A Multiple Regression-Based Method for NIPT Timing Selection and Fetal Abnormality Detection

Yansan Zhou [1, *, #], Bin Jiang [1, #], He Gao [2, #]

[1] School of mathematical sciences & school of pre-university, Dalian Minzu University, Dalian, China, 116650

[2] Computer Science and Engineering College of Dalian Minzu University, Dalian, China, 116650

* Corresponding Author Email: 2416177611@qq.com

#These authors contributed equally

Abstract. Non-invasive prenatal testing (NIPT) is a widely applied clinical technique for detecting fetal chromosomal abnormalities by analyzing cell-free fetal DNA fragments in maternal plasma. The accuracy and reliability of NIPT are affected by multiple factors, including gestational age and maternal body mass index (BMI). This paper investigates the correlation between Y chromosome concentration, gestational weeks, and BMI in male fetuses, as well as the feasibility of predicting female fetal abnormalities in the absence of Y chromosome signals. To address these issues, this article establish a comprehensive framework combining interaction regression analysis, gradient boosting decision trees (GBDT), rule-based grouping, random forests, and logistic regression models. Based on historical data, optimal NIPT time points are estimated for different BMI groups by calculating the gestational weeks at which 95% of pregnant[1] women achieve the required fetal DNA concentration, with additional safety margins introduced to enhance robustness. Experimental results show that the optimal testing window for male fetuses is approximately 11 ± 3 weeks, while the proposed model achieves a precision of 0.8178 and a recall of 0.8621 for female fetal abnormality detection. The findings provide practical clinical recommendations to minimize potential testing risks, improve detection success rates across maternal BMI groups, and offer effective judgment methods for female fetuses lacking Y chromosome markers.

Keywords: Non-Invasive Prenatal Testing, Interaction Regression, Gradient Boosting Decision Trees, Random Forest, Logistic Regression.

1. Introduction

Fetal health is a critical concern for families and healthcare providers. Genetic disorders and congenital abnormalities not only threaten infant survival but also impose heavy psychological and economic burdens. The emergence of non-invasive prenatal testing (NIPT) [1], which analyzes cell-free fetal DNA (cffDNA) fragments in maternal peripheral blood, has transformed prenatal screening by enabling earlier, safer, and more accurate detection of chromosomal abnormalities compared with traditional invasive procedures. Specifically[2], NIPT is effective in detecting trisomies such as Down syndrome (chromosome 21), Edwards syndrome (chromosome 18), and Patau syndrome (chromosome 13), as well as abnormalities in sex chromosomes [2, 3].

The reliability of NIPT depends on multiple biological and clinical factors. In male fetuses, the concentration of Y chromosome fragments must reach a threshold (typically ≥4%) for accurate testing. Previous studies have shown that gestational age and maternal body mass index (BMI) are strongly correlated with fetal DNA concentration, influencing the earliest gestational week at which reliable detection can be achieved [4]. Thus, determining the optimal testing window for pregnant women with different BMI values is essential to minimize testing failure, reduce false negatives, and improve clinical decision-making.

For female fetuses, the absence of Y chromosome markers introduces additional challenges. Their abnormality detection must rely on other chromosomal features, including trisomy-related Z-scores, GC content, sequencing depth, and related ratios [3]. This motivates the need for robust statistical

and machine learning models that can integrate multiple maternal and fetal indicators for accurate diagnosis.

This study develops and evaluates a multi-model framework for optimizing NIPT timing and fetal abnormality prediction. Our contributions are threefold:

(1) Correlation modeling: This article analyzes the relationship between Y chromosome concentration, gestational weeks, and BMI using interaction regression models, confirming significant joint effects.

(2) BMI-based stratification: By applying rule-based grouping and ensemble methods (GBDT, random forests), the study determines optimal NIPT windows across different BMI intervals, introducing safety margins to account for measurement uncertainties[4].

(3) Female fetal abnormality detection: The study designs logistic regression and boosting-based models to evaluate chromosomal indicators and achieve robust classification performance in the absence of Y chromosome data.

The proposed methods provide quantitative guidance for clinical practice, ensuring that NIPT can be performed at an early yet reliable stage, thereby improving screening accuracy, minimizing maternal risk, and supporting personalized prenatal care.

2. Methodology

2.1. Assumptions and Preliminaries

To ensure tractable modeling and analysis, the article makes the following assumptions:

(1) Fetal chromosomal abnormalities are primarily associated with sex chromosomes, chromosome 21, chromosome 18, and chromosome 13, and other chromosomal anomalies are excluded.

(2) The testing environment is stable, and external conditions do not affect the reliability of NIPT measurements.

(3) Maternal health is assumed to be normal, and factors beyond BMI, gestational age, and age are considered negligible for the purposes of this study.

2.1.1. Symbol Definitions

Let the main notations be defined as follows:

b_0 intercept term.

b_1, b_2, b_3: effect coefficients.

$\hat{b}_i$: estimated regression coefficients.

$s_{\hat{b}_i}$: standard error of the coefficient estimates.

ST: mean gestational age at which fetal DNA reaches the target concentration.

n: number of samples in a group.

m: number of samples reaching the concentration threshold within the group[5].

w_i: weight assigned to each risk component.

$s(x)$: mapping function to normalize values into [0,1] [0,1] [0,1].

2.2. Interaction Regression Model (Correlation Analysis)

To analyze the joint influence of gestational weeks and BMI on Y chromosome concentration in male fetuses, the study establishes a multiple linear regression model with an interaction term:

$$Y = b_0 + b_1 X_{week} + b_2 X_{BMI} + b_3 (X_{week} \cdot X_{BMI}) + \varepsilon \quad (1)$$

Where Y denotes Y chromosome concentration, X_{week} the gestational weeks, X_{BMI} the maternal BMI, and ε the random error term.

b0: intercept, representing the baseline level of Y chromosome concentration when both gestational weeks and BMI are set to zero (a theoretical reference point rather than a biologically realistic value).

b_1: effect of gestational age when BMI is fixed.

b_2: effect of BMI when gestational age is fixed.

b_3: interaction coefficient, expected to be negative, implying that higher BMI attenuates the effect of gestational age on Y concentration.

Model validity is tested using the coefficient of determination (R^2) and t-tests for significance.

2.3. Rule-Based Grouping (BMI Stratification)

Given that maternal BMI significantly affects the earliest gestational week at which Y chromosome concentration reaches the clinical threshold (≥4%), the study designs a rule-based grouping method to stratify pregnant women into five BMI categories. For each group, the study determines the earliest reliable NIPT week by calculating the 95th percentile gestational week at which the threshold is satisfied [6].

The grouping rule can be represented as:

$$\mathrm{f(x)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{I(x \in G_k)} \tag{2}$$

Where G_k is the k-th BMI group and $I(\cdot)$ is an indicator function. When rules are mutually exclusive, each sample belongs to exactly one group, ensuring clear clinical stratification.

2.4. Random Forest Model (Multifactor Integration)

To incorporate additional maternal characteristics such as age, height, and weight, the study extends the analysis using a random forest model [5]. This ensemble approach is robust to measurement noise and non-linear feature interactions.

Each decision tree is trained on a bootstrap sample, using mean squared error (MSE) as the splitting criterion:

$$\mathrm{MSE(f,t)} = \frac{|\mathrm{S_1}|}{|\mathrm{S}|}(\mathrm{T_1} - \overline{\mathrm{T}})^2 + \frac{|\mathrm{S_2}|}{|\mathrm{S}|}(\mathrm{T_2} - \overline{\mathrm{T}})^2 \tag{3}$$

Where S_1, S_2 are subsets after splitting by feature f at threshold t, and $T_1, T_2, \overline{T}$ represent mean attainment times.

The objective function for optimized BMI intervals is defined as:

$$\mathrm{F} = \mathrm{Ar} + \mathrm{Bd} \tag{4}$$

Where R is the mean risk coefficient, D the deviation from the target attainment ratio (≥95%), and α, β are weighting parameters[7].

2.5. Gradient Boosting and Logistic Regression (Female Fetus Abnormality Detection)

For female fetuses, in which Y chromosome markers are absent, the study constructs classification models based on chromosomal and maternal features (Z-scores of chromosomes 21, 18, and 13; X chromosome metrics; GC content; sequencing depth; BMI).

2.5.1. Gradient Boosting Decision Trees (GBDT)

The boosting procedure iteratively fits decision trees to pseudo-residuals:

(1) Compute residuals:

$$\mathrm{r_{im}} = -\left[\frac{\partial \mathrm{L}(\mathrm{y_i}, \mathrm{F_{m-1}(x_i)})}{\partial \mathrm{F_{m-1}(x_i)}}\right] \tag{5}$$

(2) Fit a tree to residuals and update the model:

$$\mathrm{F_m(x)} = \mathrm{F_{m-1}(x)} + \eta \sum_{\mathrm{j=1}}^{\mathrm{J}} \gamma_{\mathrm{jm}} \mathrm{I}\left(\mathrm{x} \in \mathrm{R_{jm}}\right) \tag{6}$$

2.5.2. Logistic Regression

To complement tree-based models, the study applys logistic regression [6] for binary classification:

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0+\beta^T x)}} \quad (7)$$

This ensures interpretability and provides baseline probability estimates for abnormality detection.

3. Experiments and Results

3.1. Data Preprocessing

The dataset, obtained from the 2025 Higher Education Press Cup National Undergraduate Mathematical Contest in Modeling (CUMCM 2025), The dataset includes maternal clinical characteristics (gestational weeks, BMI, height, weight, age) and sequencing-based indicators (chromosome Z-scores, GC content, read depth). To ensure data quality, the following preprocessing steps were applied:

Removal of missing or abnormal values (e.g., negative concentrations, BMI outliers, GC content outside 40–60%).

Exclusion of extreme maternal records (age <19 or >45 years, height <1.4 m or >1.9 m).

For male fetuses, retention of the earliest measurement where Y chromosome concentration exceeded 4%.

For female fetuses, exclusion of cases with incomplete chromosomal data[8].

This ensured stable input quality and prevented bias due to measurement errors or outliers.

3.2. Male Fetuses: Correlation Analysis

The article first applied the **interaction regression model** to analyze Y chromosome concentration as a function of gestational weeks and maternal BMI. The estimated regression equation was:

$$Y = -0.00004X_{week} + 0.000032X_{BMI} + 0.00643(X_{week} \cdot X_{BMI}) + 2.0469 + \varepsilon \quad (8)$$

All coefficients were statistically significant ($p < 0.05$), with the interaction term confirming that higher BMI weakens the positive effect of gestational weeks on Y concentration. The adjusted correlation coefficient was $R = 0.388$, demonstrating moderate predictive strength[9].

3.3. Optimal Testing Window via BMI Stratification

To determine optimal NIPT timing across BMI groups, the study applied rule-based grouping. Pregnant women were divided into five BMI intervals with equal spacing. For each group, the gestational week at which 95% of fetuses achieved the Y concentration threshold was estimated.

Figure 1 shows the relationship between BMI groups and optimal NIPT time. Results indicated that the optimal testing week for most groups was approximately 11 weeks (±3), with slight variations depending on BMI. Introducing a 0.5-week safety margin ensured robustness against sequencing errors [10].

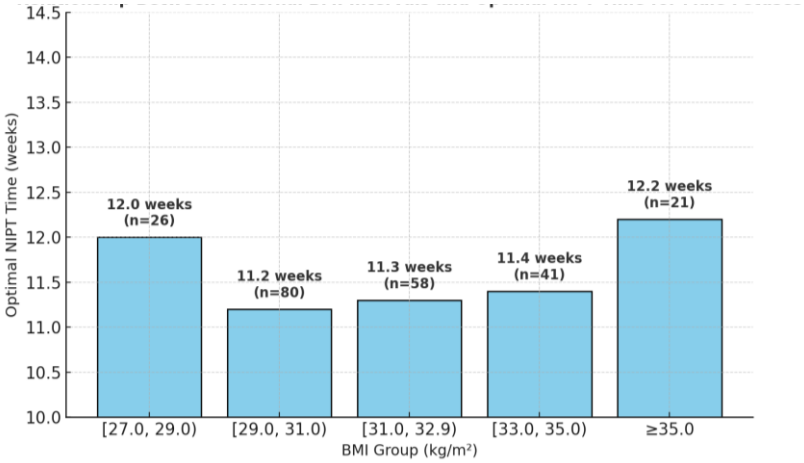


Figure 1. Relationship between maternal BMI intervals and optimal NIPT time for male fetuses.

3.4. Multifactor Random Forest Modeling

Extending the analysis, the study incorporated maternal age, height, and weight into a random forest model (200 trees, maximum depth = 8). The model predicted attainment times for each BMI subgroup and refined the grouping intervals[11].

The optimized BMI intervals were determined as:

[27.00, 29.29], [29.34, 30.95], [31.05, 32.76], [32.87, 34.72], [34.93, 36.93]

Figure 2, all groups consistently indicated an optimal NIPT week of 11 weeks, confirming the robustness of the stratification.

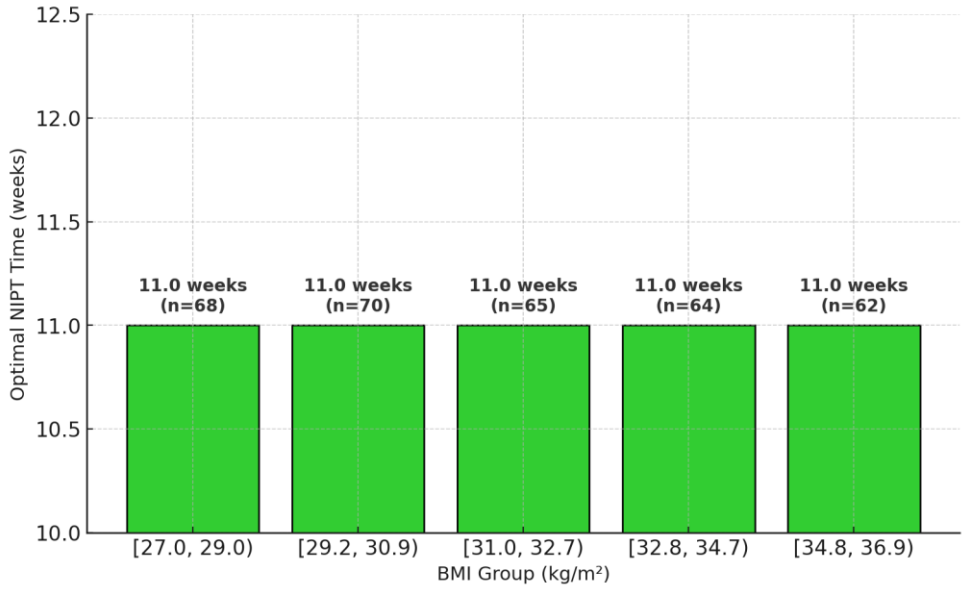


Figure 2. Refined BMI groupings and predicted optimal NIPT times using random forest modeling.

3.5. Female Fetuses: Abnormality Detection

For female fetuses lacking Y chromosome markers, the study constructed predictive models based on chromosomal and sequencing features (chromosomes 21, 18, 13, and X; GC content; sequencing depth; BMI).

The gradient boosting decision tree (GBDT) and logistic regression models were evaluated. Performance metrics are summarized in Table 1.

Table 1. Performance of abnormality detection models for female fetuses.

Class	Precision	Recall	F1-score	Support
Normal (0)	0.892	0.961	0.925	77
Abnormal (1)	0.250	0.600	0.143	10
Overall Accuracy	–	–	–	86.21%

3.6. Summary of Findings

(1) Male fetuses: Significant correlation exists between gestational weeks, BMI, and Y chromosome concentration. The optimal NIPT window is around 11 weeks ± 3.

(2) BMI stratification: Rule-based grouping and random forest models provide clinically interpretable stratifications, ensuring ≥95% reliability across BMI groups[12].

(3) Female fetuses: GBDT and logistic regression models achieve promising accuracy (86.2%), offering effective alternatives to Y-based detection.

These findings provide strong empirical evidence to guide clinical recommendations, reduce false negatives, and support risk-aware prenatal decision-making.

4. Discussion

4.1. Interaction Regression Model

The interaction regression model effectively quantifies the combined influence of gestational weeks and BMI on Y chromosome concentration.

Strengths: The model is interpretable and statistically robust, clearly indicating that higher BMI attenuates the positive impact of gestational weeks. Coefficient significance tests (t-tests) confirm the reliability of the results.

Limitations: The approach assumes homoscedastic errors and normally distributed residuals, which may not hold in practice given sequencing failures and measurement errors. Furthermore, the model is limited to linear interactions and cannot easily incorporate nonlinear maternal factors such as age and height.

Improvements: Robust regression techniques or weighted fitting could mitigate the impact of abnormal data. Extending the model with polynomial or spline terms could better capture nonlinear patterns while controlling multicollinearity.

4.2. Rule-Based Grouping

Rule-based grouping provides an intuitive clinical tool to stratify BMI and recommend NIPT timing.

Strengths: The approach is simple, easy to implement, and provides clear clinical guidance without requiring complex computation. It ensures that ≥95% of cases in each group achieve the Y concentration threshold.

Limitations: The grouping is heuristic and may fail to capture fine-grained variations. It primarily considers BMI and overlooks other influential maternal characteristics (e.g., maternal age, conception method).

Improvements: Hybrid grouping strategies that combine clustering (e.g., K-means) with clinical rules could improve objectivity and granularity. Additionally, adaptive grouping schemes could dynamically adjust for measurement errors[13].

4.3. Random Forest Model

The random forest model integrates multiple maternal characteristics to refine the prediction of optimal NIPT timing.

Strengths: The ensemble approach captures nonlinear interactions, handles noisy data, and automatically ranks feature importance. The model confirms BMI as the dominant predictor while incorporating additional factors such as age and height.

Limitations: Despite its predictive accuracy, the model lacks interpretability, making it challenging for clinicians to understand why specific recommendations are made. Performance may also degrade for small subgroups, such as women with very high BMI.

Improvements: Model explainability can be enhanced by incorporating SHAP values or feature attribution methods. Oversampling or weighting strategies may improve robustness in underrepresented subgroups.

4.4. Gradient Boosting and Logistic Regression for Female Fetuses

Gradient boosting and logistic regression models were used to detect female fetal abnormalities in the absence of Y chromosome markers.

Strengths: The boosting framework is well-suited for handling complex feature interactions, while logistic regression provides interpretable probability estimates. The combined approach achieves an accuracy of 86.2%, demonstrating practical feasibility.

Limitations: Sensitivity to rare abnormal cases remains low, as shown by reduced precision in minority classes. Both models are sensitive to missing or biased sequencing features such as abnormal GC content. Training efficiency is also a concern when sample sizes increase.

Improvements: Missing-value imputation methods and data augmentation techniques can improve robustness. Incorporating XGBoost or hybrid models may enhance both accuracy and computational efficiency.

4.5. Overall Evaluation

Comparative analysis highlights complementary strengths across models. Interaction regression provides interpretability, rule-based grouping supports clinical decision-making, random forests ensure predictive robustness, and boosting methods enable reliable female abnormality detection. However, challenges remain in integrating heterogeneous data sources, handling sequencing noise, and improving interpretability for clinical adoption. Future work should focus on hybrid frameworks that combine the interpretability of regression models with the predictive power of ensemble learning, while expanding validation on larger and more diverse clinical datasets.

5. Conclusions

This paper proposes a multi-model integrated framework, incorporating interaction regression, random forests, and other models, to optimize the timing of non-invasive prenatal testing (NIPT) and detect fetal abnormalities. It also provides quantitative clinical recommendations tailored to different maternal BMI groups and fetal types.

The study found that gestational weeks and maternal BMI jointly influence the concentration of the Y chromosome, with a significant negative interaction effect. For male fetuses, the optimal timing for NIPT is approximately 11 ± 3 weeks, which is stable across all BMI subgroups; introducing a safety margin can further reduce the risk of test failure. Relevant models offer recommendations for populations with different BMI levels, ensuring that over 95% of cases meet the detection threshold. When there is no Y chromosome signal, the relevant models achieve an accuracy of 86.2% in detecting abnormalities in female fetuses.

This method can provide practical guidelines for prenatal testing. In the future, the model will be optimized through validation with multi-center data, integration of advanced algorithms, and other means, and its application will be extended to fields such as oncology, laying a foundation for personalized clinical decision-making.

References

[1] Shi Weihui, Xu Chenming. The Application Value of Non-invasive Prenatal Testing in the Diagnosis of Maternal Complications and Comorbidities in Obstetrics [J]. Practical Journal of Obstetrics and Gynecology, 2025, 41 (08): 617-620.

[2] Jiang Liya, Lu Shaokai, Du Jiayin, et al. Development and Application of Non-invasive Prenatal Testing Technology [J]. Clinical Medical Research and Practice, 2025, 10 (23): 191-194.

[3] Yu Dandan, Li Fengjin, Yao Xinyu, et al. Efficacy Analysis of NIPT in Detecting Common Chromosomal Aneuploidies in Fetuses of Different High-Risk Groups [J]. China Family Planning and Obstetrics and Gynecology, 2025, 17 (03): 72-76+82.

[4] Qian Yue, Jin Kerqin, Hu Min, et al. Screening Value of Expanded Non-invasive Prenatal Testing for Pregnant Women at High Risk of Down Syndrome [J]. Zhejiang Clinical Medicine, 2025, 27 (02): 180-183.

[5] Li Na, Li Yipei, Yang Weixian, et al. Analysis of the Application Value of Non-invasive Prenatal Testing Technology in the Diagnosis of Fetal Sex Chromosome Aneuploidies [J]. China Maternal and Child Health Care, 2024, 39 (24): 4963-4966.

[6] Li Shanshan, Zhang Meng, Lv Wei, et al. Prenatal Diagnosis and Genetic Analysis of NIPT Cases with Chromosome 9 Abnormalities [J]. Journal of Clinical Laboratory Medicine, 2024, 21 (22): 3380-3387. DOI: CNKI: SUN: JYYL.0.2024-22-023.

[7] Wu Jianfang, Liang Lisheng, Zheng Zhi, et al. Analysis of the Results of Non-invasive Prenatal Testing for Trisomy 21 Screening Based on Single Indicator MOM Value [J]. China Prenatal Diagnosis Journal (Electronic Edition), 2024, 16 (03): 30-35.

[8] Pan Shuqing, Pan Xiaoli, Ge Lizha, et al. The Application Value of Non-invasive Prenatal Testing in Prenatal Screening [J]. Zhejiang Medicine, 2024, 46 (17): 1881-1884.

[9] Li Jingjing, DAVID Lubasi Nalisa, Sun Zongli. Study on the Current Status of Non-invasive DNA Prenatal Testing among Pregnant Women in a Certain Area of Henan Province from 2021 to 2022 [J]. Modern Disease Prevention and Control, 2024, 35 (08): 613-616+625.

[10] Zhang Hongyun. Research on the Development and Applicability Evaluation of Non-invasive Prenatal Screening Technology for 66 Dominant Single Gene Diseases [D]. Central South University, 2023.

[11] Yuan Xiying. Clinical Application Study on Fetal Chromosome Aneuploidies Screening and Prenatal Diagnosis in Twin Pregnancy with Non-invasive Prenatal Testing [D]. Central South University, 2023.

[12] Cheng Yan, Ma Wen, Quan Qiuning, et al. The Application Value of DNA Non-invasive Prenatal Testing and Color Doppler Ultrasound Examination in the Screening of Fetal Chromosome Abnormalities in High-Risk Pregnant Women [J]. Journal of Clinical Laboratory Medicine, 2024, 21(02):217-220.

[13] Wang Yalin. Non-invasive prenatal testing: Towards the next decade [J]. Clinical Pediatrics Journal, 2024, 42(01):15-19.