A Hybrid Modeling Framework for NIPT: Integrating Discrete-Time Survival Analysis and XGBoost for Enhanced Prenatal Screening

Xiumei Liu[1, *, #], Kai Zheng[2, #], Zhibo Cui[3, #]

[1] School of Science, Yanshan University, Qinhuangdao 066004, China

[2] Electrical Engineering, Yanshan University, Qinhuangdao 066004, China

[3] School of Science, Yanshan University, Qinhuangdao 066004, China

[*] Corresponding author: 1791629046@stumail.ysu.edu.cn

[#] These authors contributed equally.

Abstract

This study combines discrete-time survival analysis regression trees with machine learning algorithms, specifically logistic regression and XGBoost, to develop an integrated predictive framework aimed at optimizing the clinical application of Non-Invasive Prenatal Testing (NIPT). The research demonstrates that this hybrid approach effectively models the time-to-threshold progression of Y-chromosome concentration in male fetuses and improves the classification accuracy of fetal abnormalities in female fetuses by leveraging key biomarkers such as chromosome Z-scores and maternal physiological parameters. The incorporation of SHAP values enhances the interpretability of the model, providing transparent explanations for predictions. Sensitivity analysis confirms the robustness of the model under varying parameter settings, supporting its potential for personalized prenatal screening protocols and improved clinical decision-making.

Keywords

Non-Invasive Prenatal Testing (NIPT), discrete-time survival analysis, XGBoost, SHAP values, fetal abnormality classification, predictive modeling.

1. Introduction

Non-Invasive Prenatal Testing (NIPT) has established itself as a cornerstone of modern prenatal care, fundamentally shifting the paradigm for screening fetal chromosomal abnormalities. By analyzing cell-free DNA (cfDNA) in maternal plasma, NIPT offers a highly sensitive and specific method for detecting common aneuploidies, thereby reducing the need for invasive diagnostic procedures [1]. The reliability of this test is critically dependent on the fetal fraction, a key performance parameter which has been demonstrated to be significantly influenced by factors such as gestational age and maternal weight [2]. Further research into the determinants of fetal fraction at 11-13 weeks' gestation confirmed that a range of maternal and fetal characteristics can impact its level, underscoring the necessity of considering these variables for accurate test interpretation and clinical counseling [3]. The evolution of NIPT has also facilitated its integration into comprehensive first-trimester screening protocols. Combining NIPT with a fetal anomaly scan has been proposed as a patient-friendly and efficient integrated screening model, enhancing early detection rates while maintaining a focus on patient safety and choice [4].

The clinical scope of NIPT has rapidly expanded beyond the initial focus on trisomies 21, 18, and 13. So-called "expanded NIPT" now encompasses the detection of fetal sex chromosome aneuploidies [11] and chromosomal microdeletion/microduplication syndromes [8]. Large-scale clinical studies, such as one involving 5,696 pregnant women, have provided robust evidence for the performance of expanded NIPT in a real-world setting, demonstrating its effectiveness in identifying a broad spectrum of chromosomal imbalances [10]. This utility extends to more complex pregnancy scenarios, including twin pregnancies, where research has validated its application for screening chromosomal abnormalities, thereby addressing a significant clinical need [12]. The growing adoption of NIPT has also prompted investigations into patient perspectives, with studies exploring pregnant women's awareness and the factors influencing their choice to undergo testing, which is crucial for understanding its implementation in clinical practice [9]. Moreover, the potential applications of NIPT continue to broaden, with research indicating its value in diagnosing obstetric maternal complications and comorbidities, suggesting a future role beyond traditional genetic screening [1]. A substantial body of literature, including meta-analyses and large cohort studies, continues to affirm the high performance and clinical effectiveness of NIPT in prenatal screening across diverse populations [13].

The advancement of complex data analysis in NIPT parallels innovations in predictive modeling and machine learning seen in other medical fields. For instance, the construction of survival analysis models and nomograms for AIDS patients with related diseases illustrates the power of sophisticated statistical tools in prognosticating complex medical outcomes [5]. In the realm of data handling, the development of specialized models like the Random Effects-Expectation Maximization Regression Tree has shown significant promise in dealing with the intricate, hierarchical structure often inherent in medical system data [6]. Furthermore, cutting-edge approaches that integrate multiple techniques, such as k-shape clustering combined with a TCN-Attention-XGBoost framework for baseline load estimation, showcase the potential of hybrid machine learning models to achieve high predictive accuracy in complex, time-series-related applications [7]. These methodological advances provide a valuable conceptual and technical toolkit that can be leveraged to optimize the interpretation of NIPT data, improve risk stratification, and ultimately enhance predictive models in prenatal diagnostics.

The main contributions of this study can be summarized as: 1) developing a comprehensive, integrated prediction model that synergizes time series analysis with advanced machine learning algorithms to significantly improve the accuracy of interpreting NIPT results and forecasting outcomes; 2) conducting a multifaceted analysis that not only evaluates the impact of key clinical variables-such as fetal fraction determinants, maternal demographics, and specific clinical indications-on test performance but also rigorously optimizes the model architecture to mitigate overfitting and enhance generalizability, drawing on methodologies from related fields; and 3) establishing a robust, data-driven framework for clinical decision-making that provides actionable insights for optimizing the application of both standard and expanded NIPT in diverse patient populations, including singletons and twins, and for a variety of conditions, from common aneuploidies to microdeletions, while also considering patient choice factors.

The structure of this paper is as follows: the first part is the introduction, which presents the research background, the current status and expanded applications of NIPT, and key contributions of the study; the second part is the methodology, detailing the data collection process, feature engineering based on established and novel clinical factors, and the implementation of the hybrid prediction model; the third part focuses on results and analysis, showcasing the model's performance metrics, validation results against clinical outcomes, and comparative evaluations with existing screening methods; the fourth part discusses the clinical implications and findings, reviewing the insights gained for different patient subgroups and the

significance of various factors in optimizing NIPT clinical pathways; and the fifth part concludes the paper, emphasizing the overall contributions and proposing future directions for research in NIPT data analysis and clinical integration.

2. Related Theories

Discrete-time survival-analysis regression trees represent a machine learning approach that extends traditional survival analysis to handle censored data within a tree-based framework. This method segments the population into distinct subgroups based on covariate patterns that influence the hazard of an event occurring within discrete time intervals. Unlike conventional survival models that assume proportional hazards, this technique non-parametrically captures complex, non-linear interactions between predictors and the time-to-event outcome. It operates by recursively partitioning the data to maximize the difference in survival experiences between nodes, thus identifying homogeneous risk profiles. This makes it particularly valuable for uncovering specific subpopulations with unique risk trajectories without imposing strict functional form assumptions on the baseline hazard.

XGBoost is a highly efficient and scalable implementation of gradient boosted decision trees, designed for speed and performance. It builds an ensemble model by sequentially adding decision trees, where each new tree corrects the errors made by the previous ones. A key feature is its regularization term within the objective function, which helps control model complexity and reduces overfitting. XGBoost can handle missing data natively and supports parallel processing, making it suitable for large-scale datasets. Its flexibility allows it to be applied to various prediction tasks, including classification and regression, often achieving state-of-the-art results in many machine learning competitions and practical applications.

SHAP is a unified framework for interpreting the output of any machine learning model based on cooperative game theory. It assigns each feature an importance value for a particular prediction, representing the marginal contribution of that feature to the model's output compared to the average prediction. The method satisfies desirable properties like local accuracy and consistency, ensuring that explanations are both faithful to the model and intuitively understandable. By providing both global interpretability (feature importance across the dataset) and local interpretability (reasoning for individual predictions), SHAP helps bridge the gap between complex model performance and human understanding, making black-box models more transparent and trustworthy.

3. Experiments

3.1. Discrete-time Survival-analysis Regression Trees

Survival analysis is capable of simultaneously characterizing the time to event and the outcome, while effectively handling censored data. In this study, "the first time Y-chromosome concentration reaches 4%" is defined as the target event, with "gestational week" as the time variable. Covariates such as BMI are introduced into the model to delineate the probability of reaching the threshold across gestational weeks and to compare differences in the time-to-event distribution among pregnant women with different characteristics [5]. The specific process is as follows:

$$y_{ij} = \begin{cases} 1, t_{ij} = T_i \\ 0, else \end{cases} \tag{1}$$

Equation (1) indicates that each pregnant woman is recorded as an independent "individual–gestational week" unit at each testing time point. If an individual is censored, all subsequent observations are marked accordingly, resulting in a binary response sequence corresponding to the covariate vector.

The outcome event is defined as the first time the Y-chromosome concentration reaches 4%, and the time variable is the gestational week when this threshold is first attained. Since each NIPT test for a pregnant woman occurs at discrete gestational time points rather than continuously, a discrete-time survival analysis approach is adopted. A model is constructed using the complementary log-log (cloglog) link function, with the linear predictor expressed as:

$$\eta_{ij} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_{ij} \tag{2}$$

Here, beta represents the coefficients to be estimated, reflecting the direction and magnitude of the influence of each covariate on the probability of reaching the threshold. The conditional probability of first reaching the threshold at gestational week w is:

$$h_{ij} = 1 - \exp\{-\exp(\eta_{ij})\}, 0 < h_{ij} < 1 \tag{3}$$

Based on this, the survival probability (probability of not reaching the threshold before week w) and the cumulative probability of reaching the threshold for an individual are obtained as:

$$\begin{aligned} S_i(w) &= \prod_{t_{ij} \le w} (1 - h_{ij}) \\ P_i(w) &= 1 - S_i(w) = 1 - \prod_{t_{ij} \le w} (1 - h_{ij}) \end{aligned} \tag{4}$$

For parameter estimation of beta, the generalized linear model approach is used to maximize the likelihood function, which is formulated as:

$$\ell(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} [y_{ij} \log h_{ij} + (1 - y_{ij}) \log(1 - h_{ij})] \tag{5}$$

All continuous variables were standardized before model fitting to ensure numerical stability.

3.2. XGBoost and SHAP

When building a model for classifying abnormal female fetuses, considering that abnormality is a binary variable, logistic regression is first employed as a baseline model for preliminary judgment. Here, let the sample feature vector be denoted as $\mathbf{X} = (x_1, x_2, \cdots, x_p)$, and the label $y \in \{0, 1\}$ indicate whether the fetus is abnormal. The logistic regression model is as follows:

$$P(y = 1 \mid x) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)\}} \tag{6}$$

To avoid multicollinearity, an L2 regularization term is introduced into the loss function, resulting in the penalized likelihood function:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} [y_i \log P(y_i|x_i) + (1 - y_i) \log (1 - P(y_i|x_i))] - \lambda \|\beta\|_2^2 \quad (7)$$

where λ represents the regularization strength.

The model training adopts 10-fold cross-validation, with the maternal code as the grouping factor to ensure that records from the same pregnant woman do not appear simultaneously in both the test and training sets. Additionally, HC3 robust standard errors are used in parameter significance tests to ensure more reliable confidence intervals and p-values under small-sample conditions, thereby improving the model's robustness.

Since logistic regression can only provide the probability of abnormality, the XGBoost model is further introduced to enhance nonlinear modeling capabilities and identify specific conditions under abnormality. Its objective function consists of two parts:

$$\mathcal{O}bj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (8)$$

where l is the logistic loss function, and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ is the regularization term used to control the complexity of the trees. The model iteratively optimizes a series of CART base learners to approximate the negative gradient direction of the objective function.

Based on the coefficient values of the variables obtained from logistic regression, the importance ranking of indicators influencing female fetal abnormality can be determined. Therefore, to reduce the complexity of the XGBoost model and improve interpretability, the top eight features with the largest absolute coefficient values from the logistic regression are selected as inputs for XGBoost in subsequent analysis. After the final model is trained, feature importance ranking and SHAP values are used to interpret and evaluate the results, revealing the relationship between feature values and prediction outcomes [7].

Here, SHAP is based on the Shapley value from game theory, which allocates the contribution of each feature to the predicted output. A higher score indicates greater importance. For a given prediction function , the SHAP value for a sample is defined as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left(f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S)\right) \quad (9)$$

where F represents the entire set of features, and S is a subset of features. SHAP values measure the marginal contribution of each feature to the model's prediction at the sample level, with the sign indicating the direction of increasing or decreasing the probability of abnormality.

4. Results and Analysis

4.1. Results and Analysis of Discrete-time Survival-analysis

After obtaining the time-to-threshold prediction model, a regression tree model was further constructed based on the four main influencing factors (BMI, height, weight, age) of pregnant women carrying male fetuses to group the pregnant women [6]. Within each group, individual probability curves of reaching the threshold were first calculated, and then the average probability curve for the group was derived by taking the mean within the group.

The average probability curves for each group, obtained using Python, are shown in Figure 1. The significant differences in the shapes of the curves indicate that the regression tree effectively distinguishes the Y chromosome threshold progression among pregnant women with different characteristics. Groups 1–3 reached the threshold rapidly in the early stages, while Group 4 exhibited a relatively delayed progression.

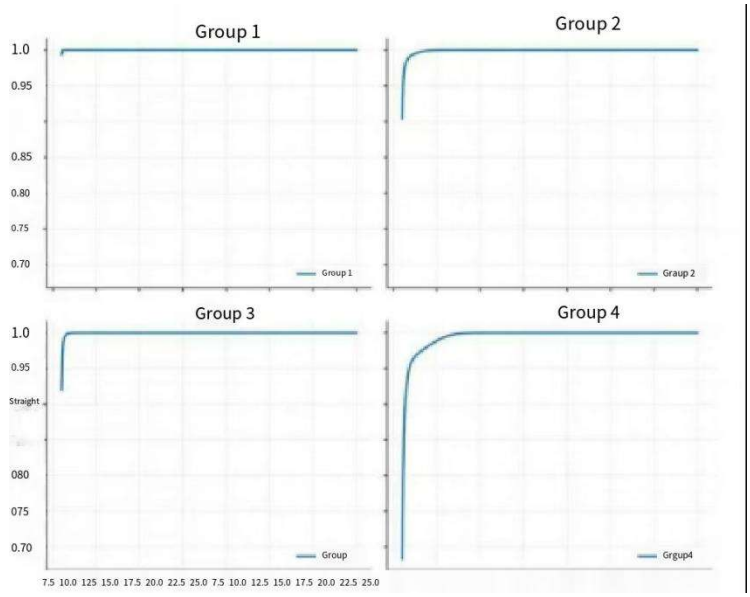

Figure 1. Average Probability Curves of Reaching the Threshold for Each Group

Finally, basic statistical characteristics of each group were calculated using Python and are presented in Table 1.

Table 1. Basic Statistical Characteristics of Each Group

Group Number	Average Gestational Week of First Reaching Threshold	Earliest Gestational Week with Y≥4%	Latest Gestational Week with Y≥4%	Number of Pregnant Women
1	14.2285708	11	21.28571	25
2	13.21953181	11	24.71429	177
3	14.22857175	11	24.28571	40
4	16.3371428	11.14286	24	25

As shown in Table 1, Group 1 and Group 3 have similar average gestational weeks for reaching the threshold. However, Group 1 exhibits a more concentrated time to threshold, indicating a more stable progression of Y chromosome concentration reaching the threshold within the

group. In contrast, Group 3 shows a larger time span, reflecting greater variability. Group 2 has a relatively early average time to threshold but a large time span, suggesting significant individual differences among the pregnant women in this group. Group 4 has the latest first threshold time, indicating an overall delayed progression. These group differences demonstrate the effectiveness of the tree model in distinguishing the patterns of Y chromosome concentration reaching the threshold among pregnant women with different characteristics, providing a basis for further determining the optimal timing for NIPT.

Based on our model, we conducted a sensitivity analysis to examine the robustness of the identified optimal NIPT timing with respect to the weighting of the loss function. Specifically, we perturbed the weights for the risks of missed detection, false positives, and timing cost, respectively, and compared the fluctuations in the optimal NIPT timing across different groups. The results in Table 2 show that, within the defined range of weight variations, the optimal NIPT timing for all groups remained largely unchanged, indicating that the model is robust and reliable with respect to the weight settings.

Table 2. Sensitivity Analysis Results under Different Parameter Perturbations

Parameter Settings	Group	Optimal Gestational Week (weeks)
λ_1、λ_2、λ_3=0.05,0.05,0.9	1	13
	2	14
	3	13
	4	14
λ_1、λ_2、λ_3=0.1,0.05,0.85	1	13
	2	14
	3	13
	4	14
λ_1、λ_2、λ_3=0.15,0.35,0.5	1	13
	2	14
	3	13
	4	14

4.2. Results and Analysis of XGBoost and SHAP

The logistic regression function fitted in this study is as follows:

$$P(y=1|\mathbf{X})=\frac{1}{1+\exp\{-(-2.421715-0.505995x_1+\cdots+0.024200x_p)\}} \tag{10}$$

As shown in Figure 2, Chromosome 18 Z-score, unique ratio, raw read count, and uniquely mapped read count are core features for predicting fetal abnormalities. An increase in these values is significantly positively correlated with the risk of fetal abnormalities. Maternal physiological state variables, on the other hand, have an indirect impact on the prediction results. SHAP analysis further clarifies the direction and magnitude of the influence of each key feature on the risk assessment of abnormalities, providing interpretable support for chromosome abnormality prediction.

The impact of features on the determination of female fetal abnormalities, as calculated using Python, is shown in Table 3. The top five features in terms of importance are: X chromosome Z-score, maternal BMI, gestational week at testing, Chromosome 18 Z-score, and duplicate read ratio.

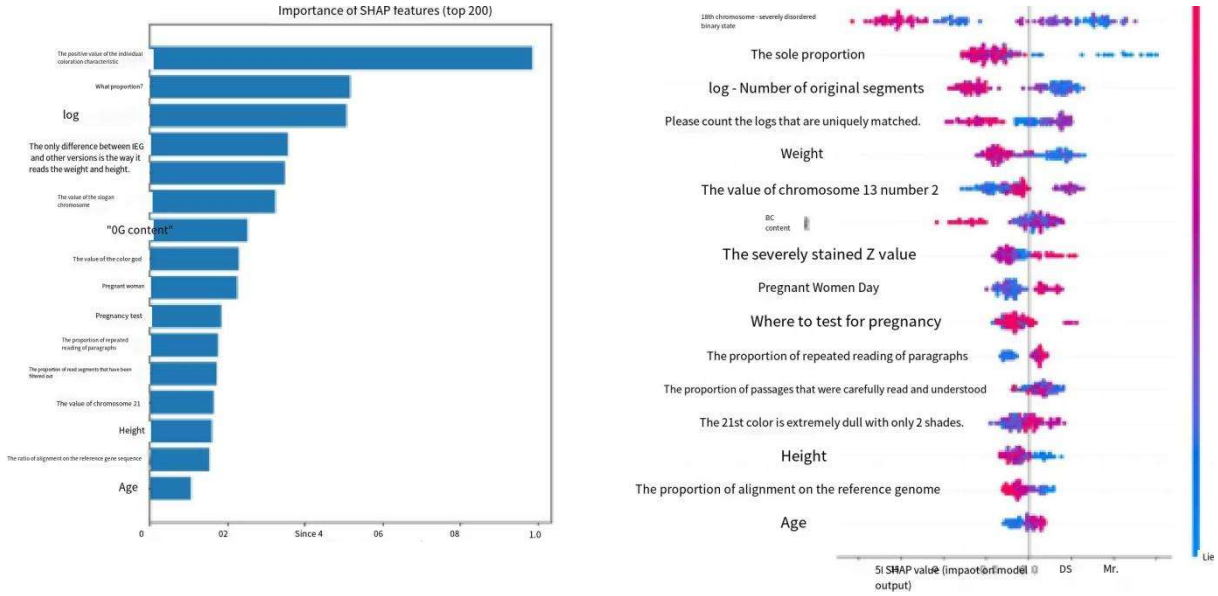

Figure 2. Feature Importance of the XGBoost Model for Fetal Abnormalities (Left) and Corresponding SHAP Value Distributions (Right)

Table 3. Importance Evaluation Values of Each Indicator

Indicator Name	Importance Value	Indicator Name	Importance Value
Maternal BMI	0.104649	Heigh	0.055986
X Chromosome Z-score	0.103187	Age	0.053360
Chromosome 18 Z-score	0.074192	GC Content	0.053253
Duplicate Read Ratio	0.072089	Unique Ratio	0.052742
Mapping Ratio to Reference Genome	0.066980	log(Raw Read Count)	0.050388
Filtered Read Ratio	0.062417	Weight	0.044150
Gestational Week at Testing	0.061547	Chromosome 21 Z-score	0.043112
Chromosome 13 Z-score	0.058960	log(Uniquely Mapped Read Count)	0.042987

This study employs the XGBoost model to determine abnormalities in female fetuses. Leveraging the model's strong capability to capture nonlinear relationships and feature interactions, the trained XGBoost model takes the X chromosome Z-score, maternal BMI, gestational week at testing, Chromosome 18 Z-score, and duplicate read ratio as core input features. Through weighted integration, it outputs the probability of female fetal abnormalities. Based on the optimized threshold or clinical requirements, the final abnormality determination is made, yielding the importance evaluation values.

After reducing the maternal BMI value by 30% and re-inputting it into the model from research question four, the top five features in terms of importance for determining female fetal abnormalities remain the X chromosome Z-score, maternal BMI, gestational week at testing, Chromosome 18 Z-score, and duplicate read ratio. This indicates that the model possesses a certain degree of robustness, as the determination results do not change significantly when the maternal BMI value undergoes minor perturbations.

5. Conclusion

This study combines discrete-time survival analysis regression trees with advanced machine learning methodologies, including logistic regression and XGBoost, to establish a comprehensive predictive framework for optimizing the clinical application of Non-Invasive Prenatal Testing (NIPT). The integrated approach effectively captures the temporal progression patterns of Y-chromosome concentration thresholds in male fetuses while simultaneously addressing the complex challenge of fetal abnormality classification in female fetuses. By leveraging the distinct advantages of each analytical technique-particularly the survival model's capacity to handle time-to-event data with censoring and XGBoost's superior ability to capture nonlinear relationships and feature interactions-the framework provides robust solutions for determining optimal NIPT timing across different maternal subgroups and identifying key biomarkers for fetal abnormality prediction. The incorporation of SHAP values further enhances clinical interpretability, enabling transparent explanation of model predictions at both individual and population levels. Validation through rigorous sensitivity analysis confirms the model's stability under parameter perturbations, demonstrating its potential for reliable implementation in diverse clinical settings to support personalized prenatal screening protocols and improve patient outcomes.

References

[1] Shi Weihui, Xu Chenming. The Application Value of Non-Invasive Prenatal Testing in the Diagnosis of Maternal Complications and Comorbidities in Obstetrics. Journal of Practical Obstetrics and Gynecology, 2025, 41(8): 617-624.

[2] Wang E, Batey A, Struble C, et al. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. Prenatal Diagnosis , 2013, 33(7): 662-666.

[3] Ashoor, G., Poon, L., Syngelaki, A., Mosimann, B., & Nicolaides, K. H. Fetal fraction in maternal plasma cell-free DNA at 11-13 weeks' gestation: effect of maternal and fetal factors. Fetal Diagnosis and Therapy , 2012, 31(4): 237–242.

[4] Srebniak, M. I., Knapen, M. F. C. M., & Joosten, M. Patient-friendly integrated first trimester screening by NIPT and fetal anomaly scan. Molecular Cytogenetics , 2021, 14(1): 3.

[5] Chen Yiwei, Pan Kejun, Maimaitiaili Wubuli, et al. Survival analysis of AIDS patients with AIDS-related diseases or symptoms and construction of related influencing factors and a nomogram prediction model. Journal of Tropical Medicine , 2025, 25(03): 301-307+330+284.

[6] Li Weinan. Application of Random Effects-Expectation Maximization Regression Tree Model in Medical System Structure Data and Model Evaluation. Guangdong Pharmaceutical University, 2019. DOI:10.27690/d.cnki.ggdyk.2019.000015.

[7] Shen Jie, Xing Haijun, Yu Qian, et al. Baseline Load Estimation Method Based on k-shape Clustering and TCN-Attention-XGBoost. Acta Energiae Solaris Sinica , 2025, 46(08): 676-685. DOI:10.19912/j.0254-0096.tynxb.2024-0661.

[8] Xie Xiaoyuan, Li Kuotao, Wu Fang, et al. Application of Expanded Non-invasive Prenatal Testing in the Detection of Fetal Chromosomal Microdeletion/Microduplication Syndromes. Chinese Journal of Birth Health & Heredity , 2024, 32(03): 577-581. DOI:10.13404/j.cnki.cjbhh.2024.03.020.

[9] Wang Yaqun, Liu Hui. Investigation of Pregnant Women's Awareness and Choice of Non-invasive Prenatal Testing. Anhui Medical Journal , 2023, 44(07): 869-872.

[10] Zhang Chu, Wang Jianhong, Li Rui, et al. Results of Expanded Non-invasive Prenatal Testing in 5696 Pregnant Women. Laboratory Medicine , 2023, 38(06): 553-558. DOI: CNKI:SUN:SHYY.0.2023-06-003.

[11] Xu Yunzhi, Xu Xueqin, Tang Shaohua, et al. Clinical Value of Non-invasive Prenatal Gene Testing in Screening for Fetal Sex Chromosome Aneuploidy Disorders. Journal of Wenzhou Medical University , 2022, 52(04): 294-299. DOI: CNKI:SUN:WZYX.0.2022-04-007.

[12] Cheng Yuan. Application of Non-invasive Prenatal Gene Testing Technology in Screening for Chromosomal Abnormalities in 1048 Twin Pregnancies. Anhui Medical University, 2022. DOI:10.26921/d.cnki.ganyu.2022.000869.

[13] Gao Xinghua. Application Effect of Non-invasive Prenatal Gene Testing in Prenatal Screening. China Practical Medicine , 2021, 16(33): 57-60. DOI:10.14163/j.cnki.11-5547/r.2021.33.020.