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

Research on Fetal Chromosome Abnormality Risk in NIPT Based on Multivariate Nonlinear Modeling

Jianuo Wu, Zhenglin Wang*, Xinyuan Qiao

School of Information and Electrical Engineering, Hebei University of Engineering, Handan 056038, China

* **Corresponding author**: Zhenglin Wang (Email: Kirito_F1@outlook.com)

Abstract: This study addresses key challenges in fetal chromosomal abnormality risk assessment for non-invasive prenatal testing (NIPT) by establishing an integrated analytical framework combining multivariate nonlinear modeling and machine learning. The innovative contributions of this research are highlighted in several aspects: For male fetuses, a novel trivariate cubic polynomial regression model was developed to precisely quantify the complex nonlinear relationships among Y-chromosome concentration, gestational age, and maternal BMI, with a joint optimization algorithm determining personalized testing time windows—overcoming limitations of existing methods in dynamic prediction and timing selection. For female fetuses, an anomaly identification model integrating ADASYN oversampling and decision trees was constructed, effectively addressing data imbalance through multi-dimensional feature selection and significantly enhancing classification performance in the absence of Y-chromosome reference. Overall, the proposed gender-specific modeling strategy provides a more precise and personalized screening solution for clinical practice. Validation on a clinical dataset of 1,056 male and 605 female fetal cases demonstrates that this framework significantly improves risk prediction accuracy and offers an effective tool for optimizing prenatal screening strategies.

Keywords: Chromosome Abnormality Risk, Nonlinear Programming, Multivariate Polynomial Regression, Decision Tree, Prenatal Screening.

1. Introduction

Non-Invasive Prenatal Testing (NIPT) has become a cornerstone technology for preventing birth defects, enabling the screening of common chromosomal aneuploidies (such as trisomy 21, 18, and 13) through the analysis of cell-free fetal DNA (cffDNA) in maternal plasma. However, the accuracy of NIPT is highly dependent on the precise assessment of fetal DNA concentration, which is complexly influenced by various factors including gestational age (GA) and maternal body mass index (BMI). Existing studies generally indicate significant nonlinear relationships between fetal DNA concentration and these maternal physiological parameters, which are difficult to accurately capture with conventional linear models, posing substantial challenges to precise risk assessment [1-2]. Specifically, for male fetuses, while current research has recognized the trend of Y-chromosome concentration varying with GA and BMI, there is a lack of robust mathematical models capable of effectively quantifying multi-factor interactions and dynamically predicting the optimal testing timing based thereon [3]. For female fetuses, the absence of the Y-chromosome as a reference necessitates the integrated analysis of multidimensional indicators such as Z-scores, GC content, and read counts for chromosomal abnormality identification. Furthermore, the scarcity of abnormal samples in clinical data leads to severe class imbalance, significantly complicating the construction of classification models [4-5].

To systematically address these challenges, this paper proposes a hybrid modeling framework that integrates multivariate nonlinear programming with machine learning. Its marginal contributions are primarily manifested in three aspects. First, for male fetuses, this study constructs a trivariate cubic polynomial regression model that systematically quantifies, for the first time, the nonlinear dynamic relationships among Y-chromosome concentration, GA, and maternal BMI. Building upon this, a joint optimization model based on multi-constrained nonlinear programming is established, which employs a fusion of genetic algorithms and differential evolution algorithms to solve for personalized optimal testing time windows tailored to different BMI populations [6-7]. Second, for female fetuses, this study designs an integrated identification method combining the ADASYN oversampling technique with a decision tree model. By developing a multidimensional feature evaluation framework that synthesizes correlation strength, statistical significance, and inter-group consistency, key discriminative indicators are selected, effectively mitigating the data imbalance problem and enhancing the detection efficacy for abnormalities in the absence of a Y-chromosome reference [8]. Through these methodological innovations, this research aims to provide more accurate risk assessment tools and optimized personalized screening strategies for the clinical practice of NIPT.

2. Methodology

This study aims to address two core challenges in NIPT: optimizing the testing time for male fetuses and identifying chromosomal abnormalities in female fetuses. The overall methodological framework is divided into two corresponding parts.

2.1. Model for Male Fetus: Y-Chromosome Concentration Dynamics and Testing Time Optimization

2.1.1. Multivariate Polynomial Regression Model

This study employed a refined body mass index (BMI) grouping strategy, dividing the overall sample into six consecutive intervals. For each BMI interval, a three-variable

cubic polynomial regression model was systematically constructed to quantitatively analyze the nonlinear effects of gestational age, BMI, and their interaction on Y chromosome concentration. The matrix X and parameter vector β are given by Equation 1:

$$Y = X\beta + \epsilon \quad (1)$$

Y is the vector of observed Y-chromosome concentrations, X is the design matrix containing the constant term, linear, quadratic, cubic, and interaction terms of BMI and GA, β is the coefficient vector, and ε is the error term. Ordinary Least Squares (OLS) is employed to estimate the model parameters. The core principle involves minimizing the residual sum of squares (RSS), with the objective function defined as:

$$\hat{\beta} = \text{argmin}_{\beta}(Y - X\beta)^T(Y - X\beta) \quad (2)$$

Quantitative evaluation of the model requires the following key statistical indicators: Residual Standard Error (RSE) measures the average magnitude of prediction errors, while the coefficient of determination (R^2) reflects the proportion of total variation in the dependent variable explained by the model, as shown in Equations 3 and 4:

$$RSE = \sqrt{\frac{\sum e_i^2}{n-p}}, (p = 10) \quad (3)$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \overline{Y})^2} \quad (4)$$

2.1.2. Joint Optimization Model for Detection Time

This study formulates the joint optimization problem of body mass index (BMI) grouping and clinical testing timepoints as a multi-constrained nonlinear programming model. The model aims to simultaneously determine the optimal boundaries for BMI grouping and the corresponding clinical testing schedules through mathematical optimization methods, maximizing efficiency while fully ensuring statistical reliability and clinical feasibility.

The optimization objective is to minimize the weighted average time to first achievement across all groups, with the objective function defined as:

$$\min f(x) = \sum_{i=1}^{k} \left(\frac{n_i}{N}\right) \cdot t_i^e \quad (5)$$

Define k−1 individual body mass index cut-off points $c_1, c_2, \ldots, c_{k-1}$ to divide the total sample into k consecutive intervals. Among them: n_i is the sample size in the i-th group, reflecting the representativeness of this group in the overall population; N is the total sample size; t_i^{earliest} represents the earliest time point when the predicted value of Y chromosome concentration in the i-th group first reaches or exceeds the clinical threshold of 4%, which is calculated by Formula (6).

$$t_i^e = \min_{\hat{}} t \in \{[t_{\min}, t_{\max}] : \hat{Y}_i(t) \geq 0.04\} \quad (6)$$

Among them, $\hat{Y}_i(t)$ is the predicted concentration at time t of the three-variable cubic polynomial model corresponding to the i-th group.

To ensure the scientific rigor of the model, the following systematic constraint framework has been established: Sample size constraints require that the number of observed samples in each subgroup must not fall below the specified threshold. For each subgroup, the estimated probability of achieving the target Y chromosome concentration at the recommended testing time point must not be less than 95%, as shown in Equations 7-9.

$$n_i \geq n_{\min} = 20, \forall i \in \{1,2, \ldots, k\} \quad (7)$$

$$P(Y_i(t_i^{\text{earliest}}) \geq 0.04) \geq 0.95, \forall i \quad (8)$$

$$P(Y_i \geq 0.04) = 1 - \Phi\left(\frac{0.04 - \hat{Y}_i}{\sigma_i}\right) \quad (9)$$

Here, σ_i is the concentration value predicted by the model and the standard deviation of the prediction error for this group.

This paper addresses the requirement for solving joint optimization problems by effectively integrating two classical evolutionary algorithms—the genetic algorithm and the differential evolution algorithm. This fusion scheme aims to fully leverage the synergistic advantages of both algorithms, significantly enhancing the convergence performance and solution accuracy. The specific steps are as follows:

Using real-number encoding, the decision variables are encoded as chromosomes:

$$\text{X} = [c_1, c_2, \ldots, c_{k-1}, t_1, t_2, \ldots, t_k] \quad (10)$$

$$\text{X}_{i,j} = a_{i,j} + (b_j - \text{X}_{r,j}) \quad (11)$$

Among them, c_i is the BMI cut-off point, and t_i is the detection time point. Based on the distribution characteristics of clinical data, reverse initialization is adopted.

Employing a mutation strategy based on the differential evolution algorithm, generate mutated individuals from the target individuals:

$$\text{V}_i = \text{X}_{r1} + F \cdot (\text{X}_{r2} - \text{X}_{r3}) + F \cdot (\text{X}_{r4} - \text{X}_{r5}) \quad (12)$$

Among them, X_{r1} to X_{r5} are distinct random individuals in the population, and F is the scaling factor.

Retain variant or original individual genes based on cross probability CR:

$$U_{i,j} = \begin{cases} V_{i,j} & \text{if rand}(0,1) \leq CR \\ X_{i,j} & \text{otherwise} \end{cases} \quad (13)$$

The penalty function method is employed to address medical constraints, with the fitness function defined as:

$$\text{Fitness(X)} = \sum_{i=1}^{k} \left(\frac{n_i}{N}\right) \cdot t_i + \lambda \cdot \sum_j \max\left(0, g_j(\text{X})\right)^2 \quad (14)$$

2.2. Model for Female Fetus: Abnormality Identification

2.2.1. Feature Selection Model

The dataset comprises 17 indicators. To systematically evaluate the discriminative power of each indicator, a multidimensional assessment framework was constructed to comprehensively quantify indicator importance across three dimensions: correlation strength, statistical significance, and intergroup consistency. Specifically, Spearman's correlation coefficient was used to calculate correlations, statistical significance was assessed, and the importance of each indicator was then synthesized as follows:

Mean Absolute Correlation (MAC): Reflects the overall strength of association between the indicator and abnormal outcomes:

$$\text{MAC}_i = \frac{1}{m} \sum_{j=1}^{m} |r_{ij}| \quad (15)$$

Among them, r_{ij} represents the Spearman correlation coefficient of the i-th index in the j-th valid group, and m is the number of valid groups.

Significantly Correlated Groups (SCG): The frequency with which a metric shows significant correlation across multiple groups:

$$\text{SCG}_i = \sum_{j=1}^{m} \mathbb{I}(p_{ij} < 0.05) \quad (16)$$

Among them, $\mathbb{I}$ is the indicator function, and p_{ij} is the p-value of the corresponding correlation coefficient.

High Correlation Group Count (HCG): Consistency of the evaluation metric exceeding the correlation threshold (0.15) across different groups:

$$\mathrm{HCG}_i = \sum_{j=1}^{m} \mathbb{I}(|r_{ij}| > 0.15) \quad (17)$$

Integrating information from the above three dimensions, emphasizing the dominant role of relevance strength while balancing distinctiveness and consistency:

$$S_{\mathrm{score}_i} = 0.5 \cdot \mathrm{MAC}_i + 0.25 \cdot \frac{\mathrm{SCG}_i}{m} + 0.25 \cdot \frac{\mathrm{HCG}_i}{m} \quad (18)$$

2.2.2. Decision Tree Classification Model

Based on the top five key indicators selected through preliminary correlation analysis, a decision tree model for identifying female fetal abnormalities was constructed. The Gini index was employed as the evaluation criterion for node splitting. The Gini index measures the purity of node samples, where a smaller value indicates greater concentration of sample categories and higher purity, expressed as follows:

$$Gini(D) = 1 - \sum(p_k)^2 \quad (19)$$

pk is the proportion of class k in node D.

The effectiveness of the decision tree model in identifying female fetal abnormalities is measured using the following metrics: accuracy, precision, and recall.

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (20)$$

$$\mathrm{Precision}_1 = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (21)$$

$$\mathrm{Precision}_0 = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \quad (22)$$

$$\mathrm{Recall}_1 = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (23)$$

$$\mathrm{Recall}_0 = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (24)$$

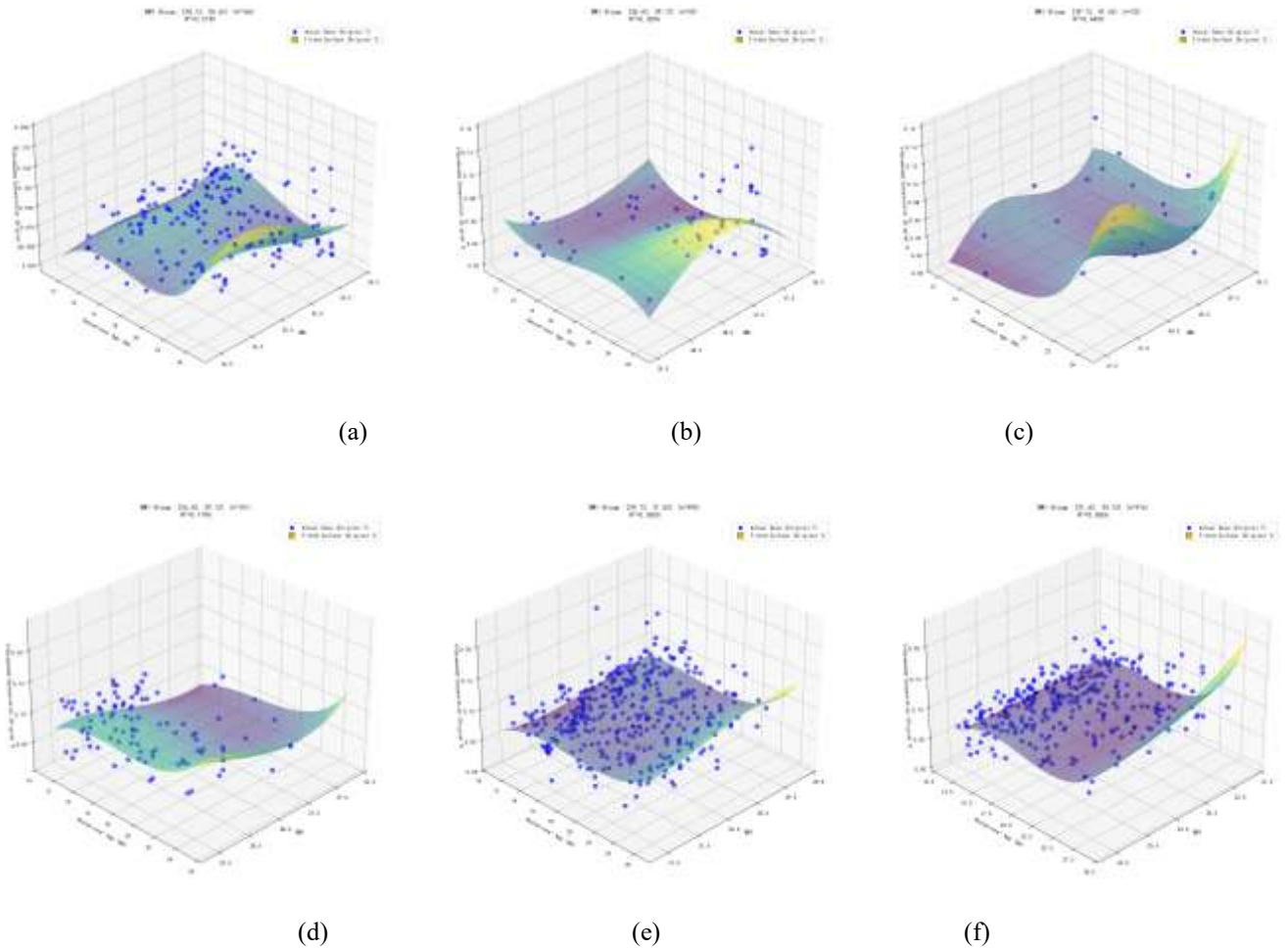
TP refers to the number of samples that are actually in the positive class and are correctly predicted as positive by the model, TN refers to the number of samples that are actually in the negative class and are correctly predicted as negative by the model, FP refers to the number of samples that are actually in the negative class but are incorrectly predicted as positive by the model, FN refers to the number of samples that are actually in the positive class but are incorrectly predicted as negative by the model.

3. Results

Based on clinical data from 1,056 male fetal samples and 605 female fetal cases, we processed the dataset using Python and programmed simulations of the aforementioned model.

3.1. Regression and Optimization Results for Male Fetal Models

This study employed a three-dimensional surface fitting method to visualize the regression models across each BMI group, yielding a more intuitive relationship map. As shown in Figure 1:



(a) (b) (c)

(d) (e) (f)

Figure 1. Y chromosome concentration vs. gestational age and BMI fit plot

As shown in Figure 1, the refined grouping strategy demonstrates outstanding model performance: as BMI intervals increase, the model's interpretability exhibits a systematic and significant enhancement, with prediction accuracy simultaneously optimized to a high degree. The highest BMI interval achieves an extremely low error rate, fully validating this modeling approach's exceptional adaptability and precise predictive capability for high-BMI populations. This provides a reliable basis for conducting personalized clinical testing.

By solving the joint optimization model using an improved genetic algorithm with differential evolution, the corresponding optimal BMI groupings and optimal detection timepoints can be obtained, as shown in Table 1 below.

Table 1. Optimal BMI grouping and testing timepoint

Optimal BMI Grouping	Optimal Testing Timepoint
Group 1: [26.62,30.62)	Group 1: 16.0 weeks
Group 2: [30.62,34.62)	Group 2: 19.0 weeks
Group 3: [34.62,38.62)	Group 3: 16.5 weeks
Group 4:[38.12,42.62)	Group 4: 18.0 weeks

Based on the results, pregnant women were categorized into four BMI groups, each corresponding to an optimal gestational week for testing. For instance, women with a BMI between 30.62 and 34.62 are recommended to undergo testing at 19 weeks, while other groups are advised to test around 16 to 18 weeks. This grouping scheme was not arbitrarily determined but derived through an optimization algorithm. It aims to ensure that the Y chromosome concentration in each group is most likely to reach the effective detection threshold of 4% or higher at the recommended time point.

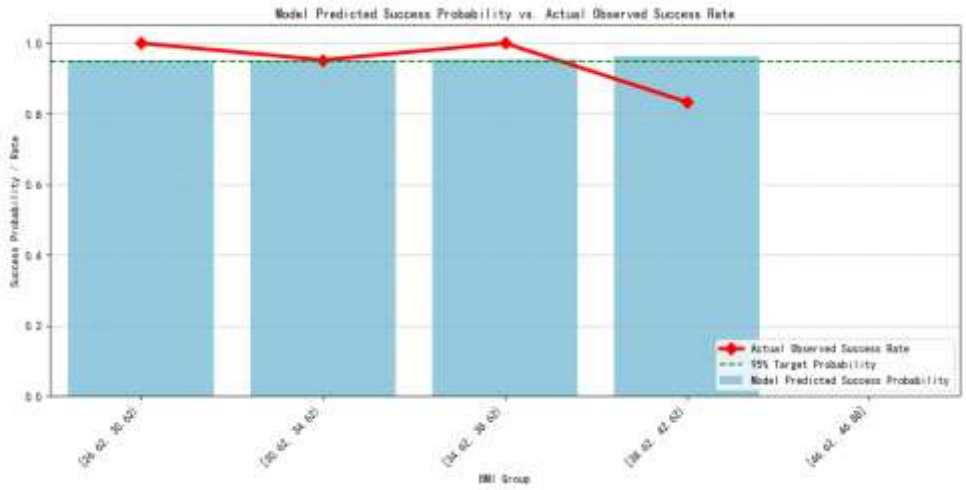


Figure 2. Comparison chart of predicted compliance rate vs. actual predicted compliance rate

Figure 2 provides compelling validation of the reliability of the aforementioned recommendation scheme. The high degree of overlap between the two curves in the figure demonstrates that the model delivers highly accurate predictions of actual clinical achievement of target outcomes. This confirms that the personalized grouping strategy not only effectively distinguishes optimal testing timepoints across different populations but also exhibits an exceptionally high success rate in real-world settings. It thus provides a scientifically sound and credible basis for decision-making in clinical practice.

3.2. Anomaly Detection in Female Fetal Models Results

Simulating the feature selection model in Section 2.2.1 yields a comprehensive score based on 17 indicators, as illustrated in Figure 3.

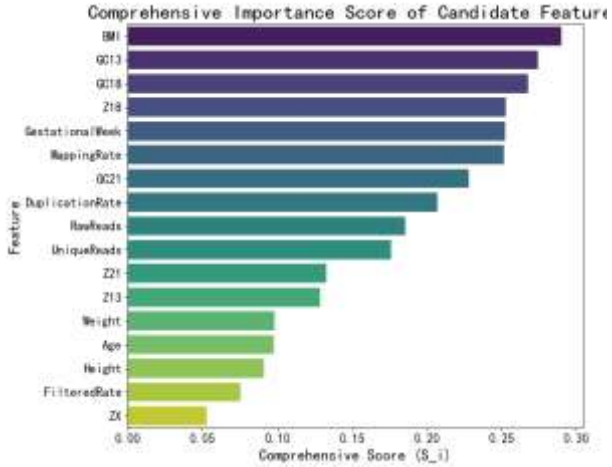


Figure 3. Composite importance score chart for each candidate indicator

A comprehensive analysis of this figure indicates that the importance of candidate indicators varies significantly when constructing a model for detecting chromosomal abnormalities in female fetuses. The top five key indicators ranked by composite score are: BMI (Body Mass Index), GC content of chromosome 13 (GC13), GC content of chromosome 18 (GC18), Z-score of chromosome 18 (Z18), and gestational week. These five indicators demonstrate the strongest discriminative power and stability in distinguishing chromosomal abnormalities in female fetuses. Consequently, they were selected as core features for subsequent decision tree model construction to ensure the diagnostic method is both effective and clinically interpretable.

The top five indicators by score were selected to train the decision tree classification model. Integrating the various indicators from Section 2.2.2, the specific results are shown in Figure 4.

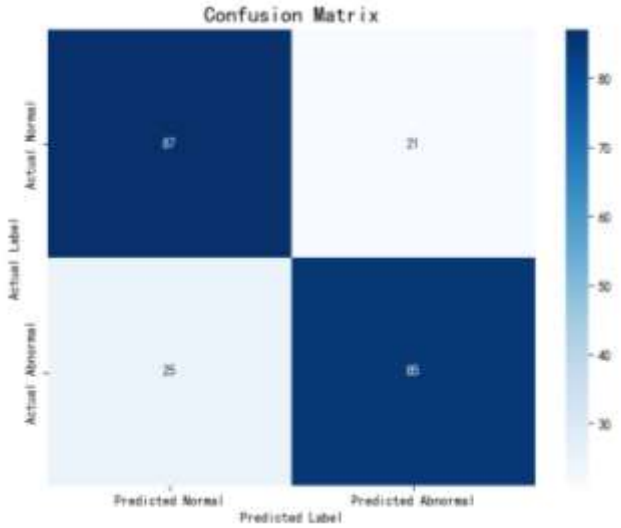


Figure 4. Confusion Matrix Chart

The model's confusion matrix results indicate that it correctly identified 85 abnormal fetuses and 87 normal fetuses. However, there were 21 cases where normal fetuses were incorrectly classified as abnormal and 25 cases where abnormal fetuses were incorrectly classified as normal. The final calculations yielded an overall accuracy rate of 78.90%. For the abnormal category, the precision rate was 80.19% and the recall rate was 77.27%. For the normal category, the precision rate was 78.36% and the recall rate was 80.56%. These results indicate that the model demonstrates high predictive accuracy overall.

Analysis of the ROC curve for the decision tree model is shown in Figure 5 below.

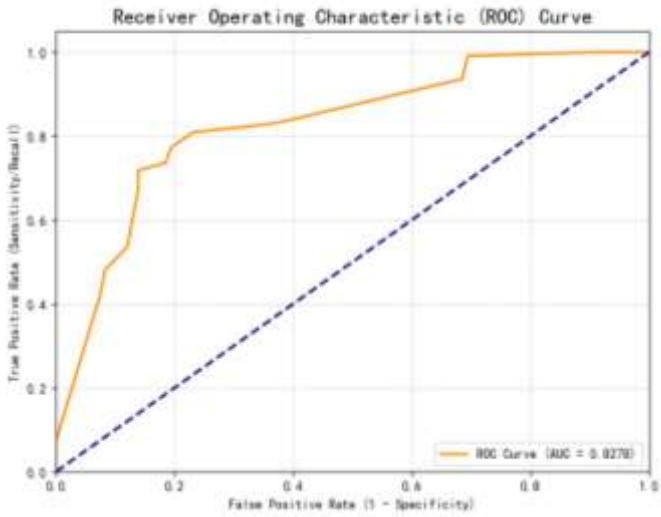


Figure 5. Receiver Operating Characteristic (ROC) Curve

The ROC curve is clearly positioned above the random guessing line and curves upward to the left, indicating the model possesses strong discrimination capabilities. The area under the curve is calculated to be approximately 0.8278, further confirming the model's excellent performance. The model achieves a high true positive rate in regions with low false positive rates, which is highly advantageous for clinical scenarios requiring strict control of false alarms.

4. Conclusions

This study establishes a gender-specific analytical

framework to address key challenges in fetal chromosomal abnormality risk assessment for NIPT. Based on clinical data from 1,056 male and 605 female fetal cases, we developed comprehensive modeling solutions for both gender groups. For male fetuses, we created a BMI-stratified trivariate cubic polynomial regression model and employed an optimized algorithm to determine personalized testing time windows. For female fetuses, we developed a feature selection system and constructed a decision tree classification model using ADASYN oversampling to handle data imbalance. Empirical results demonstrate that our framework significantly improves prediction accuracy and supports targeted clinical decision-making.

The proposed framework shows strong potential for clinical implementation, as it utilizes routinely available clinical parameters without requiring additional testing costs. The models maintain moderate computational complexity suitable for integration into existing NIPT analysis systems. Future work will focus on incorporating additional biomarkers, exploring advanced deep learning architectures, and validating the framework through multi-center clinical studies. These efforts will further enhance the model's robustness and promote the development of more precise, personalized prenatal screening solutions.

References

[1] Yu Dandan, Li Fengjin, Yao Xinyu, et al. Efficacy analysis of NIPT in detecting common fetal chromosomal aneuploidy abnormalities in different cautious populations [J]. Chinese Journal of Family Planning and Gynecotokology, 2025, (03): 72-76+82.

[2] Zhang Liangliang, Zhuo Zhaozhen, Huang Shengwen, et al. Comparative analysis of the application value of NIPT and NIPT-plus in prenatal screening [J]. Guizhou Medical Journal, 2025, (06): 942-944.

[3] Zheng Haiyan, Wu Yanhong, Zhang Xin, et al. Significance of NIPT in prenatal screening for chromosomal abnormalities and classification of high-risk Z-values [J]. Journal of Reproductive Medicine, 2023, (12): 1843-1848.

[4] Chen Qiang. Improvement and positioning application of differential evolution algorithm [J]. Beidou and Spatial Information Application Technology, 2025, (03): 28-31.

[5] Chen Xiang, Yang Zhiqiang, Yang Bing, et al. Application of LOESS in noise identification and extraction of GNSS vertical coordinate time series [J]. Journal of Geodesy and Geodynamics, 2021, (10): 1024-1029.

[6] Hu L, Pei Y, Luo X, et al. A multivariate modeling method for the prediction of low fetal fraction before noninvasive prenatal testing[J]. Science Progress, 2021, 104(4): 00368504211052359.

[7] Zhou Y, Zhu Z, Gao Y, et al. Effects of maternal and fetal characteristics on cell-free fetal DNA fraction in maternal plasma[J]. Reproductive Sciences, 2015, 22(11): 1429-1435.

[8] Becking E C, Scheffer P G, Henrichs J, et al. Fetal fraction of cell-free DNA in noninvasive prenatal testing and adverse pregnancy outcomes: a nationwide retrospective cohort study of 56,110 pregnant women[J]. American Journal of Obstetrics and Gynecology, 2024, 231(2): 244. e1-244. e18.