**Mathematical Modeling and Algorithm Application**
ISSN: 3006-0842 | Vol. 6, No. 3, 2025

# Research on the Timing Selection of NIPT and the Judgment of Fetal Abnormalities Based on Historical Data

**Xiangxiang Meng[*,#], Xingmin Zhou[#], Weiqu Cui[#]**

Department of Mechanical and Electronic Engineering College, Shandong Agriculture and Engineering University, Shandong 255300, China
[*]**Corresponding author**: Xiangxiang Meng (Email: mengmechatronics@163.com)
[#]**These authors contributed equally.**

**Abstract:** With the wide application of non-invasive prenatal testing (NIPT) technology, scientifically determining testing time points and effectively identifying fetal abnormalities have become hot topics. The integration of machine learning and mathematical modeling offers crucial support for related research. Based on actual NIPT data of pregnant women in a region, this paper focuses on two core tasks: Y chromosome concentration prediction and abnormal female fetal identification, constructing and optimizing multiple machine learning models. First, raw data was cleaned, missing values imputed and normality tested. Spearman correlation analysis explored relationships between Y chromosome concentration and relevant factors. A prediction model integrating random forest and gradient boosting regression was then built. After hyperparameter and ensemble optimization, its R2 rose from 0.5993 to 0.7268 with good significance, supporting quantitative NIPT timing selection. To address the scarcity of abnormal female fetal samples, the SMOTE algorithm balanced the database, and a random forest-based identification model was established. Results show the models significantly enhance detection reliability and sensitivity, providing a scientific basis for individualized clinical non-invasive prenatal screening decisions.

**Keywords:** NIPT detection; Random Forest; SMOTE algorithm; Y chromosome concentration.

## 1. Introduction

With the rapid development of current genomic sequencing technology and bioinformatics, non-invasive prenatal testing, as a detection technology based on fetal free DNA in the mother's peripheral blood, has been widely applied worldwide. In addition, with the extensive development of artificial intelligence and the massive increase in data volume, machine learning models have further advanced. Under this background, Guy Blanc et al. [1] proposed the distribution lifting theorem on the original machine learning model, providing an attribute framework for understanding the performance of the algorithm under the daytime distribution and enhancing the generalization ability of the machine learning model. In addition, Sharma et al.[2] using transformation methods such as DeepInsight, omics data with independent variables in tabular (table-like, including vector) form can be turned into image-like representations, enabling CNNs to capture latent features effectively. This approach not only enhances predictive power but also leverages transfer learning, reducing computational time, and improving performance. In terms of technological application, Wang et al. [3] Binary logistic regression analysis and restricted cubic spline (RCS) methods were used to assess the association between individual serum mineral levels and false-positive NIPT results, and concluded that there was a strong correlation between false positive NIPT and restricted placental chimerism. Subsequently, Wang et al. [4] conducted a further study on the pregnant woman population in Heilongjiang Province using a large sample of clinical data and found that NIPT enables non-invasive screening for fetal chromosome aneuploidy. Guillet et al. [5] studied the deterioration of ITP and the risk of NITP. To conclude, women with ITP do not increase their risk of severe bleeding during pregnancy. NITP is associated with NITP history and the severity of maternal ITP during pregnancy. Finally, some studies, including Ashoor et al. And Pergament et al. [6][7] have pointed out that the accuracy of NIPT detection is influenced by multiple factors such as the BMI of pregnant women, gestational age, and the proportion of fetal DNA.

In recent years, machine learning and statistical modeling methods have been widely introduced into the field of NIPT data analysis. Nguyen et al. [8] present an efficient computational approach to create positive samples with autosomal trisomy from negative samples. The results indicated that VINIPT is a powerful tool for detecting autosomal trisomy from low-coverage data; Zhang et al. [9] used the gradient boosting tree algorithm to predict the proportion of fetal cfDNA, significantly improving the accuracy of concentration estimation. Meanwhile, Gao et al. [10] constructed a fetal abnormality determination model based on deep neural networks, achieving parallel recognition of multiple chromosomal abnormalities. To address the issue of sample imbalance, He W. et al. [11] introduced the SMOTE oversampling technique to enhance the stability of model training and verified the effectiveness of the algorithm in small sample anomaly detection. In addition, foreign scholars have also attempted to combine ensemble learning with feature screening techniques to enhance the interpretability and clinical transferability of models. Zhao et al. [12] found through feature importance analysis that BMI of pregnant women, placental DNA fraction and GC content were the main variables affecting the test results, which was highly consistent with clinical practice.

Comprehensive analysis shows that the current research has achieved remarkable accomplishments in the evaluation of the detection accuracy of non-invasive prenatal testing (NIPT), the optimization of model algorithms, and the

interpretation of features. However, there are still several deficiencies: Firstly, quantitative research on the detection time points is relatively limited, and most studies only rely on empirical judgment or single-factor regression analysis for estimation.

Secondly, in the absence of the Y chromosome as a reference, the abnormality determination model for female fetuses still needs to be further improved in terms of accuracy and robustness. In view of this, based on the integration of relevant research results at home and abroad, this study utilized the actual NIPT detection data of pregnant women in a specific region, and constructed an ensemble learning model integrating random forest regression and gradient boosting regression to address the two core issues of predicting Y chromosome concentration in male fetuses and determining chromosomal abnormalities in female fetuses. In addition, this study also adopted the SMOTE algorithm to balance the sample distribution, and further improved the performance of the model through hyperparameter optimization and significance testing. The objective of this study is to provide new quantitative analysis methods and data support for the optimization of NIPT detection time points and the identification of fetal abnormalities, in order to promote the intelligent development of non-invasive prenatal testing technology in the field of precision medicine.

# 2. Method

## 2.1. Spearman correlation coefficient

In statistics, the spearman rank correlation coefficient named after Charles Edward Spearman, also known as the Spearman correlation coefficient. It is often represented by the Greek letter ρ. It is a non-parametric indicator that measures the dependence of two variables. It uses a monotonic equation to evaluate the correlation between two statistical variables, and its mathematical expression is:

$$\rho = 1 - \frac{6\sum_{t=1}^{n} d_i^2}{n(n^2-1)} \quad (1)$$

## 2.2. Random forest

Random forest is a machine learning algorithm, a (parallel) ensemble algorithm composed of decision trees, belonging to the Bagging type. By combining multiple weak classifiers, the final result is voted on or averaged, making the overall model result have high accuracy and generalization performance, as well as good stability. In this study, random forest regression was implemented by MATLAB, and its mathematical expression is

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (2)$$

Here, B represents the number of decision trees. $T_b(x)$ represents the prediction result of the B-th decision tree. For each decision tree during training, the self-sampling method is adopted to randomly draw samples from the original training set in a relaxed manner, and the optimal splitting feature is selected through the curvature splitting criterion. And the size of the smallest leaf node should be controlled between 1 and 10 to avoid overfitting, minimize the root mean square error of the split node:

$$MSE = \frac{1}{n_{node}}\sum_{i \in node}(y_i - \bar{y}_{node})^2 \quad (3)$$

Among, $\bar{y}_{node}$ It is the mean value of the target variable of the samples within the node. Specifically as shown in Figure 1.

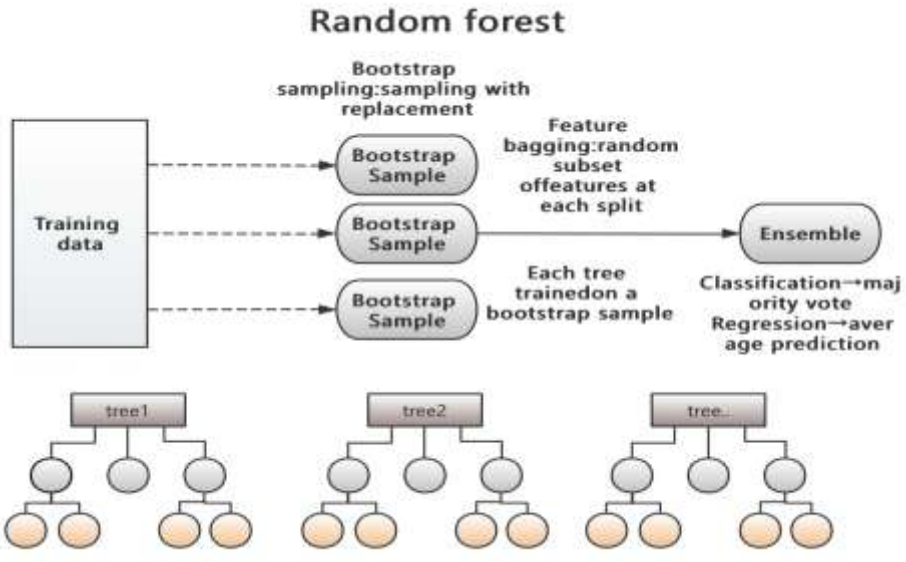


**Figure 1.** Structure of Random forest

Like many machine learning methods, the random forest model also has numerous parameters that need to be adjusted. These parameters have a significant impact on the model's performance. Inappropriate parameters often reduce the accuracy of the model's classification or prediction. Among them, the number of trees growing in the random forest model and the number of sample predictors at each split node have a significant impact on the model's performance and are also selected as the subsequent hyperparameter optimization targets. Subsequently, gradient boosting regression is used to optimize the model.

## 2.3. Gradient boosting regression(GBR)

Gradient boosting regression is an ensemble learning technique that builds a powerful prediction model by combining multiple weak prediction models. This study adopts the least squares lifting algorithm, and its objective function is

$$\min_f E(y - f(x))^2 \quad (4)$$

In each iteration, new model$h_m(x)$fits the current residuals:

$$h_m(x) = arg\min_h \sum_{i=1}^{n}\left[r_{m-1,i} - h(x_i)\right]^2 \quad (5)$$

Among $r_{m-1,i}$is the residuals after (m-1) iterations. The final prediction is the weighted sum of all weak learners:

$$\hat{y} = \sum_{m=1}^{M} \eta h_m(x) \quad (6)$$

In the formula, η represents the learning rate(sets 0.01-0.2), M represents the number of iterations.

## 2.4. SMOTE algorithm

Chawla et al. proposed the SMOTE algorithm in 2002. Its core idea is to synthesize new samples through linear interpolation in the feature space of minority class samples, thereby expanding the number of minority class samples. Suppose the sample set of the minority class is $S = \{x_1, X_2, X_3, \ldots . X_n\}$, for any one of the samples $X_i$, randomly select a sample from its k-nearest neighbor set$X_{zi}$, randomly select a sample from its k-nearest neighbor set:

$$X_{new} = X_i + \lambda(X_{zi} - X_i) \quad (7)$$

Among, λ is a randomly generated proportional coefficient used to control the interpolation position of the synthetic sample between the two. When λ=0.5, the synthetic sample is located at the midpoint between the original sample and its neighboring samples. By performing interpolation generation in the feature space, SMOTE avoids the problems of sample redundancy and overfitting caused by random replication, effectively improving the representativeness of minority class samples. This data is sourced from the National College Students Mathematical Modeling Competition, Question C. Among the female fetal samples, the abnormal proportion was only 10.42%, indicating a significant data imbalance. After balancing the training set through the SMOTE algorithm, the random forest classification model is adopted for training and validation.

# 3. Results

## 3.1. The establishment of simulation model

This study modeled and simulated the NIPT detection data based on the MATLAB platform. Firstly, the original samples were subjected to data cleaning and standardization processing, eliminating low-quality and abnormal samples. For male fetal samples, a model ensemble prediction framework integrating random forest regression and gradient boosting was constructed, and a prediction model for fetal Y chromosome concentration was established.

Since the K-S test results indicated that most variables did not follow a normal distribution, using the Spearman correlation analysis to quantitatively analyze the correlation between physiological indicators such as BMI, gestational weeks, and age of pregnant women and Y chromosome concentration, and eight significant factors were screened out. Subsequently, the model hyperparameters were optimized through grid search and cross-validation. The optimal parameters include: the number of decision trees 150, the minimum leaf node 3, the learning rate 0.05, and the maximum number of splits 100. After integration and optimization with the gradient boosting regression model, the model's coefficient of determination R2 increased from 0.5993 to 0.7268, with an increase of 21.3%. The F-test results indicated that the model was statistically highly significant.

The feature importance analysis shows that the concentration of the X chromosome, the detection date and the Z value of the Y chromosome are the main factors affecting the concentration of the Y chromosome. The visualization results show that the predicted values of the model are highly consistent with the actual values, which can well depict the nonlinear relationship between fetal DNA concentration and the physiological characteristics of pregnant women, providing a scientific basis for the quantitative selection of NIPT detection time points. For female fetuses samples, due to the abnormal samples accounts for only 10.42% of the total sample size, data sets imbalance problems. To prevent the model from biased towards normal samples during the training process, this paper adopts the SMOTE algorithm to balance the data. After the synthetic abnormal samples were generated by SMOTE, the data set was divided into the training set and the test set at a ratio of 8:2, and stratified sampling was adopted to keep the category ratios consistent. Subsequently, a model for determining chromosomal abnormalities in female fetuses was constructed using the random forest classification model to identify and predict the abnormal risks.

## 3.2. Analysis of experimental results

This study systematically evaluated the model performance through multiple indicators such as R-squared, p-value, accuracy rate, precision rate, recall rate, F1 score, and the area under the ROC curve.

In the male fetal sample, the regression model integrating random forest and gradient boosting performed exceptionally well, with a coefficient of determination R2 of 0.7268, an increase of 0.1275 compared to the initial model. The results indicated that the BMI of pregnant women, the date of detection, the concentration of X chromosome, the last menstrual period and the Z value of Y chromosome had a significant impact on the concentration of Y chromosome, verifying the important role of these indicators in the accuracy of NIPT detection. After the model was optimized, its ability to fit the trend of Y chromosome concentration changes was significantly enhanced, providing a quantitative basis for the optimal detection time at different stages of pregnancy.

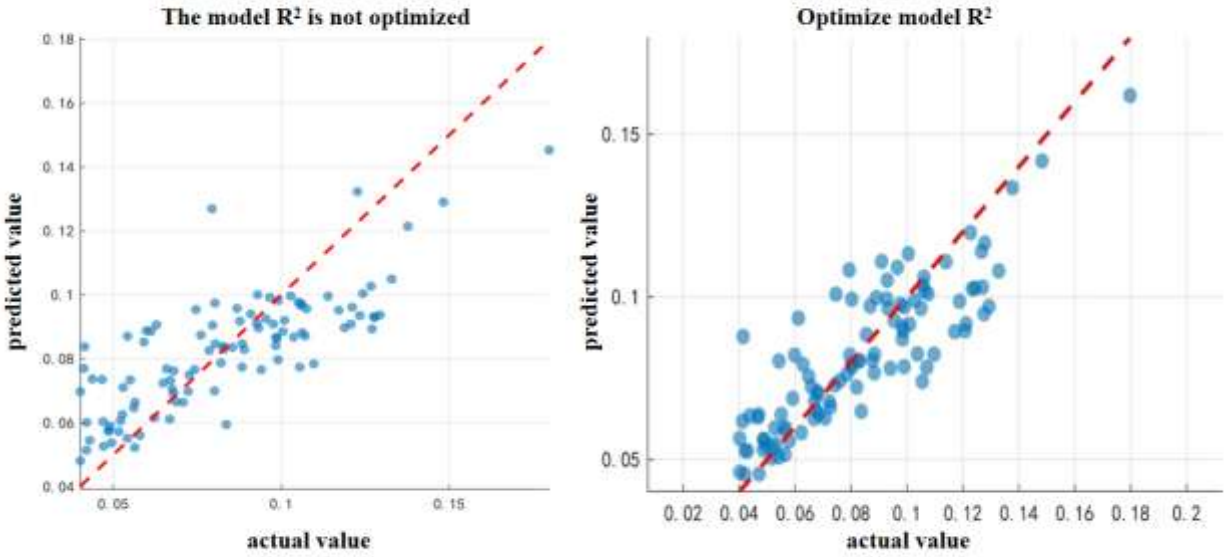


**Figure 2.** Model optimization results comparison

Based on the relationship between the actual Y concentration and the predicted Y concentration in Figure 2, the final performance R2=0.7269. Compared with the original performance R2=0.5993, the performance has improved by 0.1277, with an improvement rate of 21.3%. The R2 has increased from approximately 0.6 to approximately 0.73, indicating that the model's ability to explain the dependent variable has been significantly enhanced, and the optimization effect is quite obvious.

**Table 1.** Significance Verification

| Character | Correlation coefficient | P-value | Significant |
|---|---|---|---|
| Concentration of X chromosome | 0.34085 | 0.00042641 | Significant |
| Test date | -0.4771 | 3.4848e-07 | Significant |
| The Z value of the Y chromosome | 0.20003 | 0.042778 | Significant |
| Last menstrual period | -0.25035 | 0.010757 | Significant |
| Pregnancy BMI | -0.11082 | 0.026511 | Significant |
| Test gestational age | 0.065166 | 0.51312 | Non-significant |
| The GC content of chromosome 18 | -0.17683 | 0.073968 | Non-significant |
| The GC content of chromosome 21 | -0.12931 | 0.193 | Non-significant |

According to Table 1, there are a total of 5 features that have a significant correlation with the target variable. For instance, the detection date shows a moderately strong negative correlation with a correlation coefficient of -0.47771, and the P-value is extremely small, which has the greatest impact. The X chromosome concentration shows a moderate correlation coefficient of 0.34085, and the positive correlation value is much less than 0.05, indicating high reliability.

In the task of female fetal anomaly determination, the AUC of the random forest model balanced by SMOTE on the training set reached 0.9972, and on the test set, it was 0.7532. After threshold optimization($\tau^*$=0.492), The F1 score has increased from 0.400 to 0.462, indicating that the model has achieved a better balance between recall rate and precision rate. The cross-validation results show that the average F1 value is 0.8842, indicating good model stability. Table 2 below.

**Table 2.** Optimize before and after comparison

| Evaluation index | Before optimization(τ=0.5) | After optimization(τ∗=0.492) | Rangeability |
|---|---|---|---|
| Precision rate(Prec) | 0.3846 | 0.4348 | 13.05% |
| Recall rate(Rec) | 0.4167 | 0.5 | 19.99% |
| F1 | 0.4 | 0.4615 | 15.38% |

The random forest model was used to analyze the importance of features and generate the ranking of feature importance. As show in table 3.

**Table 3.** Feature importance ranking

| Influence factor | Value |
|---|---|
| The BMI of pregnant women | 0.1833 |
| The GC content of chromosome 13 | 0.1390 |
| The only number of read segments for comparison | 0.1070 |
| The Z value of the X chromosome | 0.1000 |
| The GC content of chromosome 21 | 0.0918 |
| The GC content of chromosome 18 | 0.0904 |
| The GC content | 0.0747 |
| The value of chromosome 21 | 0.0520 |

The BMI of pregnant women, the GC content of chromosome 13, and the number of read segments for unique alignment ranked the top three, with significantly higher importance than other features, and were the most critical influencing variables in the model. The Z value of the X chromosome, the GC content of chromosome 21, and the GC content of chromosome 18 are in the middle level and belong to secondary but still influential variables. The importance of subsequent features such as the Z value of chromosome 13 and the Z value of chromosome 18 gradually decreases and has a relatively weak impact on the model. From this, it can be concluded that the BMI of pregnant women is the main factor for abnormalities in female fetuses. The content of chromosome 13, GC, and the number of unique contrast segments also affect chromosomal abnormalities in female fetuses, and the larger the values, the greater the possibility of abnormalities in female fetuses.

# 4. Conclusions

Based on real NIPT detection data, a dual-model system for predicting Y chromosome concentration in male fetuses and determining abnormalities in female fetuses was constructed. Research has found that pregnant women BMI and gestational age has significant effects on the Y chromosome concentration, which can be used as important reference index of detecting point; The random forest combined with the SMOTE algorithm performs exceptionally well when dealing with imbalanced clinical data. The optimized model can effectively enhance the accuracy of NIPT detection and the scientific nature of time point selection, providing a new idea for precise prenatal screening.

Random forests combined with gradient promote regression integrated modeling method in the analysis of NIPT detection has high stability and interpreted. Its advantages include: a good ability to fit nonlinear relationships and multi-feature interactions; Robustness against missing data; Visualizing the importance of model features enhances clinical interpretability.

Because this model of random forests for the stability of the noise data are limited. The next step of this research is to introduce a noise filtering processing module and build a multi-model fusion framework in combination with the gradient boosting tree support vector machine model to enhance the data's anti-interference ability. Secondly, combined the sample size of multiple clinical institutions, and included the data of pregnant women from different regions, ages and underlying diseases. Enhanced the generalization ability through feature engineering to optimize the model. Finally, introduce the SHP/LIME interpretability algorithm. Ultimately, Formed an integrated analysis system to promote the transformation and application of scientific research achievements in clinical practice.

# References


[1] Blanc G, Lange J, Strassle C, et al. A Distributional-Lifting Theorem for PAC Learning[J]. arXiv preprint arXiv,2025,2506.16651.

[2] Sharma A, Lysenko A, Jia S, et al. Advances in AI and machine learning for predictive medicine[J]. Journal of Human Genetics, 2024, 69(10): 487-497.

[3] Wang G, Zhang Y, Wu T, et al. Serum Zn and Ca correlate with false positive non-invasive prenatal test results[J]. American Journal of Translational Research, 2025, 17(3): 2135.

[4] Wang C, Zhang H, Liu Y, et al. Results and indications changes by prenatal diagnosis of 3,458 pregnant women: a 10-year retrospective study[J]. Annals of Medicine, 2025, 57(1): 2563753.

[5] Guillet S, Loustau V, Boutin E, et al. Immune thrombocytopenia and pregnancy: an exposed/nonexposed cohort study[J]. Blood, 2023, 141(1): 11-21.

[6] Ashoor G., Syngelaki A., Poon L. C., et al. Fetal fraction in maternal plasma cell-free DNA at 11–13 weeks'gestation: relation to maternal and fetal characteristics. [J].Journal of ultrasound Obstet Gynecol, 2013, 41(1): 26–32.

[7] Pergament E., Cuckle H., Zimmermann B., et al. Single-nucleotide polymorphism-based noninvasive prenatal screening in a high-risk and low-risk cohort. [J].Journal of Obstetrics and Gynecology, 2014, 124: 210–218.

[8] Nguyen T H, Nguyen H V, Pham M D, et al. An efficient computational method to create positive NIPT samples with autosomal trisomy[C]2024 12th International Conference on Bioinformatics and Computational Biology (ICBCB). IEEE, 2024: 126-131.

[9] Zhang H., Gao Y., Jiang F., et al. Noninvasive prenatal testing for trisomy 21, 18 and 13—Clinical experience from 146,958 pregnancies. [J].Journal of ultrasound Obstet Gynecol, 2015, 45(5): 530–538.

[10] Gao F., Lin J., Liu X., et al. Application of deep learning in prenatal screening and genetic diagnosis. [J].Journal of Frontiers in Genetics, 2021, 12: 667.

[11] He W., Wang X., Li J., et al. SMOTE-based data augmentation for imbalanced NIPT datasets. [J].Journal of IEEE Access, 2022, 10: 45062–45074.

[12] Zhao Y., Chen H., Zhang X., et al. Feature importance and interpretability in machine learning models for prenatal testing. [J]. Journal of Computers in Biology and Medicine, 2023, 156: 106760.