

Machine Learning-Based Optimization of Timing and Intelligent Determination of Fetal Abnormalities in Non-Invasive Prenatal Testing

Mianjie Huang^{*,#}, Zihan Lin[#], Cheng Xie[#], Mingying Yu, Zihao Zeng

School of Information Technology and Engineering, Guangzhou College of Commerce, Guangzhou 511363, China

*Corresponding author: Mianjie Huang (Email: fr_kazy@163.com)

[#]These authors contributed equally.

Abstract: Non-Invasive Prenatal Testing (NIPT) is a crucial technique in modern prenatal screening, holding significant value in detecting fetal chromosomal abnormalities. However, accurately predicting fetal Y chromosome concentration and optimizing the testing timing remain challenges in current clinical practice. This study employs machine learning methods to analyze the relationship between maternal physiological indicators and fetal Y chromosome concentration, aiming to construct a predictive model for optimizing NIPT testing strategies. The research begins with rigorous data quality control and preprocessing, including handling missing values, correcting outliers, and standardizing data. Through Pearson correlation analysis and coefficient of variation selection, gestational age, maternal BMI, and age were identified as core features. Addressing the severe multicollinearity among features, the study innovatively proposes a Gradient Boosting Random Forest hybrid model, effectively combining the anti-multicollinearity advantage of Random Forest with the sequential optimization capability of Gradient Boosting. Model validation results demonstrate that the hybrid model has significant predictive ability. Feature importance analysis further verifies the core roles of gestational age, BMI, and age in prediction. This research provides a new method for the intelligent prediction of fetal Y chromosome concentration in NIPT, contributing positively to improving detection accuracy and supporting clinical decision-making.

Keywords: Non-Invasive Prenatal Testing (NIPT); Machine Learning; Multidimensional Complex Bayesian Optimization.

1. Introduction

Non-Invasive Prenatal Testing (NIPT) is an important technique for screening fetal chromosomal abnormalities by analyzing cell-free fetal DNA in maternal peripheral blood. Known for being non-invasive, safe, and highly sensitive, NIPT has become a key component of prenatal screening. With its widespread application, enhancing the accuracy and timeliness of detection has become a focus of current research. Fetal Y chromosome concentration is a key indicator reflecting the amount of fetal DNA in maternal blood. Its variation is closely related to various factors such as gestational age and maternal physiological status. Therefore, building a model that can accurately predict Y chromosome concentration is significant for optimizing the timing of NIPT and improving the detection rate of fetal abnormalities. Currently, some studies have attempted to use traditional statistical methods like linear regression to establish relationships between Y chromosome concentration and maternal indicators[1-2]. However, due to common issues in medical data such as multicollinearity, nonlinear relationships, and individual differences, traditional models often struggle to achieve satisfactory results. With the advancement of machine learning technologies, algorithms like Random Forest and Gradient Boosting have shown great potential in medical data prediction, particularly excelling in handling complex feature relationships and resisting interference. Based on real clinical data, this study aims to optimize the prediction model for fetal Y chromosome concentration in NIPT using machine learning methods. The research begins with data quality control and feature selection, then proposes a Gradient Boosting Random Forest hybrid model to address

multicollinearity issues[3-4]. The model's performance is comprehensively evaluated through statistical tests and visualization analysis. The results not only provide a reliable predictive tool for clinical practice but also offer theoretical support for the intelligent decision-making of NIPT testing strategies.

2. Data Import and Preprocessing

The data is from <https://www.mcm.edu.cn/>. First, analyze the relationship between fetal Y chromosome concentration and maternal indicators; second, identify which indicators can predict Y chromosome concentration; finally, establish a prediction model and verify its effectiveness.

2.1. Data Quality Control

To ensure data quality, we adopted a strict data quality control process to guarantee the reliability of the analysis results. First, a comprehensive quality assessment was conducted on the raw data to identify and handle missing values. For missing values of Y chromosome-related indicators in female fetus data, they were filled with 0 based on biological principles. For missing values of other numerical variables, a median imputation strategy was used to reduce the impact of outliers.

When handling outliers, the interquartile range (IQR) method was employed, with boundaries set as $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$. Outliers beyond the boundaries were replaced with boundary values. This method maintains data integrity while avoiding the influence of extreme values on the model. For data standardization, to ensure all indicator columns can participate in calculations, the test date was uniformly converted to a standard format, the gestational age

was converted from "weeks + days" to total days, and the number of pregnancies greater than 3 was uniformly adjusted to 3.

We did not perform Z-score standardization or Min-Max normalization on numerical variables mainly based on the following considerations:

Many indicators in the dataset are discrete variables, such as the number of pregnancies and the number of blood draws. These variables have clear counting meanings themselves, and standardization may destroy their original numerical significance and interpretability.

Continuous variables, such as maternal BMI, age, and gestational age, have small numerical fluctuations and no order of magnitude differences.

The random forest algorithm is insensitive to feature scales, so good prediction results can be obtained without data standardization. Maintaining the biological significance of the original data facilitates clinical interpretation and application.

2.2. Feature Selection

Based on consulting the Encyclopedia of Chinese Medicine: Clinical Medicine [1], we identified 7 indicators closely related to maternal pregnancy status in the preprocessed dataset, which we refer to as physiological indicators. These indicators include gestational age, maternal BMI, age, number of pregnancies, number of blood draws, number of deliveries, and IVF pregnancy. They are all basic physiological or clinical information of pregnant women, and all may affect the fetal DNA concentration in maternal blood through biological mechanisms. Therefore, we believe they are suitable for fitting with fetal Y chromosome concentration.

There are clear reasons for selecting these indicators: gestational age reflects the fetal development stage; maternal BMI and age represent the maternal physiological state; the number of pregnancies and deliveries show the maternal

reproductive history; the number of blood draws reflects technical operation factors; IVF pregnancy indicates the impact of assisted reproductive technology. These 7 indicators cover the main aspects affecting fetal Y chromosome concentration, laying a foundation for establishing an accurate prediction model[5-6].

3. Pearson Correlation Analysis and Feature Selection

After completing data preprocessing and feature selection, the Pearson correlation coefficient was used to conduct correlation analysis on the selected indicators, evaluating the strength of the linear relationship between each physiological indicator and fetal Y chromosome concentration. The formula for the correlation coefficient is:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

where r_{xy} is the Pearson correlation coefficient between variables x and y . Since the correlation coefficient heatmap can show the correlation strength between Y chromosome concentration and each physiological indicator, a correlation coefficient heatmap was drawn after calculating the correlation coefficients.

Analysis of the heatmap (as shown in Figure 1) revealed that compared with other indicators, Y chromosome concentration has a strong positive correlation with the number of blood draws and gestational age, while it has a negative correlation with maternal BMI and age, both showing stronger negative correlations compared to other physiological indicators. Based on these results, we set a threshold of absolute correlation coefficient greater than 0.1 for feature selection, and finally selected four indicators—gestational age, BMI, age, and number of blood draws—as core features.

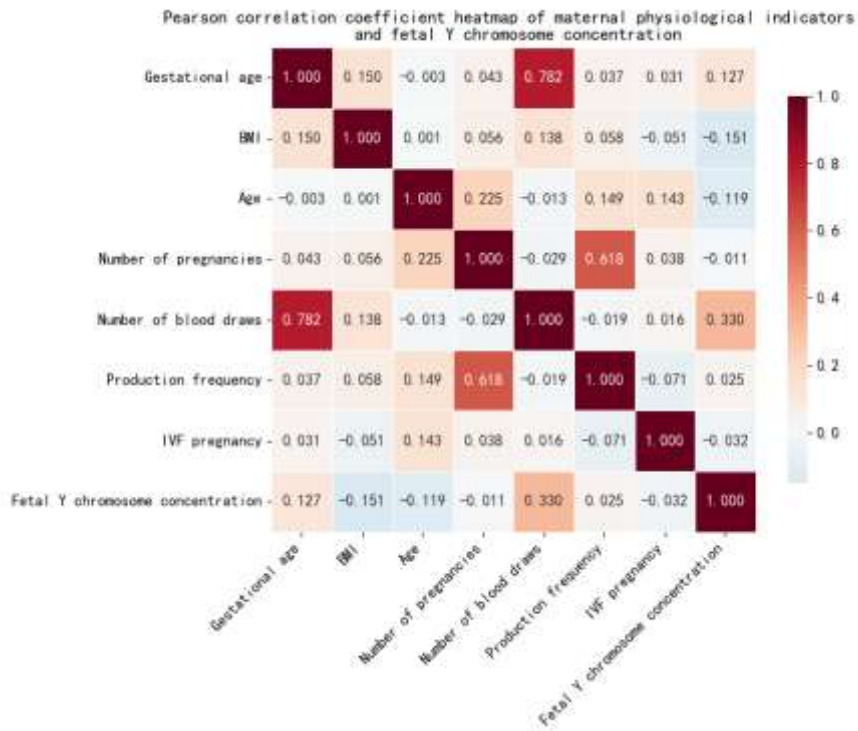


Figure 1 Pearson correlation coefficient heatmap of maternal physiological indicators and fetal Y chromosome concentration

The figure 1 shows the Pearson correlation coefficients between maternal physiological indicators (gestational age,

BMI, age, etc.) and fetal Y chromosome concentration, with values ranging from -1 to 1.

4. Descriptive Statistical Analysis

4.1. Statistical Description of Core Features

Descriptive statistical analysis provides basic information

for understanding data distribution characteristics and variable properties. Based on various statistics, we can comprehensively understand the distribution characteristics of each independent variable.

Table 1 Results of descriptive statistical analysis of core features

Statistic	Gestational Age (days)	Maternal BMI (kg/m ²)	Age (years)	Number of Blood Draws
Sample Size	1082	1082	1082	1082
Mean	117.92	32.29	28.94	2.49
Standard Deviation	28.53	2.97	3.66	1.13
Minimum Value	77	20.7	21	1
Maximum Value	203	46.88	43	5
Median	112	31.81	29	2
Skewness	0.608	0.864	0.615	0.03
Kurtosis	-0.793	2.649	1.169	-1.339
Coefficient of Variation	0.242	0.092	0.126	0.453

The descriptive statistical results (as shown in Table 1) indicate that the average gestational age of the pregnant women in the sample is 117.92 days, covering a wide range from the first trimester to the second trimester, which is consistent with the clinical application of NIPT. The average maternal BMI is 32.29, slightly higher than the normal range, but the coefficient of variation is small, indicating that the body mass index is relatively concentrated in the sample. The average age is 28.94 years old, which is consistent with the common characteristics of the current childbearing age. The average number of blood draws is 2.49, but its coefficient of variation is as high as 0.453, indicating significant differences in the distribution of this indicator, which may be affected by various factors such as individual health status and pregnancy management. Therefore, the number of blood draws may behave less stably in regression analysis. Overall, the basic characteristics of the sample are in line with clinical

expectations, but the fluctuation in the number of blood draws is relatively large, and its potential impact deserves further attention.

4.2. Coefficient of Variation Selection

Based on the above descriptive statistical results, we further used the coefficient of variation selection method to ensure the stability of features. The coefficient of variation is an important indicator to measure the degree of data dispersion, with the formula $CV = \sigma/\mu$, where σ is the standard deviation and μ is the mean. Coefficient of variation selection (as shown in Table 2) is a key step to ensure model stability. By eliminating features with excessively large coefficients of variation, the impact of data noise on model performance can be effectively reduced.

Table 2 Stability analysis results of feature selection using coefficient of variation

Feature	Coefficient of Variation	Selection Result	Selection Basis
Gestational Age	0.242	Retained	$CV \leq 0.3$
Maternal BMI	0.092	Retained	$CV \leq 0.3$
Age	0.126	Retained	$CV \leq 0.3$
Number of Blood Draws	0.453	Eliminated	$CV > 0.3$

Based on the selection criterion of coefficient of variation ≤ 0.3 , the number of blood draws was eliminated due to its excessively large coefficient of variation, and three core features—gestational age, maternal BMI, and age—were finally retained. This selection strategy effectively improves the stability and generalization ability of the model, ensuring the reliability of the prediction results. The theoretical basis for coefficient of variation selection is that features with

excessively large coefficients of variation often contain more noise information, which may lead to model overfitting and affect prediction accuracy and stability.

5. Multicollinearity Diagnosis

After determining the final feature set, we need to test for multicollinearity between these features, which is crucial for subsequent model selection.

5.1. Variance Inflation Factor (VIF) Analysis

Multicollinearity is a common problem in multiple regression analysis, which may lead to unstable parameter estimation and difficulty in interpretation. This study used the variance inflation factor (VIF) to diagnose multicollinearity between variables. The VIF formula is:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (2)$$

where VIF_j is the variance inflation factor of the j -th independent variable, and R_j^2 is the coefficient of determination when the j -th independent variable is regressed on all other independent variables, with a value range of $[0, 1]$.

Table 3 VIF diagnosis results of three core features

Feature	VIF Value	Multicollinearity Degree
Gestational Age	5.82	Severe multicollinearity
Maternal BMI	52.45	Severe multicollinearity
Age	4.96	Severe multicollinearity

The VIF diagnosis results (as shown in Table 3) indicate that all three core features have severe multicollinearity problems, with maternal BMI having the highest VIF value of 52.45. This severe multicollinearity is not uncommon in medical data, mainly due to the complex internal correlations between physiological indicators. The increase in gestational age is accompanied by systematic changes in maternal physiological status, and BMI is closely related to factors such as age and reproductive history.

5.2. Limitation Analysis of Linear Regression Models

Given the severe multicollinearity revealed by the VIF diagnosis results, we need to reconsider the modeling strategy. For regression problems, we initially intended to use a multiple linear regression model to establish the relationship between fetal Y chromosome concentration and maternal physiological indicators. However, in-depth analysis revealed some insurmountable limitations of this method.

First, there is severe multicollinearity between core indicators, which is difficult to eliminate. This is because physiological indicators in medical data have natural biological correlations. For example, maternal age, BMI, gestational age, and other indicators interact with each other, forming a complex network of relationships. Forcing the elimination of these correlations will destroy the biological significance and clinical interpretation of the indicators. Common multicollinearity handling methods, such as principal component analysis and ridge regression, can mathematically solve the multicollinearity problem but will cause the features to lose their biological significance. In addition, physiological indicators in the medical field usually

have nonlinear relationships, threshold effects, and complex interaction effects, and the assumptions of linear regression are difficult to adapt to these characteristics.

Based on the above-mentioned problems, we ultimately abandoned the linear regression method and chose the random forest algorithm, which has a natural ability to resist multicollinearity. Through bootstrap sampling and random feature selection mechanisms, random forest can effectively solve the multicollinearity problem while retaining the biological significance and clinical interpretability of the features.

6. Construction of Gradient Boosting Random Forest Hybrid Model

6.1. Algorithm Design Principle

Faced with severe multicollinearity problems and complex nonlinear relationships, this study adopted a gradient boosting random forest hybrid model as the core prediction algorithm. This innovative ensemble learning method combines the anti-multicollinearity advantages of random forest and the sequential optimization ability of gradient boosting [2].

Random forest naturally resists multicollinearity through bootstrap sampling and random feature selection mechanisms. The algorithm constructs B decision trees, each trained on a different bootstrap sample:

$$\hat{y}_{\text{RF}} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (3)$$

where: $\hat{y}_{\text{RF}}$ is the prediction value of the random forest; B is the total number of decision trees; $T_b(x)$ is the prediction result of the b -th decision tree for input x .

A gradient boosting mechanism is introduced for sequential optimization, and the prediction effect is gradually improved by fitting the residuals of the previous round of the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (4)$$

where $F_m(x)$ is the model prediction value after the m -th round of iteration, $F_{m-1}(x)$ is the model prediction value after the $(m-1)$ -th round, γ_m is the learning rate of the m -th round, and $h_m(x)$ is the weak learner that fits the residuals in the m -th round.

Finally, the hybrid model integrates the prediction results of random forest and gradient boosting through weighted average:

$$\hat{y}_{\text{hybrid}} = \alpha \cdot \hat{y}_{\text{RF}} + (1 - \alpha) \cdot \hat{y}_{\text{GB}} \quad (5)$$

where $\hat{y}_{\text{hybrid}}$ is the final prediction value of the hybrid model, $\hat{y}_{\text{RF}}$ is the prediction value of the random forest, $\hat{y}_{\text{GB}}$ is the prediction value of gradient boosting, and α is the weight parameter. In this study, $\alpha = 0.6$, indicating that the random forest dominates the final prediction.

6.2. Feature Importance Analysis

After constructing the model, we verified the contribution of each feature in the prediction through feature importance analysis.

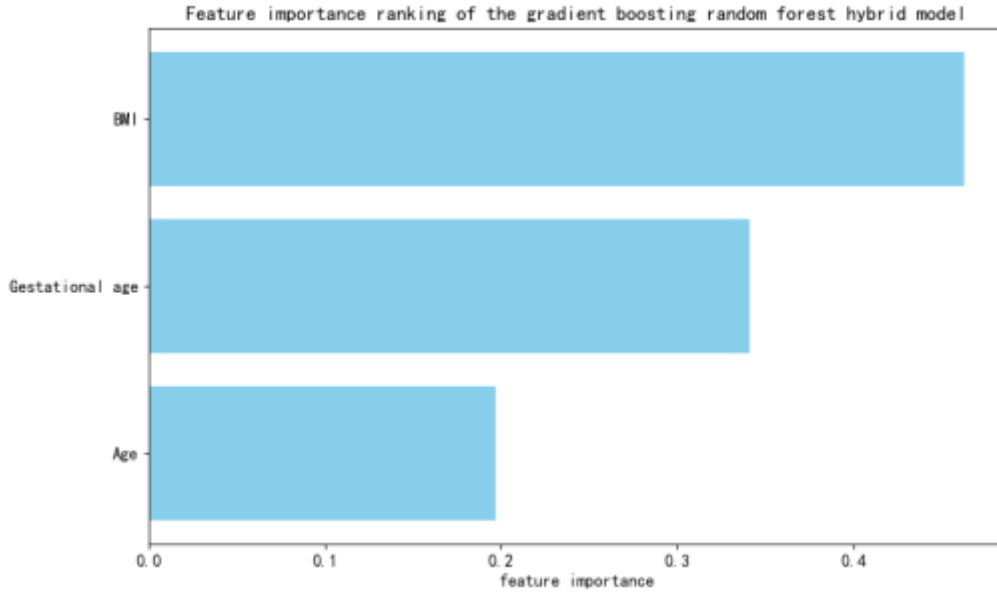


Figure 2 Feature importance ranking of the gradient boosting random forest hybrid model

The figure 2 shows the relative importance ranking of each feature in the gradient boosting random forest hybrid model, with the horizontal axis representing feature importance values.

The feature importance bar chart shows the relative importance ranking of each feature in the gradient boosting random forest hybrid model (as shown in Figure 3). The analysis results indicate that maternal BMI, an indicator calculated from height and weight, has the highest importance score, reflecting its strong correlation with the maternal physical condition during pregnancy and playing a core role in predicting fetal Y chromosome concentration. The importance score of gestational age is 0.338, reflecting the important impact of fetal development time on Y chromosome concentration. As gestational age increases, the concentration of fetal DNA in maternal blood shows regular changes. The importance score of age is 0.199, reflecting the comprehensive impact of maternal age on fetal development and maternal physiological status. The sum of the importance scores of the three features is close to 1.0, indicating that the model effectively captures key prediction information, providing reliable feature priority guidance for clinical applications.

7. Model Performance Evaluation and Statistical Testing

Based on the aforementioned gradient boosting random

forest hybrid model, this study established a comprehensive model evaluation system. This system aims to verify the statistical effectiveness and clinical application value of the model. The evaluation system includes three levels: statistical significance testing, prediction performance analysis, and residual diagnosis. We particularly focus on the overall significance of the model and the statistical significance of key features.

In model evaluation, two statistical methods were used: the F-test to evaluate the overall significance of the model, and the t-test to analyze the statistical significance of each feature. Through these two methods, the statistical performance of the model can be verified from multiple angles.

7.1. F-test Result Analysis

The F-test is used to evaluate the overall significance of the model, with the test hypothesis being H_0 : the model has no predictive ability ($R^2 = 0$) vs. H_1 : the model has predictive ability ($R^2 > 0$). The formula for the F-statistic is:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \quad (6)$$

Where k is the number of independent variables, n is the total number of samples, and $n - k - 1$ is the degree of freedom of the error term.

Table 4 F-test results

Dataset	R^2	F-statistic	p-value	Significance
Training Set	0.629	486.58	1.11×10^{-16}	Significant
Test Set	0.145	12.03	2.64×10^{-7}	Significant
Overall Data	0.535	413.55	< 0.001	Significant

The F-test results (as shown in Table 4) indicate that the overall data has $R^2 = 0.535$, which can explain 53.5% of the variation in the target variable. In the medical field, it is common for R^2 to be relatively low, mainly due to the

complexity and individual differences of biological systems. The R^2 of traditional medical prediction models is usually between 0.2 and 0.4; relatively speaking, the R^2 value of this study has reached a good level. Significance mainly depends

on the p-value, with values less than 0.05 considered significant. The p-values of all F-statistics are much less than 0.05, strongly rejecting the null hypothesis, proving that the model has extremely significant predictive ability.

7.2. T-test Result Analysis

On the basis of verifying the overall significance of the model, we further tested the statistical significance of each feature. A t-test was performed on each feature to evaluate its

statistical significance in the model. The t-statistic is calculated based on feature importance and its standard error:

$$t_j = \frac{\text{Importance}_j}{\text{SE}(\text{Importance}_j)} \quad (7)$$

Where t_j is the t-statistic of the j-th feature, Importance_j is the importance score of the j-th feature in the model, and $\text{SE}(\text{Importance}_j)$ is the standard error of the importance score of the j-th feature.

Table 5 t-test results

Feature	Importance	Standard Error	t-statistic	p-value	Significance
Maternal BMI	0.463	0.022	20.96	1.35×10^{-14}	Significant
Gestational Age	0.338	0.026	13.25	4.78×10^{-11}	Significant
Age	0.199	0.023	8.54	6.28×10^{-8}	Significant

The t-test results (as shown in Table 5) indicate that the p-values of the three core features are all much less than 0.05, indicating that these features have important statistical significance in predicting fetal Y chromosome concentration and are highly significant. Among them, maternal BMI has the highest t-statistic, reflecting its core position in the model; the t-statistics of gestational age and age both exceed 8.0, showing strong statistical significance. These results indicate

that in clinical applications, maternal BMI, gestational age, and age are the most critical physiological indicators among the influencing factors and deserve special attention.

7.3. Visualization of Model Prediction Effect

In addition to statistical testing, we also used visualization methods to intuitively evaluate the prediction effect and residual distribution characteristics of the model.

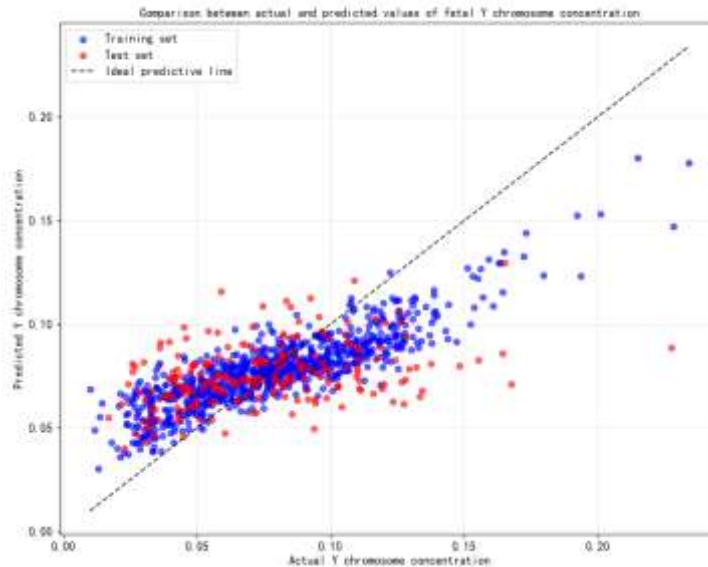


Figure 3 Comparison between actual and predicted values of fetal Y chromosome concentration

The figure 3 shows the scatter plot of actual vs. predicted values of fetal Y chromosome concentration for the training set and test set, with the dashed line representing the ideal prediction line.

The scatter plot of actual vs. predicted values intuitively shows the prediction accuracy and fitting effect of the model (as shown in Figure 4). Most data points of the training set and test set are distributed near the ideal prediction line, indicating that the model has a good fitting effect. The data points of the training set are more closely clustered around the

ideal prediction line, while the data points of the test set are relatively scattered, which is consistent with the general characteristics of machine learning models. From the figure, it can be observed that the model has a good prediction effect when the Y chromosome concentration is in the range of 0-20 ng/mL, and there is a certain prediction deviation in the high concentration range, which may be related to the relatively small number of high-concentration samples. Overall, the model has a good prediction effect, providing a reliable prediction tool for clinical applications.

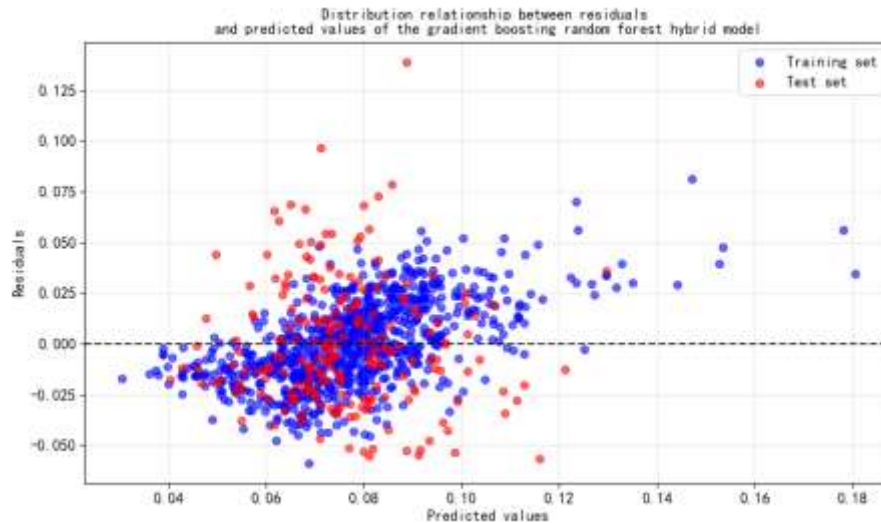


Figure 4 Distribution relationship between residuals and predicted values of the gradient boosting random forest hybrid model

The figure 4 shows the distribution of model prediction errors for the training set and test set, with the horizontal axis representing predicted values and the vertical axis representing residuals.

The residual analysis plot shows the distribution pattern and randomness characteristics of the model prediction errors (as shown in Figure 5). The residual points are relatively randomly distributed near the zero line, with no obvious systematic deviation or heteroscedasticity, indicating that the model assumptions are basically satisfied. The residuals of the training set are relatively concentrated, while the residuals of the test set are slightly scattered, which is consistent with the aforementioned overfitting phenomenon. From the residual distribution, it can be seen that most prediction errors are controlled within the range of ± 10 ng/mL, and the existence of a few extreme values may be related to individual physiological differences or data quality. The random distribution characteristics of the residuals indicate that the model successfully captures the main patterns in the data without missing important systematic information.

8. Conclusions

This study systematically analyzed the relationship between maternal physiological indicators and fetal Y chromosome concentration using machine learning methods and constructed a Gradient Boosting Random Forest hybrid prediction model. The results indicate that gestational age, maternal BMI, and age are the core features influencing Y chromosome concentration. The proposed hybrid model demonstrated good predictive performance and statistical significance on both the training and test sets, providing data support for optimizing the timing of NIPT. However, this study has certain limitations. Firstly, the sample source is relatively concentrated, which may affect the model's generalizability. Secondly, the number of samples with high Y chromosome concentration is small, leading to limited prediction accuracy in extreme value ranges. Additionally, the biological mechanisms underlying the feature relationships are not fully elucidated, and the model's interpretability needs

further enhancement. Future research will focus on the following aspects: First, expanding the sample size and data sources to improve the model's generalizability. Second, incorporating more potential influencing factors (such as maternal metabolic indicators and genetic background) to further optimize the feature set. Third, exploring advanced algorithms like deep learning to enhance the fitting capability for complex nonlinear relationships. Finally, integrating clinical practice to promote the application of the model in real-time testing and personalized diagnosis, ultimately achieving intelligent and precise NIPT testing procedures.

References

- [1] Xie Q Y, Wei S Y, Du H, et al. On-off-on electrochemiluminescence biosensor with dual-cycle amplification for ultrasensitive prenatal detection of thalassemia mutations using lanthanide europium nanorods[J]. *Analytica Chimica Acta*, 2025, 1378344712-344712.
- [2] Daglar K H, Kaymak D. The 'Y sign' as a novel qualitative marker for prenatal detection of aortic coarctation: a retrospective cohort study. [J]. *BMC pregnancy and childbirth*, 2025, 25(1): 1132.
- [3] Smart B H, Johnston M, Sands T M, et al. The Function of Non-invasive Prenatal Testing (NIPT) Request Forms in the Australian Context[J]. *Health Care Analysis*, 2025, (prepublish): 1-25.
- [4] Vedel C, Steensberg J, Petersen B O, et al. Prenatal detection and outcome of major heart defects in a country with universal screening.[J]. *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 2025.
- [5] Zhang Zhidong. Application of Gradient Boosting Random Forests and Parallel Fuzzy Decision Trees in System Modeling and Pattern Recognition [D]. Xidian University, 2023.
- [6] Chen Yewang, Cao Hailu, Chen Yi, Kang Zhao, Lei Zhen, Du Jixiang. A Survey of DBSCAN Acceleration Algorithms for Massive Data [J]. *Computer Research and Development*, 2023, 60(9): 2028-2047.