosome Concentration vs. BMI and Gestational Age

Table 1 Results of Linear Model Parameters

Variables	Coefficient	t-statistic	P-value
β	-0.05742	-1.00115	0.316756
J	0.398879	15.98413	0.000001
K	-0.14718	-2.8588	0.004252
C	-0.11345	-1.96011	0.049983
AA	-0.01179	-0.56633	0.571173
AC	-0.16421	-2.82904	0.004669
AD	-0.01338	-0.21707	0.828156

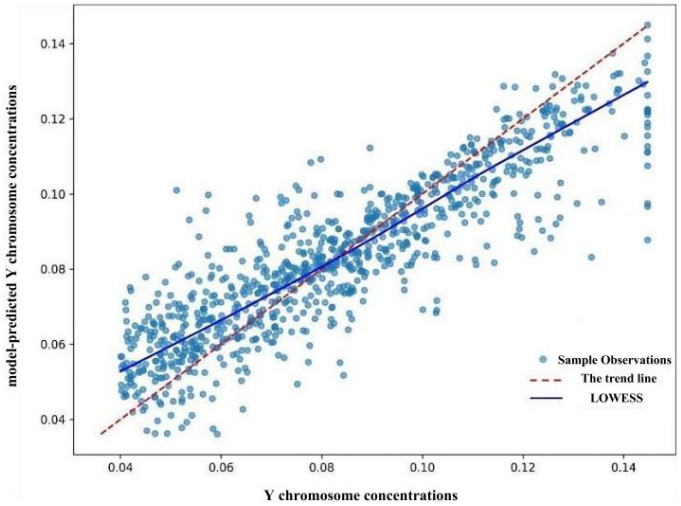

Figure 3 Graph of Actual vs. Simulated Y Chromosome Concentration and Fitting Trend

3.2 Preliminary Determination of NIPT Timing Based on Survival Analysis and Error Analysis

After dividing pregnant women carrying male fetuses into five categories according to BMI using the K-means++ algorithm, the optimal NIPT timing for each category, as predicted by the survival analysis model, is shown in Table 2.

Table 2 Predicted Optimal NIPT Timing for Pregnant Women Carrying Male Fetuses

Cluster	Predicted Optimal NIPT Timing
Cluster1	12.9
Cluster2	13.3
Cluster3	13.9
Cluster4	13.9
Cluster5	16.1

The curve showing the probability of "not meeting the standard" as gestational age progresses in Figure 4 intuitively demonstrates the differences among the various BMI groups. For all groups, the probability of not meeting the standard decreases with gestational age, but the rate and timing of this decrease vary, providing a basis for individualized timing selection.

Monte Carlo simulation results for detection error are presented in the form of a heat map in Figure 5. The results clearly indicate that the error rate is positively correlated with both error levels and BMI levels. Notably, the fifth group with the

highest BMI (BMI≥35.2) consistently shows the highest error rate under all error levels, confirming that pregnant women with high BMI (obesity) are more sensitive to detection errors.

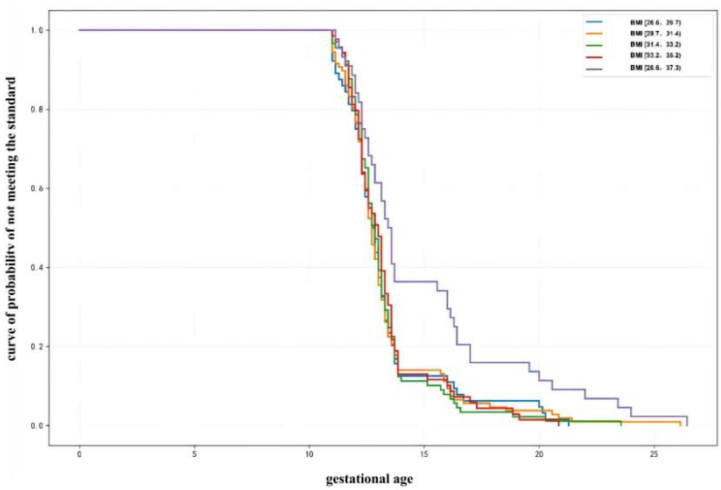

Figure 4 Curve of Probability of Not Meeting the Standard vs. Gestational Age

Monte Carlo simulation results for detection error are presented in the form of a heat map. The results clearly indicate that the error rate is positively correlated with both error levels and BMI levels. Notably, the fifth group with the highest BMI (BMI≥35.2) consistently shows the highest error rate under all error levels, confirming that pregnant women with high BMI (obesity) are more sensitive to detection errors.

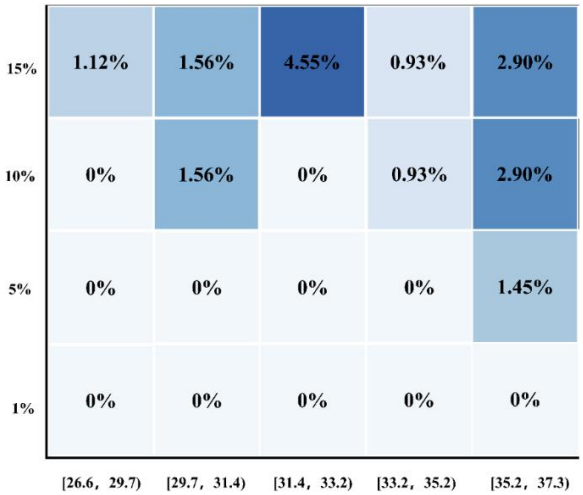

Figure 5 Heat Map of NIPT Detection Error Rates Across Different BMI Groups

3.3 NIPT Timing Optimization Based on a Dual-Outcome Model

By introducing a bivariate generalized linear mixed model that simultaneously accounts for both "concentration meeting the standard" and "error-free detection" outcomes, we obtained more robust and optimized best NIPT timings for each BMI group, as shown in Table 3.

Table 3 Results for Best NIPT Timings in Problem Three

Cluster	Predicted Optimal NIPT Timing
Cluster1	16.0
Cluster2	17.3
Cluster3	19.2
Cluster4	21.1
Cluster5	12.0

Further analysis of the impact of detection error on the selection of optimized timings revealed that: in the low error rate (0-5%) range, there are significant differences in the best timing among groups, highlighting the necessity of personalized strategies. As the error rate increases, the optimal timings for each group gradually converge (toward 12 weeks), indicating that when error becomes the dominant factor, the flexibility brought by grouping is weakened and the detection strategy is forced to become unified. This trend is clearly reflected in the line chart of best timing for different BMI groups versus error rate in Figure 6

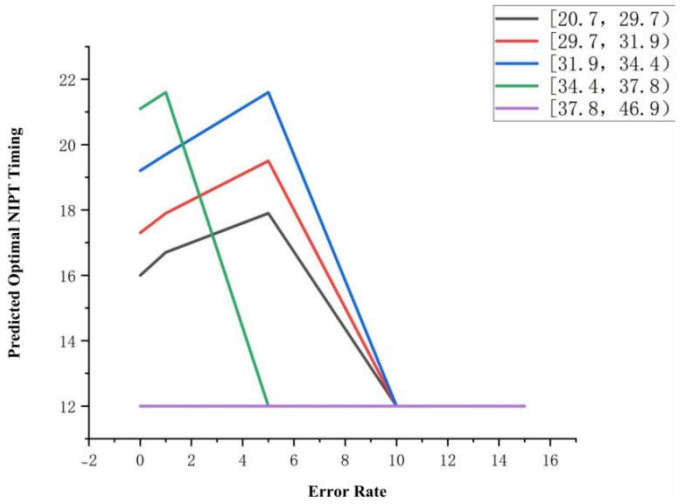

Figure 6 Line Chart of Best Timing for Different BMI Groups versus Error Rate

3.4 Ensemble Learning-Based Classification of Chromosomal Abnormalities in Female Fetuses

This study constructed a two-layer stacked model for the classification of chromosomal abnormalities in female fetuses. Feature importance analysis in Figure 7 showed that the Z-scores for chromosomes 18, 21, and 13 were the top three most important features for predicting aneuploidy, whereas indicators such as maternal BMI and GC content were relatively less important.

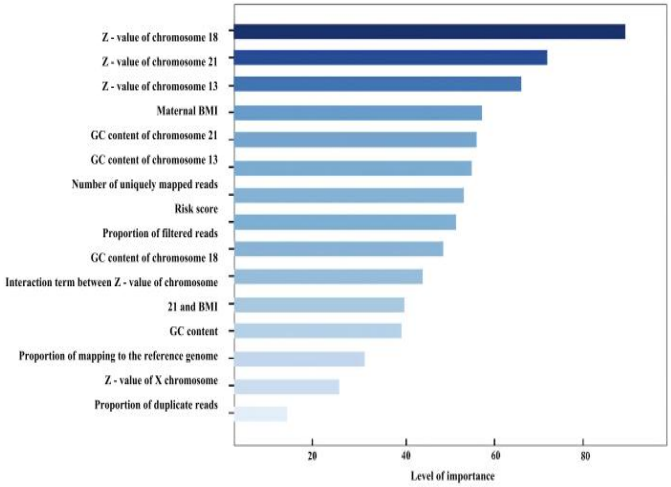

Figure 7 Feature Importance Ranking for Female Fetus Abnormality Prediction

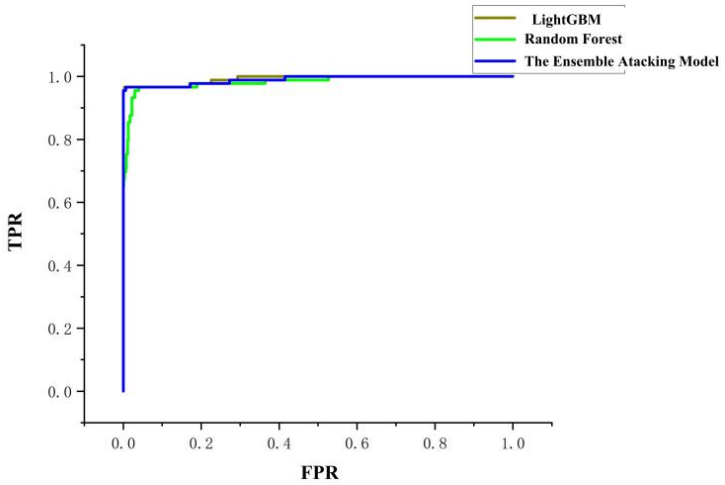

Figure 8 ROC Curve

To evaluate the model performance, we plotted the ROC curve in Figure 8. The results show that the ensemble stacking model (blue curve) significantly outperforms the individual Random Forest (green curve) and LightGBM models, with the largest area under the curve (AUC) very close to 1.

The quantitative evaluation metrics for the model are shown in Table 4, all of which demonstrate outstanding performance. The AUC reached 0.9903, with an accuracy of 0.9889, indicating that the model has extremely high discriminative power and accuracy. Both precision (0.9451) and recall (0.9663) were also at high levels, indicating that the model accurately identified abnormal samples while maintaining a very low miss rate. Finally, the model set optimal decision thresholds based on BMI stratification (0.4300 when BMI≤24, and 0.4800 when BMI>24), further enhancing its adaptability and robustness in practical applications.

Table 4 Problem 4 Model Solution Results

AUC	Accuracy	Precision	Recall	F1-Score
0.9903	0.9889	0.9451	0.9663	0.9556

4 CONCLUSION AND DISCUSSION

This study systematically investigates strategies for optimizing the timing of non-invasive prenatal testing (NIPT) based on individual maternal characteristics, particularly body mass index (BMI), and presents an efficient intelligent identification method for detecting fetal chromosomal aneuploidies in female fetuses. Gestational week, BMI, age, and number of pregnancies are significant factors influencing fetal Y chromosome concentration. Among these, gestational week is positively correlated with concentration and serves as the most critical physiological basis for determining the timing of NIPT testing. By integrating K-means++ clustering with survival analysis and a generalized linear mixed model for dual-outcome variables, we successfully identified personalized optimal NIPT testing times for pregnant women carrying male fetuses with varying BMI levels (at 16.0, 17.3, 19.2, 21.1, and 12.0 weeks, respectively). This approach minimizes risk while ensuring testing accuracy. Measurement error significantly impacts NIPT results, with pregnant women of higher BMI being more sensitive to such errors. As error rates increase, the benefits of personalized testing times gradually diminish.

Model evaluation and outlook: The models employed in this study offer multiple advantages. The linear mixed model effectively handles correlations in longitudinal data; K-means++ and survival analysis provide data-driven justifications for timing selection; the dual-outcome variable model optimizes decision-making from multiple dimensions; while the stacking ensemble model significantly enhances the robustness and accuracy of classification tasks by integrating different model strengths. However, there are limitations, such as the linear mixed model's sensitivity to outliers, necessitating comprehensive data preprocessing.

The analytical framework and predictive models established in this study offer strong potential for broader application. The framework can optimize various NIPT testing scenarios by adjusting input features and parameters to specific situations, and its core algorithmic concepts, such as multi-factor modeling and ensemble learning, can be applied to other areas of medical testing and health assessment, such as non-invasive glucose monitoring and fetal malformation risk assessment, thereby providing robust technical support and practical reference for precision medicine.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Budis J, Gazdarica J, Radvanszky J, et al. Non-invasive prenatal testing as a valuable source of population specific allelic frequencies. Journal of Biotechnology, 2019, 299: 26.

[2] Liu L L, Zheng M L, Zou B J, et al. Highly Sensitive and Specific Detection of SRY Gene for Non–invasive Prenatal Diagnosis. Chinese Journal of Analytical Chemistry, 2017, 45(10): 1043.

[3] Dasgupta S, Roy P, Ghosh C, et al. A prospective cohort study of adverse fetomaternal outcomes among overweight, obese, and normal weight pregnant women with term singleton pregnancy in a tertiary care hospital. Asian Journal of Medical Sciences, 2022, 13(4): 42351.

[4] Wang X H. The analysis of sculpture image classification in utilization of 3D reconstruction under K-means++. Scientific Reports, 2025, 15(1): 1949.

[5] Leivayamaguchi V, Alvares D. A Two-Stage Approach for Bayesian Joint Models of Longitudinal and Survival Data: Correcting Bias with Informative Prior. Entropy, 2020, 23(1): 50.

[6] Tang A M, Tang N S, Yu D L. Bayesian semiparametric joint model of multivariate longitudinal and survival data with dependent censoring. Lifetime Data Analysis, 2023, 29(4): 608.

[7] Abushattal M, Alhomaidat F, El Yabroudi M. Effects of intersection control types on driver yielding behavior to cyclists using mixed logit modeling. Scientific Reports, 2025, 15(1): 9801.

[8] Zhang J H, Xu Z J, Xiong Z, et al. Android malware detection based on feature fusion and the improved stacking ensemble model. Journal of Computer Virology and Hacking Techniques, 2025, 21(1): 546.

[9] Wang J J, Jiang L, Zhang L, et al. An Integrated Stacking Ensemble Model for Natural Gas Purchase Prediction Incorporating Multiple Features. Applied Sciences, 2025, 15(2): 778.

[10] Diamantopoulou M J, Özçelik R, Kalkanli Genç Ş. Evaluation of the random forest regression machine learning technique as an alternative to ecoregional based regression taper modelling. Computers and Electronics in Agriculture, 2025, 239(PA): 110964.
[11] Chen S Y, Shao W Q, Gao S P, et al. High impedance fault feeder selection method based on MPA-VMD and LightGBM in distribution networks. International Journal of Electrical Power and Energy Systems, 2025, 171: 110985.