Research on the Optimization of Pregnant Women's Characteristics and NIPT Detection Strategy Based on Data Driven Modeling

Yangsen Li *

School of Digital and Intelligent Industry (School of Cyber Science and Technology), Inner Mongolia University of Science & Technology, Inner Mongolia, China, 10127

* Corresponding Author Email: lee_123456@foxmail.com

Abstract. This article focuses on key issues in non-invasive prenatal testing and establishes multiple regression/GAM, BMI clustering risk assessment, machine learning regression prediction, and classification models. By conducting relevant analysis to evaluate the relationship between gestational age, BMI, and Y chromosome concentration in pregnant women, multiple linear regression, second-order polynomial regression, and generalized additive model (GAM) were constructed. The results showed that the GAM model outperformed other models in R^2, better capturing the nonlinear effects of gestational age and BMI on Y concentration; The significance test showed that these factors had a statistically significant impact on Y concentration ($p<0.05$). Using K-means clustering to group pregnant women according to BMI, calculate the gestational weeks when the first Y concentration reaches 4% in each group, and construct a comprehensive risk function to consider the impact of the probability of reaching the standard and the detection time point on the risk, and solve for the optimal detection time point for each group. The results showed that there were significant differences in the first gestational age of different BMI groups, and the risk model recommended optimized detection time points for each group.This study provides a systematic, data-driven approach to optimize NIPT detection strategies by integrating regression analysis, clustering, and machine learning models. By revealing the nonlinear impact of gestational age and BMI on Y chromosome concentration and proposing differentiated detection time points for pregnant women with different BMI levels, the research not only improves the scientific basis for clinical decision-making but also reduces the risk of false negatives or delayed diagnoses. The findings contribute to more precise, personalized prenatal testing strategies, thereby enhancing the reliability and effectiveness of NIPT in practical medical applications.

Keywords: Non Invasive Prenatal Testing, Risk Assessment, Generalized Additive Model, K-means Clustering.

1. Introduction

In non-invasive prenatal testing, fetal free DNA in the blood of pregnant women can be used to determine whether the fetal chromosomes are abnormal [1]. Compared with traditional amniocentesis, NIPT has low risk, early detection, and high accuracy, making it widely used. In practical operation, differences in BMI, gestational age, and other factors among different pregnant women can affect the duration of fetal Y chromosome concentration elevation, thereby affecting whether the detection meets the standard [2]. Early detection may fail due to insufficient DNA content, while late detection may miss the optimal intervention time [3].

Therefore, it is necessary to conduct a reasonable analysis of individual differences and testing time points among pregnant women, in order to optimize the testing arrangement and abnormal risk assessment of NIPT [4]. From the characteristics of pregnant women, BMI Factors such as gestational age and gestational age can affect the timing of achieving fetal Y chromosome concentration [5]. Early detection may fail due to insufficient DNA content, while late detection may miss the optimal intervention time; From the perspective of fetal gender, the determination of female fetal abnormalities requires a comprehensive analysis based on multiple indicators of the X chromosome and chromosomes 13/18/21. By combining the characteristics of pregnant women and fetal gender,

the optimal testing timing and grouping strategy can be analyzed and determined, thereby improving testing accuracy and reducing potential risks.

This study systematically analyzed NIPT data using statistical regression, cluster analysis, and machine learning methods. The introduction of GAM enhances the ability to characterize variable relationships and provides reference for personalized detection time points through BMI grouping and risk models. The proposed method balances predictive performance and practicality, providing a quantitative basis for clinical decision-making. However, the model relies on existing data for construction, and some parameter settings still require more clinical validation. In the future, model stability and generalization value can be improved by increasing samples and features, optimizing algorithms.

In recent years, a number of studies have focused on improving the reliability of NIPT by examining maternal characteristics and technical parameters. However, most existing studies either focus on linear regression analysis of maternal features or employ black-box machine learning models, which lack interpretability and often fail to provide actionable guidance on optimal detection timing.

Compared with these studies, this paper makes the following innovations: (1) it introduces the Generalized Additive Model (GAM) to flexibly capture the nonlinear effects of gestational age and BMI on fetal Y chromosome concentration while retaining interpretability; (2) it proposes a BMI-based clustering risk assessment model to identify group-specific optimal detection time points, which addresses individual heterogeneity more effectively; and (3) it integrates statistical regression, clustering, and machine learning prediction methods to provide a comprehensive, data-driven framework that balances accuracy, interpretability, and clinical applicability. These contributions highlight the novelty of this research and its potential to support personalized prenatal testing strategies.

2. Study on the concentration of fetal Y chromosome

2.1. Pearson correlation

In order to explore the relationship between fetal Y chromosome concentration and various indicators of pregnant women, this study used Pearson correlation coefficient to measure the linear correlation between variables and visualized it through a heatmap [6] [7]. The analysis in this section starts from the perspective of detecting key indicators such as gestational age and maternal BMI, and calculates their correlation coefficients with Y chromosome concentration.

The Pearson correlation coefficient can be used to measure the degree of correlation between two variables, and its calculation formula is:

$$r(x,y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad (1)$$

Among them, X_i and Y_i respectively represent a certain indicator value (such as gestational age, BMI) and the corresponding Y chromosome concentration value of the pregnant sample in the i-th observation, while X and Y represent the sample mean of the indicator sequence and the Y chromosome concentration sequence. By calculating r_{XY} pairwise for all variables, a complete correlation matrix can be obtained. Using Python language code, the following results are obtained as shown in Table 1:

Table 1. Correlation Matrix

Index	Chromosome concentration	Detecting gestational age	Pregnant women's BMI	age	height	weight
Y chromosome concentration	1	0.0948	-0.1586	-0.1001	-0.099	-0.1853
Detecting gestational age	0.0948	1	0.1505	-0.0175	0.0176	0.1272
Pregnant women's BMI	-0.1586	0.1505	1	-0.0184	0.0656	0.7582
age	-0.1001	-0.0175	-0.0184	1	0.0441	0.0064
height	-0.099	0.0176	0.0656	0.0441	1	0.6178
weight	-0.1853	0.1272	0.7582	0.0064	0.6178	1

Visualize the correlation coefficient matrix in the form of a heatmap: Figure 1 shows that the correlation between Y concentration and gestational age is positive (red), and negative (blue) with BMI and body weight, confirming that BMI is the main negative influencing factor. The correlation between multiple variables such as body weight and BMI is 0.76, indicating potential collinearity.

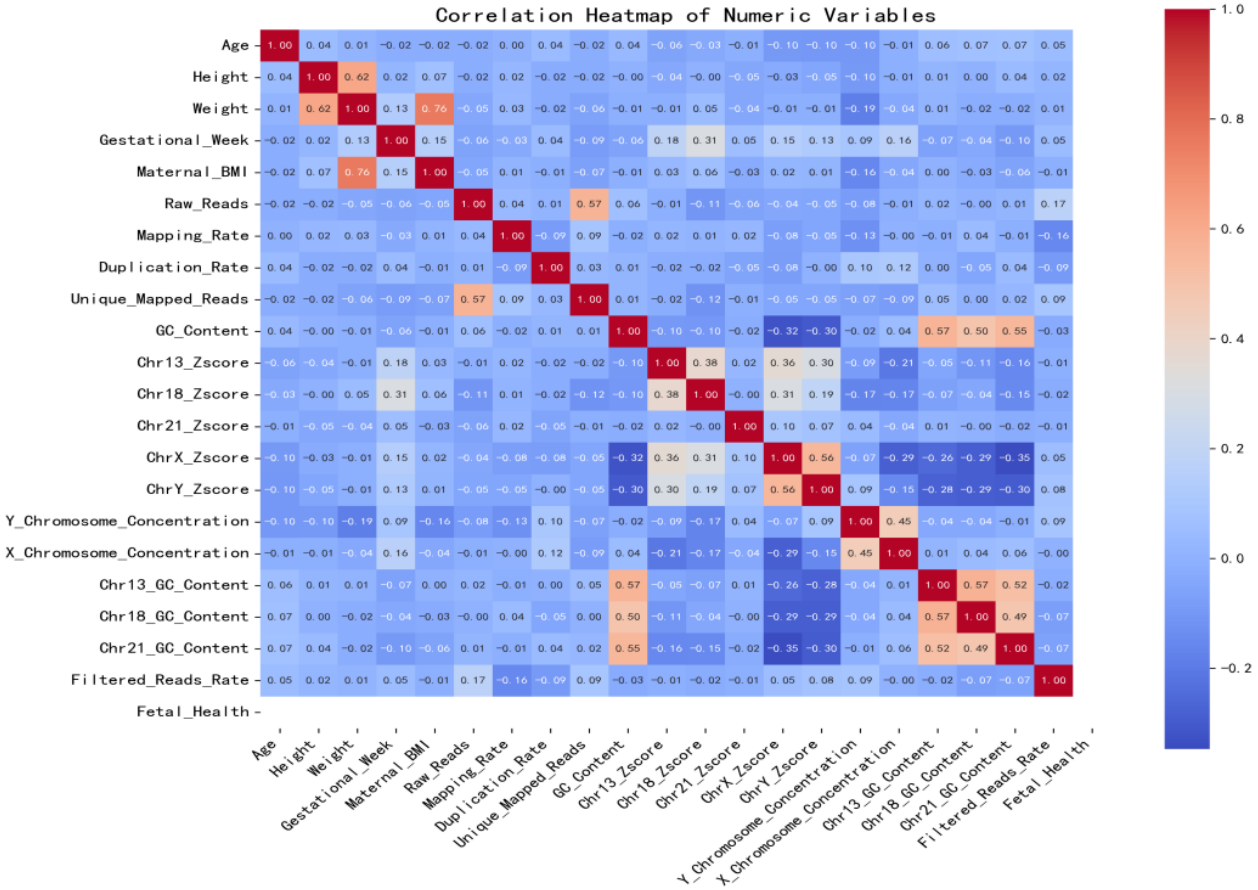
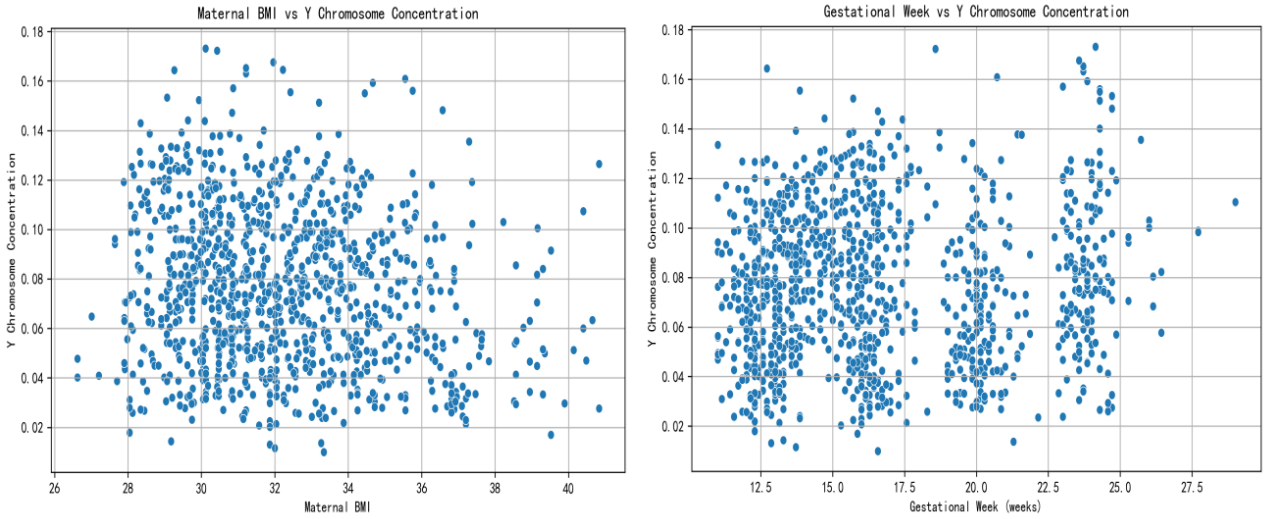


Figure 1. Heat map of numerical variable correlation after data cleaning



(a) (b)

Figure 2. Scatter plot of the relationship between BMI and gestational age and Y chromosome concentration

2.2. Model comparison

Based on Figures 1 and 2, this article established several regression models to analyze influencing factors: multiple linear regression, second-order polynomial regression, and generalized additive model. Multiple linear regression: Y represents Y chromosome concentration, B represents BMI, G represents gestational age, and other variables are denoted as $X_1, X_2, \ldots, X_p$. The model is as follows:

$$Y = \beta_0 + \beta_1 B + \beta_2 G + \sum_{j=1}^{p} \beta_{2+j} X_j + \varepsilon \quad (2)$$

β_0 is the intercept term; β_i is the regression coefficient of each variable; ε is the error term, assuming $\varepsilon \sim N(0, \sigma^2)$.

Second order polynomial regression:

$$Y = \beta_0 + \beta_1 B + \beta_2 G + \beta_3 B^2 + \beta_4 G^2 + \beta_5 (B \times G) + \sum_{j=1}^{p} \beta_{5+j} X_j + \varepsilon \quad (3)$$

Generalized Additive Model (GAM):

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_n(X_n) + \varepsilon \quad (4)$$

$f_i(\cdot)$ is a smoothing function (such as a spline function) that allows for any smooth nonlinear relationship between each variable and Y.

Python language can be used to calculate the correlation coefficients of various variables and information indicators such as AIC and GCV of GAM models. From Figure 3, it can be seen that the Generalized Additive Model (GAM) is more flexible, allowing for a smooth nonlinear relationship between each variable and Y without the need to assume a specific functional form beforehand [8]. When modeling, the main variables such as "gestational age" and "maternal BMI" are used as smoothing terms, while retaining the intercept term. By summing up the effects of each variable to fit Y, the pseudo R^2 of the model is 0.148. According to Table 2, the GAM model is more effective in capturing nonlinear trends than simple linear models.

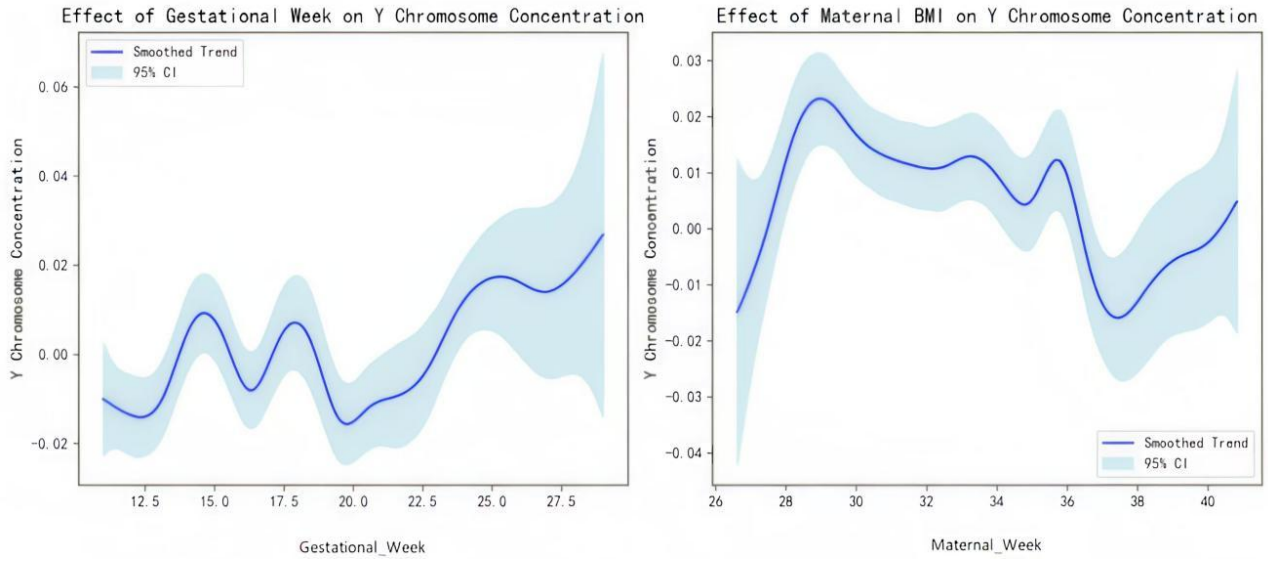


Figure 3. Smooth trend of the relationship between BMI and gestational age and Y chromosome concentration

Table 2. Comparison of Three Models

Model	R^2	$MSE(\times 10^{-3})$	$AIC(\times 10^{6})$	$GCV(\times 10^{-3})$
Multiple linear regression	0.0498	0.927	—	—
Second-order polynomial regression	0.0613	0.915	—	—
GAM model	0.1481	—	1.207	0.889

2.3. Solution of GAM Model

Firstly, the main variables are selected from the cleaned data, with the pregnant woman's "detected gestational age" and "BMI" as independent variables (X), and the fetus's "Y chromosome concentration" as the dependent variable (Y). When building the model, add a smoothing function to each independent variable to capture any possible nonlinear relationship between them and Y [9]. After the fitting is completed, the statistical indicators of GAM, including pseudo R^2, AIC, GCV, etc., can be viewed to evaluate the fitting effect and predictive ability of the model. At the same time, to ensure the reliability of the model, it is necessary to analyze the residuals, including independence and normality tests, to determine whether the model assumptions are valid and ensure the correctness of the results.

From Figure 4 and Figure 5, the residuals fluctuate up and down around the 0 axis without any obvious regular trend, and the mean of the residuals is close to zero. Most of the residuals are concentrated near the zero point, indicating that the model prediction deviation is small and the overall fitting effect is good.

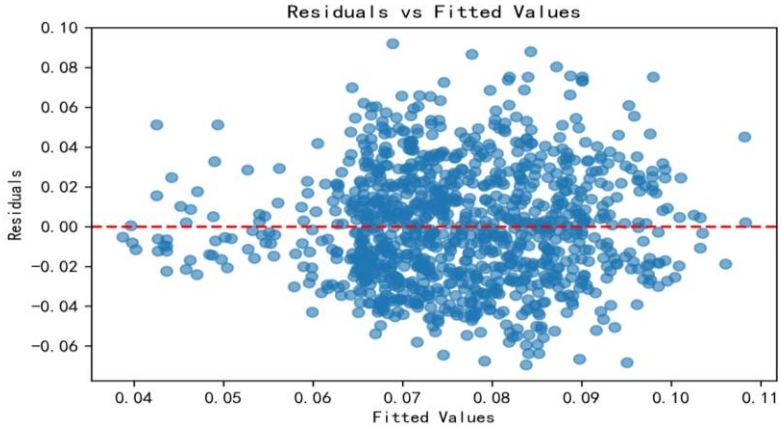
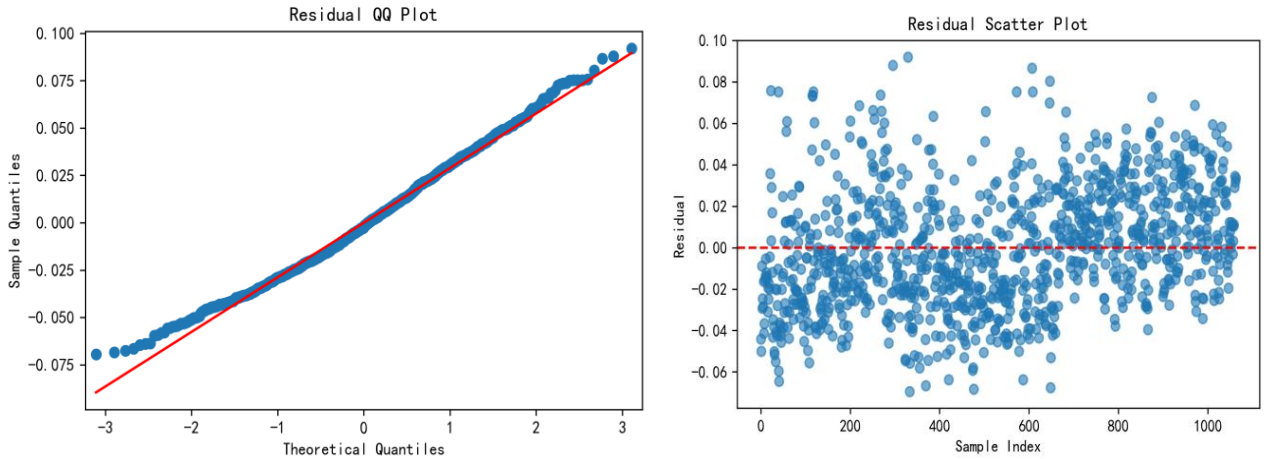


Figure 4. GAM residuals vs fitted values



(a) (b)

Figure 5. GAM residual QQ plot and scatter plot

2.4. Analysis of GAM Model Results

The joint analysis of gestational age and BMI on Y chromosome concentration showed that gestational age had a significant positive effect (concentration increased from 0.02 at 15 weeks to 0.04 at 27.5 weeks), while BMI showed a negative inhibitory effect (concentration decreased from 0.03 to -0.04 with increasing BMI); The confidence interval of the model shows a stable trend in the middle and late stages of pregnancy, but the variability increases when the BMI is high (>35). It is recommended to prioritize testing between 20-26 weeks of pregnancy in clinical practice.

3. The effect of BMI of male pregnant women on fetal Y chromosome concentration

3.1. K-Means clustering model grouping

When exploring the relationship between BMI grouping of male fetal pregnant women and the optimal timing of NIPT testing, it is necessary to consider both individual differences among pregnant women (the impact of BMI on the rate of Y chromosome concentration increase) and the uncertainty of testing errors [10]. If relying solely on traditional statistical methods such as linear regression, it is easy to overlook the multimodal characteristics of BMI distribution, which makes it difficult to group reasonably and results in a lack of representativeness. If neural networks and other "black box" models are directly used, although the prediction accuracy may be high, their interpretability is weak, making it difficult to provide actionable grouping and timing recommendations for medical testing.

Taking into account the objectives of this article, a modeling approach that balances interpretability and robustness should be adopted. This article first uses the K-Means clustering model

to perform K-Means clustering (4 clusters) on the average BMI of all pregnant women. Re label the cluster numbers in ascending order according to the center of each cluster. Subsequently, the minimum and maximum BMI values of pregnant women within each group were calculated, and four BMI grouping intervals were obtained after continuous normalization. In the analysis of gestational age, calculate the distribution of the first eligible gestational age for each group of pregnant women, and model the detection risk for different gestational ages through a segmented risk function [11]. Finally, by traversing the gestational age range, the gestational age with the lowest average risk within the group is identified as the optimal NIPT testing time for this BMI group.

Minimizing the differences in body shape characteristics among pregnant women within the same group can reflect the actual stratification effect of BMI on the successful detection of gestational age in NIPT. The model is given by the following equation:

$$\min \sum_{i=1}^{n} \sum_{k=1}^{K} 1\{z_i = k\} \|x_i - \mu_k\|^2 \tag{5}$$

Among them, x_i represents the BMI of pregnant woman i, z_i is the group label to which she belongs, μ_k is the center of the k group, and the goal is to minimize the BMI variance within the group.

Therefore, we constructed a K-Means clustering model based on the BMI data of pregnant women to achieve grouping. Firstly, by using the elbow point method to analyze the relationship between k and inertia, it was found that when k=4, the decrease in inertia slowed down significantly, indicating that the number of clusters reached a better balance between intra group differences and inter group differences. Subsequently, run the K-Means algorithm with K=4 and arrange the cluster centers in ascending numerical order to obtain continuous grouping boundaries. The final determined BMI ranges are [20.7,30.9), [31.0,33.7), [33.9,38.5), [38.9,46.9).

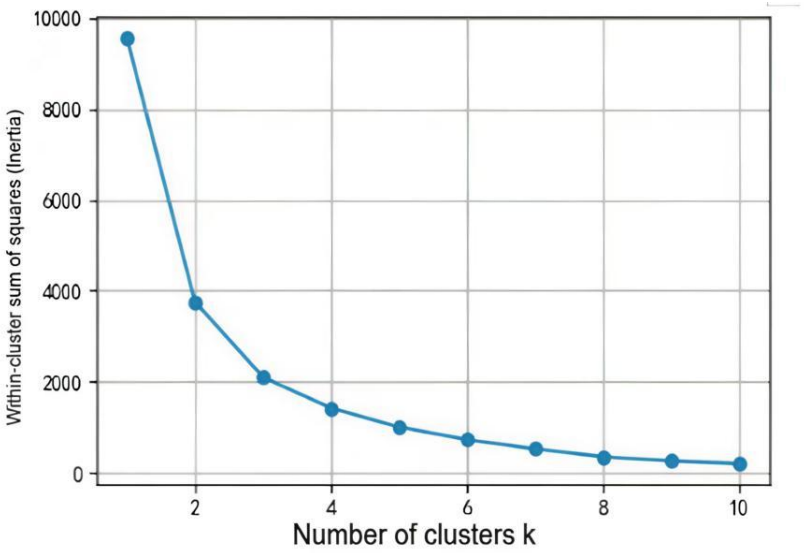


Figure 6. K-Means elbow point method analysis

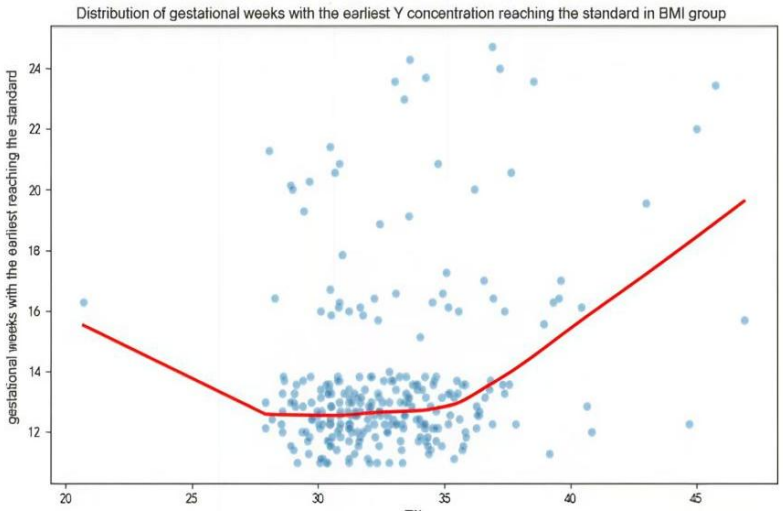


Figure 7. Distribution of gestational weeks with the earliest Y concentration reaching the standard in BMI group

According to Figure 6 and 7, as the BMI of pregnant women increases, the earliest Y chromosome concentration reaching the standard gestational age shows a delayed trend overall: most pregnant women in the low BMI group can reach the standard at 12-13 weeks, with a concentrated distribution and small fluctuations; The average delay for the moderate BMI group is 13-14 weeks, with increased individual differences, and some pregnant women need to wait until 15-16 weeks to reach the standard; The overall delay of the high BMI group is further delayed to 14-15 weeks or even later, and some pregnant women delay reaching the standard until more than 18 weeks, indicating that an increase in BMI significantly increases the risk of delayed testing.

3.2. Monte carlo

Due to detection errors in the "first standard gestational age", it is necessary to understand whether the statistical measures within different BMI groups are robust. Therefore, Monte Carlo random perturbation is introduced: Gaussian noise is added to the data in each group, and the simulation is repeated 1000 times to obtain the median mean and 95% CI.

$$\tilde{G}_g^{(b)} = \text{median}\left(\left\{G_i + \epsilon_i^{(b)} : i \in S_g\right\}\right), \quad \epsilon_i^{(b)} \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

$$\hat{G}_g = \frac{1}{B}\sum_{b=1}^{B} \tilde{G}_g^{(b)}, \quad CI_{95\%} = \text{quantile}(\tilde{G}_g^{(b)}, [0.025, 0.975]) \tag{7}$$

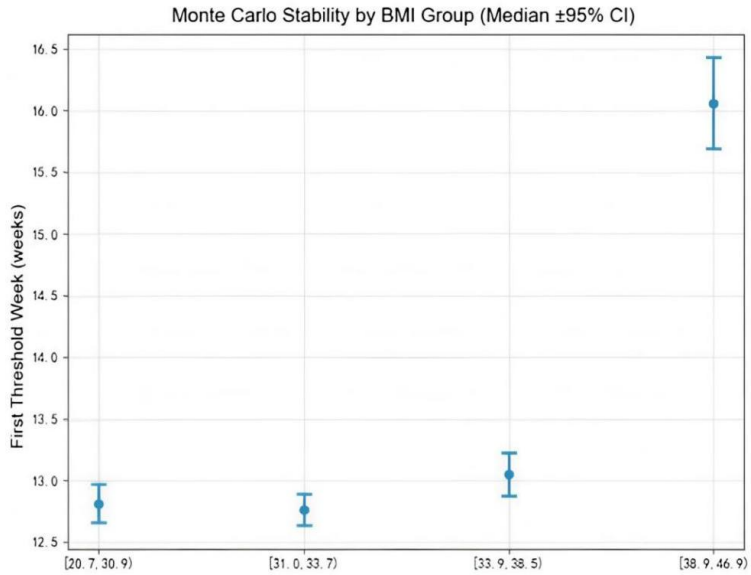


Figure 8. Monte Carlo robustness of each BMI group

As shown in Figure 8, the higher the BMI, the later the gestational age at which the first standard is reached, and the poorer the stability; Low BMI pregnant women can undergo NIPT earlier at 12-13 weeks, while high BMI pregnant women are advised to postpone testing until after 15-16 weeks to reduce the risk of failure.

3.3. Experimental result

After K-Means clustering, BMI ranges [20.70, 30.90), [31.00, 33.70), [33.90, 38.50), [38.90, 46.90] were obtained. The optimal detection time for different BMI groups is generally close, concentrated within the range of 14-16 weeks. This result is consistent with the literature report that "10-14 weeks is the optimal period for Cell Free Fetal DNA (cffDNA) detection", but in the high BMI group, the optimal gestational age for detection is slightly delayed, showing some differences. The final BMI range and corresponding optimal NIPT gestational age are shown in Table 3:

Table 3. BMI ranges and corresponding optimal NIPT detection gestational weeks

BMI group	Number of people	Average gestational age that meets the standard	Best detection time point	Minimum risk value
[20.7,30.9)	84	13.632653	13.6	0.196779
[31.0,33.7)	96	13.276786	14.1	0.215742
[33.9,38.5)	74	14.019305	14.7	0.240109
[38.9,46.9)	13	16.197802	15.4	0.240109

The results showed a positive correlation between surface BMI and the optimal timing for NIPT testing: the optimal gestational age for testing in the low BMI group was approximately 13.6 weeks, while in the high BMI group, it was postponed to 15.4 weeks. At the same time, the minimum risk value increases with the increase of BMI, with a risk value of 0.196 in the low BMI group, while the risk value of pregnant women with BMI ≥ 34 increases to about 0.24. Further analysis of detection errors revealed that the high BMI group had greater fluctuations in gestational age distribution, indicating poorer detection robustness and greater susceptibility to technical noise and experimental condition fluctuations.

It is recommended to adopt a stratified strategy in clinical testing: pregnant women with BMI<31 can undergo NIPT at 13-14 weeks to obtain stable results as early as possible; The optimal gestational age for pregnant women with a BMI between 31 and 34 is around 14 weeks; Pregnant women with a BMI ≥ 34 should appropriately delay until 14.5-15.5 weeks to ensure that the fetal DNA concentration reaches the detection threshold and reduce the risk of false negatives. At the same time, for high BMI populations, laboratories should strengthen quality control, such as increasing sequencing depth and optimizing library construction, to reduce the impact of detection errors and achieve more robust detection results.

4. Conclusions

The model proposed in this study demonstrates outstanding performance in terms of efficiency, flexibility, and stability. The K-means clustering method can efficiently complete sample grouping and has the characteristics of concise algorithm and good scalability; The Generalized Additive Model (GAM) captures nonlinear relationships between variables while maintaining good model interpretability; After further introducing ensemble learning methods such as XGBoost and random forest, the overall prediction accuracy of the model is significantly improved, and the results are more robust. In addition, the model comprehensively considers multiple variable relationships and ensures reliability through residual analysis and cross validation. The combination of SMOTE and other methods effectively alleviates the problem of data imbalance, further enhancing the practical application value of the model.

Although this study provides a comprehensive framework for optimizing NIPT detection strategies, several limitations remain. First, the model was developed based on retrospective data with limited sample size, which may restrict its generalization to broader populations. Second, the scope of maternal characteristics considered is relatively narrow, mainly focusing on gestational age and BMI, while other potentially influential factors such as ethnicity, lifestyle, or genetic background were not fully incorporated. Third, although ensemble learning methods improved predictive performance, the interpretability of some machine learning models remains limited in clinical settings.

In future work, we plan to expand the dataset by including larger and more diverse populations to enhance the robustness and external validity of the model. Additional maternal and fetal features will be integrated to improve the comprehensiveness of risk assessment. Moreover, prospective clinical trials will be conducted to validate the practical applicability of the proposed strategy. Finally, further exploration of interpretable machine learning algorithms will be pursued to balance predictive accuracy with transparency, thereby strengthening clinical decision support.

References

[1] Wang Yuanyuan, Guo Zheng, Li Guicai, etc Estimation of precipitation and its characteristics in the Three Gorges Reservoir area from 1979 to 2014 based on generalized additive model [J]. Acta Geographica Sinica, 2017, 72 (07): 1207-1220.

[2] He Sixuan, Yang Jiehao, Zhang Guoyou, etc Analysis of Vegetation Changes and Influencing Factors in Mining Damaged Areas in Central Yunnan Based on XGBoost HAP Model [J]. Surveying and Mapping Bulletin, 2025, (07):58-65.

[3] Li Xuan Research on Hospital triage model based on PCA and Random Forest [J]. China Medical Equipment, 2025, 40 (08): 38-42+76.

[4] Xing Li, Wu Shuang, Zhu Yinglin Clothing data analysis based on K-means clustering [J]. Digital Technology and Applications, 2025, 43 (04): 198-200.

[5] Zheng Xiaoliang, Dong Mengyuan, Xia Yingjie, etc Coal and gas outburst prediction based on improved Kmeans SMOTE-RF [J/OL]. Journal of Chongqing Technology and Business University (Natural Science Edition), 1-10 [September 7, 2025].

[6] Sun Y, Chen Y, Xu F, et al. Non-invasive prenatal testing for fetal chromosomal abnormalities by low-coverage whole-genome sequencing: recent progress and perspectives [J]. Frontiers in Genetics, 2021, 12: 689030.

[7] Wang L, Wang D, Xu Y, et al. Association of maternal BMI with cell-free fetal DNA fraction in non-invasive prenatal screening: a retrospective study [J]. BMC Pregnancy and Childbirth, 2022, 22 (1): 149.

[8] Zhao Q, Liu J, Chen Y, et al. Prediction of gestational diabetes mellitus using generalized additive models: a population-based study [J]. Chinese Journal of Preventive Medicine, 2023, 57 (8): 1132-1138.

[9] Li M, Zhou J, Fang H, et al. Application of machine learning models in prediction of non-invasive prenatal testing performance [J]. Scientific Reports, 2022, 12: 14587.

[10] Zhang L, Wang S, Liu J, et al. Maternal BMI classification based on K-means clustering and its association with pregnancy outcomes [J]. Maternal and Child Health Care of China, 2021, 36 (24): 5695-5699.

[11] Huang T, Ma D, Xu X, et al. Improving accuracy of NIPT by integrating maternal characteristics and sequencing features using ensemble learning [J]. Bioinformatics, 2023, 39 (3): btad079.