Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 1, No. 2, 2026

A Multivariate Statistical Modeling Approach for Non-Invasive Prenatal Testing

Gan Long *

School of Energy Engineering, Chengdu University of Technology, Chengdu, 610059, China

* **Corresponding author**: (Email: 13890744796@163.com)

Abstract: The accuracy of non-invasive prenatal testing (NIPT) is significantly influenced by the fetal fraction and maternal physiological indicators. This study aims to enhance the precision of fetal health assessment by constructing multivariate statistical and machine learning models to optimize clinical decision-making. First, to analyze the correlation between maternal factors and Y-chromosome concentration, One-way ANOVA and Ordinary Least Squares (OLS) regression were employed; results indicated that gestational age and BMI are key predictors, with Random Forest Partial Dependence Plots (PDP) utilized to interpret non-linear relationships. Second, to address BMI-based stratification and timing, the Jenks Natural Breaks method was applied to divide BMI into eight groups, identifying an optimal testing window of 11–13 weeks, validated through Monte Carlo simulations to minimize detection risks. Third, a multi-factor optimization was achieved using an improved K-means clustering algorithm integrating height, weight, and age, which identified five refined BMI categories with an optimal testing point around 14 weeks. This model quantifies the contribution of technical errors through a weighted risk assessment. In conclusion, this integrated framework shifts NIPT from single-factor analysis to multi-dimensional optimization. Future work will incorporate multi-center data and Long Short-Term Memory (LSTM) networks to predict dynamic trends in fetal DNA concentrations.

Keywords: Non-Invasive Prenatal Testing, Multivariate Statistical Models, Optimal Timing for Testing, Monte Carlo.

1. Introduction

Non-invasive Prenatal Testing (NIPT) has revolutionized clinical diagnostics by analyzing cell-free fetal DNA from maternal blood to detect chromosomal abnormalities, such as Trisomy 21, 18, and 13. The clinical reliability of NIPT is fundamentally contingent upon the fetal fraction—specifically the Y-chromosome concentration in male fetuses. For a result to be deemed accurate, this concentration must typically reach or exceed a 4% threshold. However, early detection is a double-edged sword: while detecting anomalies before the 12th week significantly minimizes medical and psychological risks, testing too early often leads to insufficient fetal fractions and sequencing failure. Since factors like maternal Body Mass Index (BMI) and gestational age significantly influence these DNA levels, a "one-size-fits-all" testing schedule often fails to balance accuracy with the need for an early intervention window.

Current literature has evolved from simple correlation studies toward integrated predictive and optimization frameworks. In the domain of concentration prediction and quality control, methodologies utilizing multivariate regression and random forests have been established to pre-emptively identify low-FF risks based on sequencing read counts and maternal phenotypes [1, 2]. Building upon these predictive foundations, the focus of clinical decision-making has shifted toward probabilistic optimization. Techniques such as Bayesian decision frameworks and Generalized Additive Models (GAM) have enabled dynamic timing recommendations that balance detection accuracy against the risk of delayed intervention [3, 4]. Furthermore, population stratification has emerged as a vital mechanism for enhancing model robustness. Recent evidence demonstrates that coupling clustering algorithms for niche-group identification with ensemble learning significantly reduces false-negative rates in high-risk cohorts [5, 6]. Collectively, these advancements suggest that transitioning from empirical rules to multi-factor, non-linear modeling is the inevitable trajectory for NIPT optimization [7].

Building upon these foundational studies, this research presents a comprehensive modeling framework to address three critical gaps in NIPT optimization. First, we investigate the quantitative correlation between maternal indicators and Y-chromosome concentration. By employing One-way ANOVA for discrete variables and Ordinary Least Squares (OLS) regression for continuous predictors, we identify the significant roles of gestational age and BMI, further utilizing Random Forest Partial Dependence Plots (PDP) to interpret non-linear interactions. Second, to solve the problem of BMI-based stratification, the Jenks Natural Breaks method is applied to categorize patients into eight optimized groups. We determine the optimal NIPT timing (11–13 weeks) for each group by minimizing potential clinical risk, validated through Monte Carlo simulations to assess the impact of detection errors. Third, we implement an improved K-means clustering algorithm that integrates multi-factor inputs including height, weight, and age. This refined model quantifies technical error contributions and employs a weighted risk assessment to establish a robust five-tier BMI classification system with a recommended testing point around 14 weeks. By shifting from empirical grouping to multi-dimensional optimization, this study provides a scientific basis for enhancing NIPT precision and clinical workflow efficiency.

2. Methods and Model Development

2.1. Correlation Analysis of Fetal Fraction via OLS and ANOV

To quantify the linear correlation between fetal Y-chromosome concentration and maternal characteristics such

as gestational age and Body Mass Index (BMI), independent variables were categorized into discrete and continuous types. For discrete indicators, integer values were assigned as group identifiers, and One-way Analysis of Variance (ANOVA) was performed; the significance of their impact on Y-chromosome concentration was evaluated using the F-test and associated p-values. For continuous indicators, a Multiple Linear Regression (MLR) model was constructed. Coefficients were estimated via Ordinary Least Squares (OLS), and the significance of each predictor was validated through t-tests and p-values.

2.1.1. Multivariate Linear Regression and Statistical Inference

Assuming linear relationships exist between various factors (e.g., gestational weeks, body mass index, etc.) and fetal Y chromosome concentration (to be validated through residual analysis), this study establishes a multiple linear regression model with these factors as independent variables and fetal Y chromosome concentration as the dependent variable to analyze their linear correlation characteristics. When considering only one variable, such as the impact of gestational weeks on Y chromosome concentration, the linear regression model between the two variables is as follows:

$$Y = \beta_0 + \beta_1 G + \varepsilon \tag{1}$$

In the formula, β_0 is the intercept, β_1 is the regression coefficient of gestational age, and ε is the random error term.

Based on this, and considering the effects of all independent variables on fetal Y chromosome concentration, the model was extended as follows:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 BMI_i + \beta_3 Age_i + \beta_4 Height_i + \beta_5 Weight_i + \beta_6 AlignRatio_i + \beta_7 DupRatio_i + \beta_8 GC_i + \varepsilon \tag{2}$$

Where $i = 1, 2, ..., n$ (n is the sample size), Y_i is the value of the independent variable for the i-th sample, $\beta_0, \beta_1, ..., \beta_8$. is the estimated regression coefficient, and ε_i is the random error term for the i-th sample.

In a multiple linear regression model, the linear influence of each variable on is quantified by the regression coefficient preceding each independent variable.

2.1.2. Non-linear Interpretation via Random Forest and PDP

(1) Random Forest Regression Model

Random Forest is an ensemble learning method that improves model accuracy and stability by constructing multiple decision trees and combining their prediction results. It is capable of handling high-dimensional data and possesses strong resistance to overfitting. First, Bootstrap sampling is performed by sampling the training set with replacement to construct M subsets, each of size n. M distinct decision trees are generated, and for each tree, the optimal features are selected for splitting. The m Bootstrap sample is:

$$\{x_i^{*m}, y_i^{*m}\}_{i=1}^{n}, m = 1,2,\cdots,M(4) \tag{3}$$

Where x_i^{*m} represents the m optimal features selected from the K feature variables provided in the problem (e.g., gestational age, maternal BMI), and y_i^{*m} is the target scalar, i.e. Y chromosome concentration.

Next, the output of the random forest is generated. Each decision tree $T_j(x)$ classifies the input sample x and outputs a class $\hat{y}_j$. The final predicted class of the random forest $\hat{y}$ is determined by a vote of all decision trees:

$$\hat{y} = vote(\{T_1(x), T_2(x), \ldots, T_m(x)\}) \tag{4}$$

In the formula, vote represents the majority vote, and $T_j(x)$ is the prediction made by the j decision tree for the sample 'x'.

Finally, the final prediction of the Random Forest is the average of the predictions from the M decision trees. That is:

$$\hat{y} = \frac{1}{M}\sum_{j=1}^{M} T_i(x) \tag{5}$$

(2) PDP Dependency Interpretation

PDP dependency graphs serve as a tool to illustrate how machine learning models are influenced by specific features, particularly when dealing with complex nonlinear models. By fixing one or several features while allowing others to take all possible values, PDP graphs calculate the model's predicted output. This approach eliminates the influence of other features, revealing the average effect of target features on model predictions. The computational formula for PDP is as follows.

$$\hat{f}(x_j) = \frac{1}{n}\sum_{i=1}^{n} f(x_j, x_{i,j}) \tag{6}$$

In this equation, $\hat{f}(x_j)$ represents the PDP value for the feature x_j, i.e. the partial dependence of the regression result on x_j as it takes on different values.

2.1.3. Statistical Inference and Significance Testing

The reliability of the proposed models was validated through a rigorous two-tier significance testing protocol.

Coefficient-level Inference: A two-tailed Student's t-test was performed on each regression coefficient β_1 to test the null hypothesis H0: $\beta_1 = 0$. This determines the independent contribution of factors such as gestational age and blood collection frequency to Y-chromosome concentration.

Model-level Validation: An F-test was employed to assess the overall significance of the regression, comparing the explained variance (SSR) against the residual variance (SSE).

For discrete variables, One-way Analysis of Variance (ANOVA) was utilized to decompose total variation $(SST = SSB + SSE)$. The ANOVA F-statistic was used to evaluate whether significant differences in Y-chromosome concentration exist across different categorical strata, thereby identifying key discrete determinants of fetal fraction. Finally, model performance was evaluated using the Coefficient of Determination (R^2) and Adjusted R^2 to quantify the proportion of variance explained while penalizing for model complexity.

2.2. BMI-Based Cohort Stratification and Optimal Timing Framework

To minimize clinical risk and diagnostic failure, we developed a two-stage optimization framework that stratifies maternal cohorts based on Body Mass Index (BMI) and identifies the optimal gestational timing for NIPT.

2.2.1. Earliest Effective Time of Threshold Attainment

To quantify the time to meet the target for each pregnant

woman, based on the chronological order of test records, we take the gestational week at which the condition $Y_{i,j} \geq \theta$ is first satisfied, i.e.:

$$T_i = \min\left\{GA_{i,j} \mid Y_{i,j} \geq 0.04, j = 1,2,\ldots,n_i\right\} \quad (7)$$

2.2.2. Objective Function for Grouping

Furthermore, whilst satisfying the basic conditions, the objective of grouping is to maximize inter-group variance and minimize intra-group variance (i.e., to adhere as closely as possible to the principle of homogeneity).

The core function for minimizing within-group variation is:

$$\text{Minimize} \quad SSW = \sum_{k=1}^{K} \sum_{i \in Group_k} (BMI_i - \mu_k)^2 \quad (8)$$

In the equation, $\mu_k = \frac{1}{|Group_k|} \sum_{i \in Group_k} BMI_i$ and μ_k represent the mean BMI of group k, whilst $|Group_k|$ represents the number of pregnant women in group k.

The inter-group maximization objective function is:

$$\begin{cases} \text{Var}(Group_k) = \frac{1}{|Group_k|-1} \sum_{i \in Group_k} (T_i - \bar{T}_k)^2 \\ \text{Var}(All) = \frac{1}{M-1} \sum_{i=1}^{M} (T_i - \bar{T})^2 \end{cases} \Rightarrow GVF = 1 - \frac{\sum_{k=1}^{K} \text{Var}(Group_k)}{\text{Var}(All)} \quad (9)$$

Where $\text{Var}(Group_k)$ is the variance of T_i for group k, and $\text{Var}(All)$ is the total variance of T_i for all valid pregnant women.

2.2.3. Objective Function for The Optimal Timing

The objective for the optimal detection time is to detect the condition as early as possible (minimizing risk) whilst ensuring compliance. For the group k, the objective function is defined as:

$$\text{Minimize} \quad Opt_{GA,k} \quad (10)$$

The objective for the optimal time is to detect as early as possible whilst ensuring compliance, i.e. to minimize risk. Depending on whether the constraints are satisfied, the solution for the optimal time is divided into two scenarios: whether there is a low-risk compliance time within the group:

Quantifying the impact of measurement error on the time to meet the target:

To investigate the effect of measurement error on the time to meet the target, and to study the deviation between the measured value and the true value, we assume that the test value $Y_{i,j}$ and the true value $Y_{i,j}^{\text{true}}$ satisfy:

$$Y_{i,j} = Y_{i,j}^{\text{true}} + \varepsilon_{i,j} \quad (11)$$

Where $\varepsilon_{i,j} \sim N(0, \sigma^2)$ (the detection error follows a normal distribution with mean 0 and variance σ^2).

The earliest true time of acceptance is defined as:

$$T_i^{\text{true}} = \min\left\{GA_{i,j} \mid Y_{i,j} - \varepsilon_{i,j} \geq \theta, j = 1,2,\ldots,n_i\right\} \quad (12)$$

This formula uses the test value minus the error to perform a secondary assessment of the time of compliance, thereby evaluating the interference of the error on the 'time of first compliance'.

2.3. Multi-factor Synergistic Optimization and Integrated Risk Assessment

To refine NIPT clinical pathways, we developed an integrated optimization framework that accounts for maternal physiological heterogeneity, sequencing-induced measurement noise, and the longitudinal dynamics of fetal fraction.

2.3.1. Individual Compliance Status

To assess whether each individual sample meets the testing accuracy requirements, the individual compliance status is defined as follows:

$$I_i = \begin{cases} 1, & Y_{\text{Concentration},i} \geq 4\% \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

Where I_i is the compliance status (binary variable) of the sample (i), and $Y_{\text{Concentration},i}$ is the Y-chromosome concentration (%) of the sample (i).

2.3.2. Population Compliance Rate

To reflect the population compliance status of a specific BMI group at a given gestational week, the following is derived: $P_{k,t}$, where $P_{k,t}$ represents the compliance rate (%) of the k th BMI group at gestational week t.

$$P_{k,t} = \frac{1}{|G_k|} \sum_{i \in G_k} I_i \cdot I(GestationalWeek_i \leq t) \times 100\% \quad (14)$$

This formula aggregates individual compliance status into a population-level indicator, reflecting the feasibility of testing for that group at t weeks, and serves as a core input for calculating potential risk. Here, G_k denotes the sample set for the k BMI group, $|G_k|$ denotes the number of samples in the group, and $I(GestationalWeek_i \leq t)$ denotes the indicator function for `"the gestational age at the time of testing for the i sample is $\leq t$.

2.3.3. Potential Risk

$Risk_{k,t}$ The potential risk is a parameter value derived by combining the "risk of subsequent testing due to non-compliance" with the "risk of induced abortion due to increased gestational age". This formula uses 12 weeks as the dividing line between the early and mid-term screening periods, applying a segmented logic to the reasonable testing window, with different potential risks for each segment.

$$Risk_{k,t}=\begin{cases} w_{early}\times\left(1-\dfrac{P_{k,t}}{100}\right) & \text{if } t\in[10,12] \\ w_{mid,non}\times\left(1-\dfrac{P_{k,t}}{100}\right)+w_{mid,week}\times(t-12) & \text{if } t\in(12,25] \end{cases} \tag{15}$$

$w_{early}=0.1$: Early non-compliance risk weight (early induced labor complication incidence <1%, low risk); $w_{mid,non}=0.8$: Mid-term non-compliance risk weight (mid-term induced labor complication incidence>5%, high risk); $w_{mid,week}=0.2$: Mid-term gestational age progressive risk weight (for every additional week of gestational age, the risk of induced labor bleeding increases by 2%, reflecting the risk escalation rate).

2.3.4. Detection of Individual and Group-Mean Errors

To quantify the interference of sequencing quality on concentration measurement, and σ_i determine individual detection error (%):

$$\begin{cases} \sigma_i=0.3\times(1-\dfrac{AlignmentRate_i}{100})+ \\ 0.2\times\dfrac{DuplicationRate_i}{100}+0.2\times\dfrac{FilterRate_i}{100}+0.3\times\dfrac{|GC_i-50|}{100} \\ \bar{\sigma}k=\dfrac{1}{|G_k|}\sum i\in G_k\sigma_i \end{cases} \tag{16}$$

Where, $AlignmentRate_i$ is the reference genome alignment ratio (%) for the i sample; $DuplicationRate_i$ is the repeat read ratio (%), and $FilterRate_i$ is the filtered read ratio (%); GC_i represents the GC content (%); $\bar{\sigma}_k$ represents the average detection error (%) for the BMI groups numbered 1 through k.

2.3.5. Core Function

The core problem of this question is to determine, for each BMI group, the gestational week at which the 'potential risk to the pregnant woman' is minimized, provided that the testing criteria are met. Therefore, the core function for this question is:

$$\min_{t_k\in T_k} Risk_{k,t_k} \tag{17}$$

Where t_k is the optimal NIPT timing for the k BMI group, and T_k is the set of valid gestational weeks: $T_k=\{t|\ P_{k,t}\geq 80\%\}$.

3. Results

3.1. Identification of Factors Influencing Fetal Free DNA Concentration

3.1.1. Multiple Linear Regression Analysis

The R^2 values before and after overall adjustment in multiple linear regression (all <0.025) were significantly low, indicating poor regression performance, which further suggests that the relationship in this study is predominantly nonlinear.

In the regression coefficient significance test results, to detect gestational age, (β=0.001264, p=0.001264) showed an extremely significant positive correlation with the target variable Y chromosome concentration, while the other regression coefficients were not significant.

3.1.2. Factor Analysis of Variance

According to the homogeneity of variance test results, age and production frequency exhibited homogeneity, while other integer-type variables showed heterogeneity. The ANOVA F-test results demonstrated that age (F=4.04, p<0.01) and blood sampling frequency (F=10.70, p<0.001) had significant effects on Y chromosome concentration.

3.1.3. PDP Interpretability Analysis

According to the random forest feature importance results, gestational age, body weight, age, ratio of comparison, BMI, and number of blood draws were the top six most important features. The final regression performance of the model showed an R^2 value of 0.797496 and an RMSE of 0.014337.

Pregnancy duration and body weight were identified as the two most significant influencing factors. Analyzing the dual-feature impact on Y chromosome concentration, the optimal combination of pregnancy duration (24.0000) and body mass index (BMI) (75.5000) yielded the highest predicted Y chromosome concentration value of 0.097606. The right Figure 1 demonstrates the influence analysis of both pregnancy duration and BMI. Overall, it is evident that increased feature quantity leads to more complex mechanisms affecting Y chromosome concentration, further illustrating the intricate nonlinear relationship between features and Y chromosome concentration.

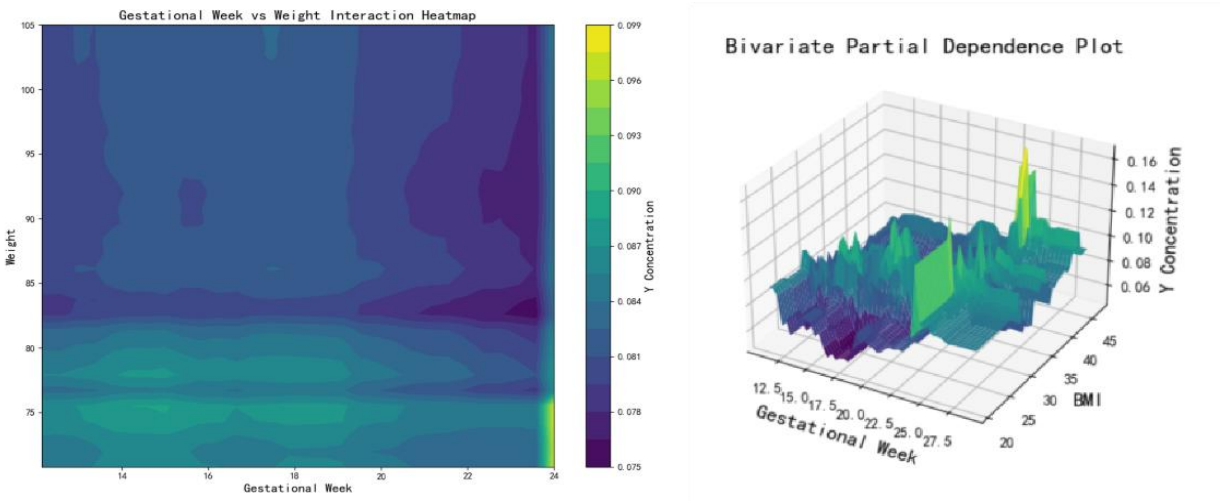

Figure 1. Two-feature factor PDP dependency graph

3.2. Optimization of Testing Time Points Based on BMI Stratification

After applying the natural breakpoint method to K iterations, we ultimately classified the BMI of pregnant women carrying male fetuses into 8 groups, with six groups classified as low-risk and two as high-risk. The high-risk group accounted for only one-quarter of the total. As shown in the BMI grouping table, the BMI intervals strictly met the coverage constraints for group classification. The mean BMI within each group clustered around 34, with the earliest gestational age meeting the criteria concentrated around 12 weeks. The optimal NIPT timing window should be between 11-13 weeks (Table 1).

Table 1. BMI grouping status

Group	BMI block	Number of participants	Group Average BMI	Earliest gestational week to meet the target within the group	Group average gestational week at which target was reached	Optimal NIPT timing (opt)	Risk label
1	[20.703, 26.619)	1	20.703125	13.714286	13.714286	13.714286	High risk
2	[26.619, 29.566)	53	28.715825	11.000000	13.164420	11.000000	Low risk
3	[29.566, 31.142)	68	30.279903	11.000000	12.850840	11.000000	Low risk
4	[31.142, 32.812)	55	31.962248	11.142857	12.719481	11.142857	Low risk
5	[32.812, 34.928)	51	33.642397	11.142857	12.921569	11.142857	Low risk
6	[34.928, 38.408)	25	35.918284	11.285714	13.108571	11.285714	Low risk
7	[38.408, 44.983)	9	39.259695	11.142857	12.412698	11.142857	Low risk
8	[44.983, 46.875]	2	45.928849	12.857143	18.785714	12.857143	High risk

3.3. Multi-factor Synergistic Optimization and Integrated Risk Assessment

The modified K-means algorithm was employed to classify the cohort of pregnant women carrying male fetuses into five equivalent BMI subgroups (range: 20.7–46.9), achieving significant intergroup heterogeneity while ensuring intra-group homogeneity. By coupling population compliance rates with segmented risk functions, the model accurately identified mutation characteristics and optimized inflection points of potential risks with increasing gestational age. The results demonstrated that under the premise of compliance rates exceeding the 80% clinical threshold (peaking at 94.3%) across all subgroups, the optimal NIPT timing for each subgroup was determined to be between 13.5–13.9 weeks (Figure 2).

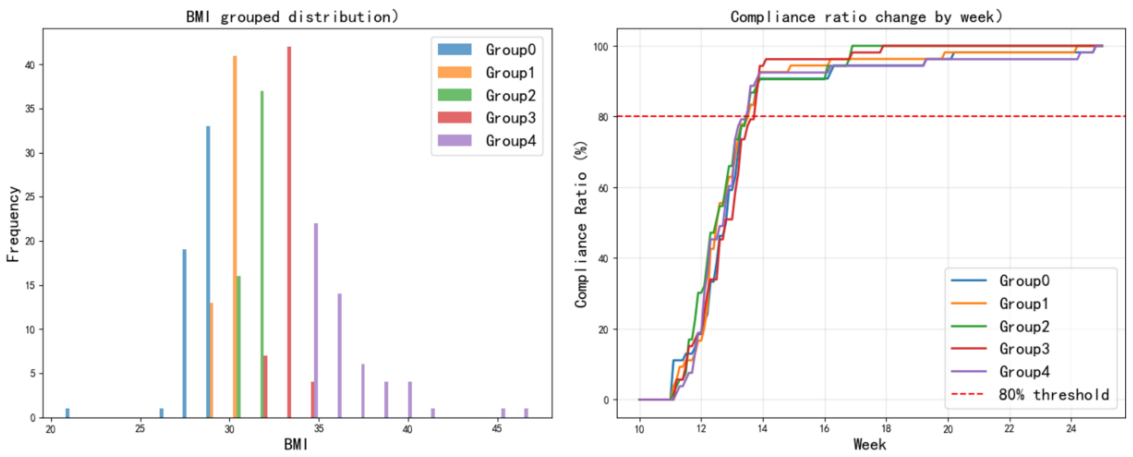

Figure 2. Model result visualization

4. Conclusions

We developed an integrated computational framework centered on multivariate linear regression, Random Forest (RF), and an augmented K-means clustering algorithm. Quantitative analysis elucidated the significant non-linear impacts of gestational age and maternal Body Mass Index (BMI) on cell-free fetal DNA concentrations. By synergizing Jenks natural breaks optimization with Monte Carlo simulations, we identified a personalized screening window ranging from 11.1 to 13.9 weeks and established a refined five-strata BMI classification system, pinpointing the global optimal NIPT timing at approximately 14 weeks. This framework effectively mitigates clinical ambiguity in timing selection and enhances diagnostic precision across heterogeneous maternal profiles.

While the model demonstrates exceptional decision stability, its statistical generalizability is currently constrained by the limited sample size within specific sub-cohorts, particularly those with extreme BMI profiles. Future research will prioritize the integration of multi-center, large-scale clinical datasets for rigorous external validation. Such efforts aim to establish a more comprehensive and clinically efficacious intelligent prenatal screening ecosystem, ultimately advancing the precision and reliability of maternal-fetal diagnostics on a global scale.

References

[1] Van Beek D M, Straver R, Weiss M M, Boon E M J, Huijsdens-van Amsterdam K, Oudejans C B M, Reinders M J T, Sistermans E A. Comparing methods for fetal fraction determination and quality control of NIPT samples: Comparing fetal fraction and quality control tools for NIPT [J]. Prenatal Diagnosis, 2017, 37(8): 769~773.

[2] Hu L, Pei Y, Luo X, Wen L, Xiao H, Liu J, Wu L, Li G, Wei F. A multivariate modeling method for the prediction of low fetal fraction before noninvasive prenatal testing [J]. Science Progress, 2021, 104(4): 00368504211052359.

[3] Ding Y, Ni K, Fan X, Yan Q. A Bayesian Decision-Theoretic Optimization Model for Personalized Timing of Non-Invasive Prenatal Testing Based on Maternal BMI [J]. Mathematics, 2026, 14(3): 437.

[4] Yang M, Fu J, Yang Z. IEEE, 2025. Optimization Method for NIPT Timing in Pregnant Women Carrying Male Fetuses Based on Semi-Supervised Learning and Bayesian Model [C]//2025 5th International Conference on Digital Society and Intelligent Systems (DSInS), 2025Haikou, China: 47~55.

[5] Marton T, Erdélyi Z R, Takai M, Mészáros B, Supák D, Ács N, Kukor Z, Herold Z, Hargitai B, Valent S. Systematic Review of Accuracy Differences in NIPT Methods for Common Aneuploidy Screening [J]. Journal of Clinical Medicine, 2025, 14(8): 2813.

[6] Teng Yuemei. Clustering Optimization and Random Forest Coupling Model for NIPT Clinical Decision-Making [J]. Technology of Smart IoT, 2026, 58(2): 145-149.

[7] Wang C, Hou D, Wang G, Wu X, Zhou Z, Yang L, Liu Y, Wang X. Progress, clinical application and challenges of non-invasive prenatal testing for monogenic diseases [J]. Frontiers in Pediatrics, 2026, 14: 1734842.