Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

A Study on Optimal NIPT Timing Selection and Intelligent Fetal Abnormality Dectection Based on Machine Learning

Mingxiang Mo

School of Electrical and Information Engineering, Jiangsu University of Science & Technology (Zhangjiagang), Zhangjiagang, China, 215600

y0219469@gmail.com

Abstract. To address the challenges of uncertain testing time points and the difficulty in determining abnormalities due to individual variations in non-invasive prenatal testing (NIPT), this paper introduces an integrated machine learning framework that amalgamates prediction, optimization, and classification. The framework is designed to identify key factors influencing the quality of testing data and, based on these factors, to offer personalized optimal testing time points and highly accurate solutions for abnormality determination. Initially, the study utilizes an XG-Boost regression model to elucidate the complex nonlinear relationships among critical quality control indicators and factors such as gestational age at testing and maternal BMI. Building upon this, a novel approach that integrates weighted K-means clustering with nonlinear programming is employed to ascertain the optimal testing window for groups of pregnant women with varying physiological characteristics. Finally, to address the challenge of classifying rare abnormal samples, an adaptively weighted and optimized random forest model is developed. Experimental results indicate that this framework effectively handles the data nonlinearity and individual heterogeneity. The final classification model achieves a high recall rate of 85% and a precision of 97%, underscoring the advanced nature of the proposed method and its potential for clinical application.

Keywords: XG-Boost regression; nonlinear optimization; random forest classification; K-means clustering; fetal abnormality detection.

1. Introduction

Non-invasive prenatal testing (NIPT) has emerged as a pivotal technology in prenatal screening, evaluating the risk of chromosomal abnormalities through the analysis of cell-free fetal DNA in maternal peripheral blood [1, 12]. The accuracy of NIPT is contingent upon a sufficient concentration of fetal DNA, a parameter influenced by various factors such as gestational age and maternal body mass index (BMI) [2]. These factors contribute to significant individual variability and nonlinear dynamic changes. Presently, clinical practice predominantly adheres to broad gestational age guidelines for NIPT, a "one-size-fits-all" approach that inadequately addresses individual heterogeneity [3, 10]. This often results in test failures or inaccurate outcomes due to insufficient DNA concentration, thereby exacerbating patients' economic and psychological burdens [11]. Although the academic community has endeavored to employ statistical methods to analyze these influencing factors, most extant studies utilize linear models, which are inadequate for capturing the intricate, nonlinear relationships between variables [4]. Furthermore, they often treat the selection of test timing and the identification of abnormalities as discrete issues, lacking a comprehensive, end-to-end solution that encompasses everything from data quality optimization to final diagnosis.

In response to the limitations of current methodologies, this study introduces a multi-stage integrated algorithm framework that amalgamates prediction, clustering, optimization, and classification. The core advantages of this framework are as follows: 1. Deep insight: Advanced machine learning models, such as XG-Boost, replace traditional linear models, facilitating a more precise depiction of complex relationships between key indicators and multidimensional features. 2. Decision optimization capability: The approach extends beyond mere prediction. Prediction models are embedded within nonlinear programming models, and, in conjunction with weighted K-means

clustering, proactively determine the "optimal" testing time for diverse population groups, thereby achieving a transition from "prediction" to "decision." 3. Robust classification performance: To address the prevalent issue of class imbalance in medical diagnosis, adaptive weighting and other strategies are employed to optimize the random forest classifier, ensuring high recall rates in identifying minority abnormal samples.

This paper seeks to address key challenges in NIPT data analysis through this integrated framework. Initially, the data will undergo systematic preprocessing, and the XG-Boost model will be employed to quantitatively assess the influence of various factors on fetal Y chromosome concentration. Subsequently, an individual-based nonlinear optimization model will be constructed and solved to ascertain the optimal NIPT timing for different groups of pregnant women. Finally, an efficient random forest classification model will be established and evaluated for the intelligent identification of chromosomal abnormalities in female fetuses, with an analysis of its key decision features.

2. Methods

To achieve the objectives outlined above, the algorithm designed in this paper begins with data preprocessing. This step includes cleaning missing values and outliers, as well as one-hot encoding categorical features (such as pregnancy method), converting the raw data into a structured format suitable for machine learning models.

$$\mathbf{x}_{\text{one-hot}} = [x_1, x_2, \text{dots}, x_k, \text{dots}, x_K]^T \quad \text{where} \quad x_k = \begin{cases} 1 & \text{if category} = k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The core modeling section is divided into three modules. The first module focuses on quantifying key factors, for which we constructed an XG-Boost regression model [5, 13, 14]. This model uses fetal Y chromosome concentration as the prediction target, with eight core features—such as gestational age, BMI, and age—selected via Spearman correlation analysis as inputs. XG-Boost iteratively trains a series of decision trees, with each new tree aiming to correct the residuals of the previous trees. Its objective function includes a loss term and a regularization term, allowing it to effectively capture nonlinear relationships among features and prevent overfitting.

$$\text{Obj} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \quad \text{where} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2}$$

$$\text{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{3}$$

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{\left(G_L + G_R\right)^2}{H_L + H_R + \lambda} \right] - \gamma \tag{4}$$

Using this model, we calculated the importance scores of each feature and conducted permutation tests to verify their significance.

The second module is optimal timing decision-making. Here, we first applied a weighted K-means clustering algorithm to group samples based on multiple dimensions such as BMI, height, and weight, determining the optimal number of clusters using the elbow method and silhouette coefficient [6].

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} \left[\omega_1 \cdot \left(B_i - \overline{B}_k\right)^2 + \omega_2 \cdot \left|\mathbf{x}_i - \overline{\mathbf{x}}_k\right|^2 \right] \tag{5}$$

For each cluster, we then established a nonlinear optimization model [7]. This model aims to minimize the gestational week at which testing occurs, with constraints including: the Y chromosome concentration predicted by the XG-Boost model must reach the clinical threshold (4%) at a certain confidence level, and the test time must fall within a specific gestational age range. We used a grid search method to solve this optimization problem, thus determining the optimal NIPT timing for each group.

The third module addresses intelligent anomaly detection. This task is defined as a binary classification problem, for which we built a random forest classification model. Given the rarity of anomalous samples, we introduced an adaptive sample weighting mechanism during model training, using a weighted Gini impurity as the node splitting criterion to increase the model's focus on minority classes [8, 15, 16].

$$w_c = \frac{N}{2 \cdot N_c} \tag{6}$$

$$\text{Gini}_w = \sum_{c=0}^{1} w_c \cdot p_c(1 - p_c) \tag{7}$$

The model inputs include the Z scores of related chromosomes, GC content, and maternal physiological indicators. To determine the optimal decision boundary, we analyzed the receiver operating characteristic (ROC) curve and maximized the Youden index, thereby identifying the optimal classification probability threshold.

$$J(\tau) = \text{TPR}(\tau) - \text{FPR}(\tau) \tag{8}$$

$$\tau^* = \arg\max_{\tau \in [0,1]} J(\tau) \tag{9}$$

3. Results

This chapter aims to present and analyze the key results of our constructed model framework on the NIPT dataset. The analysis is divided into two main sections: the first focuses on identifying factors influencing Y-chromosome concentration and optimizing the best detection time point; the second centers on evaluating the performance of the intelligent fetal chromosomal abnormality detection model for female fetuses.

3.1. Key Factor Identification for Y-Chromosome Concentration and Optimal Detection Timing

The first step in the timing optimization task is to understand which factors are decisive. Through analysis with an XG-Boost regression model, we quantified the importance of various factors affecting fetal Y-chromosome concentration. As shown in Figure 1, the model's results clearly highlight the three core factors influencing concentration: gestational week at sampling (importance 0.1835), maternal BMI (0.1817), and genome alignment rate (0.1779). These quantitative findings are highly instructive, indicating that the fetus's developmental stage (gestational week), the mother's physiological status (BMI, which affects DNA dilution via blood volume), and the quality of experimental operations (alignment rate) are key determinants of whether NIPT data is valid.

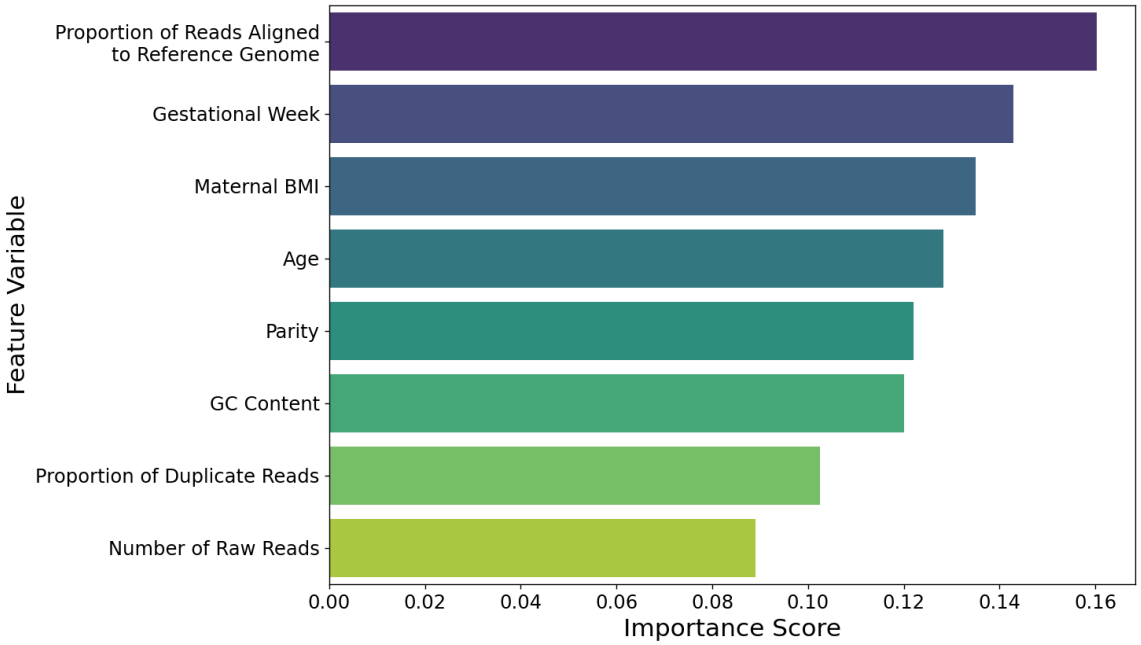

Figure 1. Feature importance plot

To further understand how these core factors influence Y-chromosome concentration, we plotted partial dependence diagrams. As shown in Figure 2, these diagrams vividly reveal the strong nonlinear relationships between variables. The impact of gestational week exhibits a clear phase: before week 21, the concentration increases relatively slowly, forming a plateau; after week 21, the concentration curve rises sharply, closely aligning with the physiological mechanism of a surge in placental DNA release during late development. The impact of maternal BMI, on the other hand, shows a threshold effect: when BMI is below 34, its inhibitory effect on concentration is relatively mild; once it exceeds the critical point of 34, concentration drops precipitously, strongly suggesting that high BMI poses significant challenges to detection success.

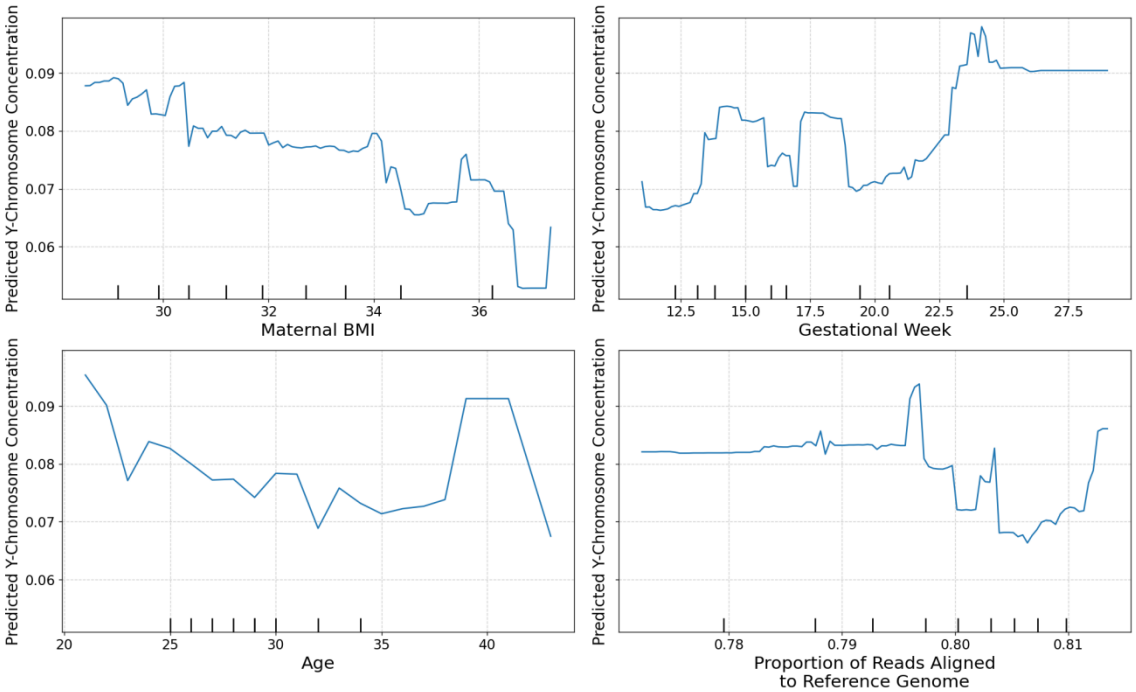

Figure 2. Partial dependence plots (PDP) of the four most important features

Based on these findings, we conducted fine-grained grouping of pregnant women using weighted K-means clustering. As shown in Figure 3, this algorithm used a combination of features including BMI, height, and weight to data-drivenly divide the samples into three internally homogenous groups.

These groups can be clearly defined as low-to-moderate BMI ([20.70, 31.78)), high BMI ([31.78, 35.90)), and extremely high BMI ([35.90, 46.88)). This data-driven grouping approach is much more scientific than relying on fixed clinical standards, laying a solid foundation for developing truly personalized detection strategies in the future.

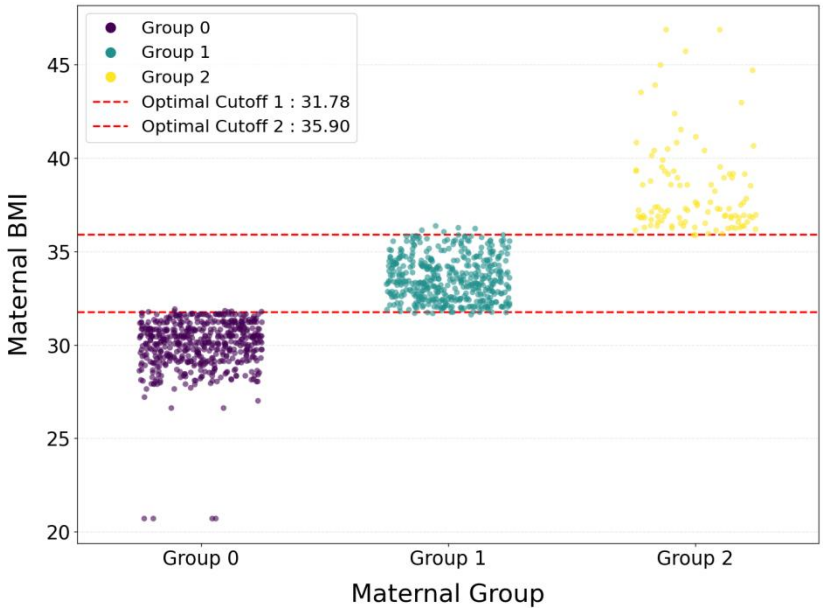

Figure 3. BMI group visualization

For these three groups, we used a nonlinear optimization model to solve for their respective optimal NIPT time points. As shown in Figure 4, the model's output decisions vary greatly—the optimal detection times are 13.3 weeks, 13.5 weeks, and 23.3 weeks for the three respective groups. This result is of significant clinical value as it quantifies the detection window for different populations: for the low-to-moderate BMI group, testing can be performed relatively early in pregnancy, while for the extremely high BMI group, the recommended testing is delayed by as much as 10 weeks. This reveals the decision logic of the optimization model: to satisfy the strict requirement that concentration must meet the standard, the model must propose a sufficiently late time point for the high-risk group, allowing time for their DNA concentration to slowly rise to a reliably detectable level.

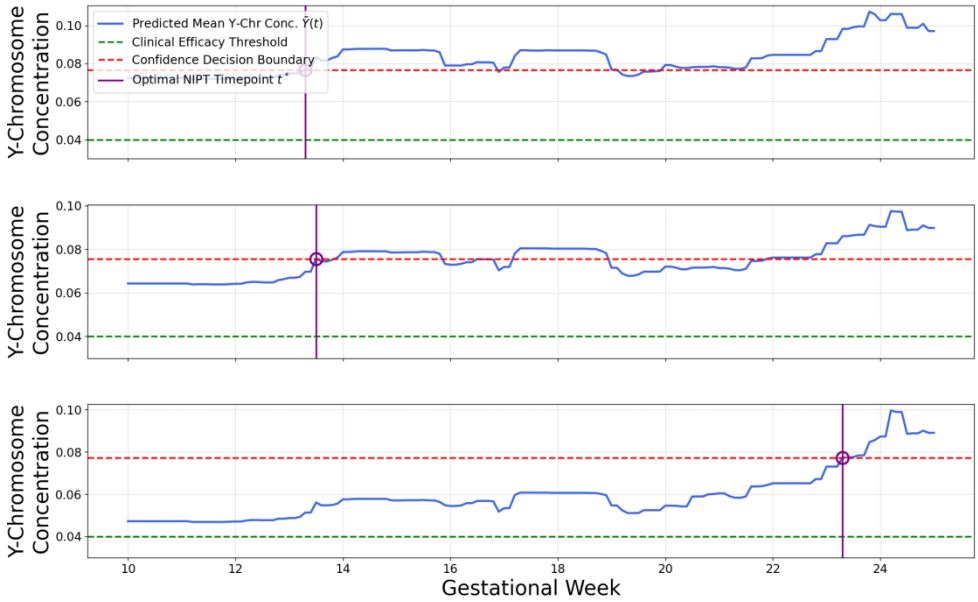

Figure 4. Optimal timing for BMI groups

To evaluate the robustness of the recommended time points, we conducted Monte Carlo simulations. As shown in Figure 5, this error analysis chart reveals the certainty of decisions for different groups. For the low-to-moderate BMI group, the simulated achievement time distribution is highly concentrated, with a narrow 95% confidence interval, indicating that the model's recommendations of 13.3 and 13.5 weeks are highly reliable and stable. However, for the extremely high BMI group, the distribution of time to threshold is extremely dispersed, with the confidence interval nearly covering the entire second trimester, indicating high inherent uncertainty in predictions for this group. Nonetheless, the model's recommended 23.3 weeks still achieved a 91.8% success rate in the simulation, demonstrating that the decision remains a robust optimal solution balancing success rate and early detection, even under high uncertainty.

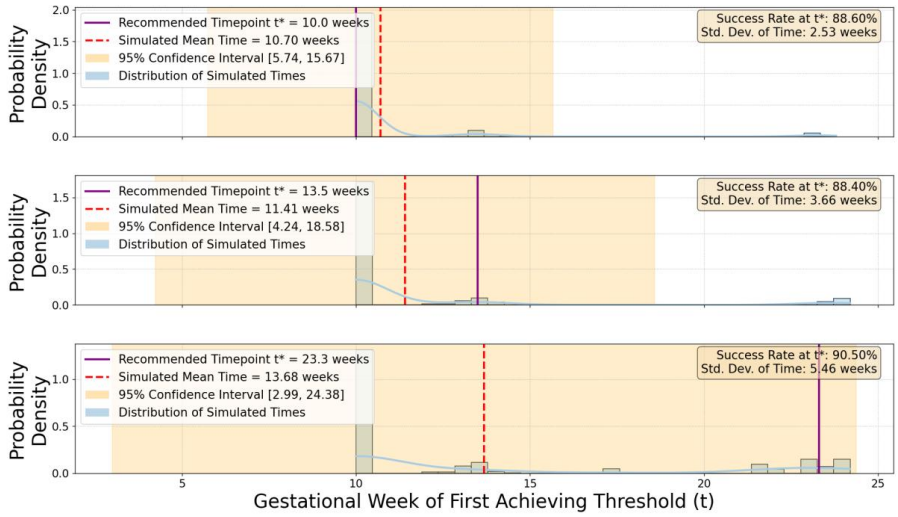

Figure 5. Error analysis plot

3.2. Intelligent Determination and Feature Analysis of Female Fetal Chromosomal Abnormalities

In the task of detecting female fetal abnormalities, our constructed random forest classification model performed exceptionally well. As shown in Figure 6, the model's ROC curve deviates significantly from the diagonal line running from the lower left to the upper right (which represents random guessing) and rises rapidly toward the upper left corner. This indicates that the model has an extremely high discriminative ability in distinguishing between normal and abnormal samples. The area under the curve (AUC) is large, further confirming the overall performance of the model. The "optimal point" found by maximizing the Youden index represents the best balance between "detecting as many abnormalities as possible (high recall rate)" and "minimizing false positives among normal samples (low false positive rate)."

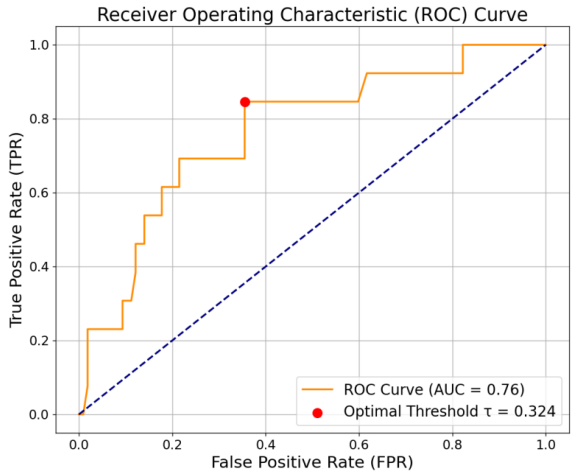

Figure 6. ROC Curve Analysis Chart

After applying the optimal threshold to the test set, the model's specific performance is clearly presented in the confusion matrix shown in Figure 7. Among the 13 true abnormal samples, the model successfully identified 11 (true positives), missing only 2 cases (false negatives), resulting in a recall rate as high as 85%. In clinical practice, a high recall rate is crucial because it directly relates to preventing missed diagnoses. Meanwhile, among samples classified as "normal," the precision reached 97%, indicating that when the model concludes a sample is normal, this result is highly reliable, effectively avoiding unnecessary panic and invasive follow-up checks for the vast majority of normal pregnant women.

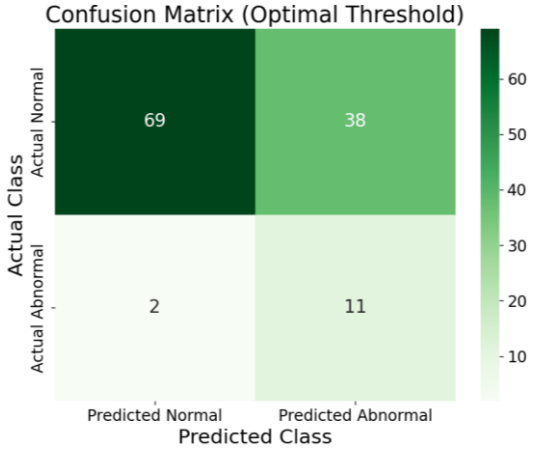

Figure 7. Confusion Matrix Illustration

Finally, we explored the key factors driving the classification of female fetal abnormalities. As shown in Figure 8, the feature importance analysis revealed a very interesting phenomenon: the traditional diagnostic indicator, Z-score, in fact had a relatively low importance score, while quality control indicators such as GC content on chromosomes 21, 13, and 18, and the concentration of the X chromosome, played a dominant role. This result suggests that our model goes beyond simple threshold judgment, learning deeper patterns: chromosomal abnormalities are not only reflected in changes to the Z-score, but are also often accompanied by subtle "data fingerprints," such as shifts in GC biases during sequencing [9-16]. This demonstrates that the model is able to mine powerful diagnostic signals from data originally considered only for quality control.

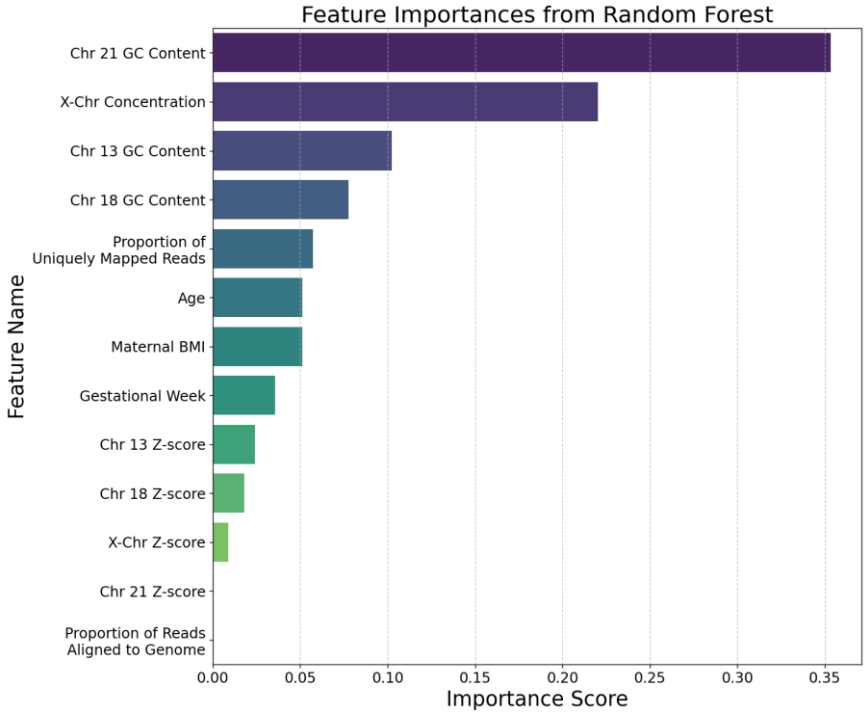

Figure 8. Feature Importance Ranking Chart

4. Conclusion

This study proposes and validates an integrated machine learning and optimization algorithm framework to address two core issues in NIPT data analysis: time point selection and anomaly detection. The main contributions of this research are as follows: First, through the XG-Boost model, nonlinear features influencing key indicators were successfully extracted and quantified from high-dimensional data, providing a reliable predictive foundation for subsequent optimization; second, the designed "weighted clustering + nonlinear optimization" paradigm can generate differentiated optimal decisions based on individual data characteristics, with its effectiveness and robustness validated through Monte Carlo simulations; finally, for classification tasks with imbalanced classes, the constructed weighted random forest model demonstrated outstanding performance, achieving a balance between high recall and high precision.

This study confirms that the organic integration of multiple machine learning algorithms can effectively solve complex biomedical data analysis problems. The strength of the model lies in its fully data-driven approach and adaptability to individual differences. However, the model's performance still depends on the quality and scale of the training data, and in certain feature ranges, such as high BMI, prediction uncertainty increases. Future research directions may focus on introducing more advanced models (such as deep learning networks) to further capture deeper patterns in the data, or extending this algorithmic framework to other similar biomedical decision-making problems.

References

[1] Zhou, C. X., He, L. L., Zhu, X. Y., Li, Z. X., Duan, H. L., Liu, W., Gu, L. L., & Li, J. (2023b). [Report content and prenatal diagnosis of non-invasive prenatal testing for sex chromosome aneuploidy]. PubMed, 58 (10), 766–773.

[2] Qiongzhen Z, JiaNaGuLi H, Shanshan D. The influence of fetal gender and maternal characteristics on fetal cell-free DNA in maternal plasma [J]. Journal of gynecology obstetrics and human reproduction, 2019, 48 (8): 653-656.

[3] Stevens C, Llorin H, Gabriel C, et al. Genetic counseling for fetal sex prediction by NIPT: Challenges and opportunities [J]. Journal of Genetic Counseling, 2023, 32 (5): 945-956.

[4] Binson V A, Thomas S, Subramoniam M, et al. A review of machine learning algorithms for biomedical applications [J]. Annals of Biomedical Engineering, 2024, 52 (5): 1159-1183.

[5] Gao, L., Feng, J., Gao, Y., Luo, L., Jiang, H., Yang, Q., Lu, J., & Guo, L. (2025). XGBoost-based model for predicting PICC occlusion risk in cancer patients: Insights from SHAP analysis. Alexandria Engineering Journal, 123, 436–447.

[6] Mim S S, Logofatu D, Leon F. Efficient Analysis of Patient Length of Stay in Hospitals Using Classification and Clustering Approaches [C] // International Conference on Computational Collective Intelligence. Cham: Springer Nature Switzerland, 2023: 675-688.

[7] Ugolkov Y, Nikitich A, Leon C, et al. Mathematical modeling in autoimmune diseases: from theory to clinical application [J]. Frontiers in Immunology, 2024, 15: 1371620.

[8] Li J, Ju J, Zhao Q, et al. Effective identification of maternal malignancies in pregnancies undergoing noninvasive prenatal testing [J]. Frontiers in Genetics, 2022, 13: 802865.

[9] Chen M, Jiang F, Guo Y, et al. Validation of fetal DNA fraction estimation and its application in noninvasive prenatal testing for aneuploidy detection in multiple pregnancies [J]. Prenatal Diagnosis, 2019, 39 (13): 1273-1282.

[10] Liehr T. Noninvasive prenatal testing (NIPT) results are less accurate the later applied during pregnancy [J]. Taiwanese Journal of Obstetrics and Gynecology, 2024, 63 (6): 892-895.

[11] Chen Y, Yang F, Shang X, et al. A study on non-invasive prenatal screening for the detection of aneuploidy [J]. Ginekologia Polska, 2022, 93 (9): 716-720.

[12] Liu S, Chang Q, Yang F, et al. Non-invasive prenatal test findings in 41,819 pregnant women: results from a clinical laboratory in southern China [J]. Archives of Gynecology and Obstetrics, 2023, 308 (3): 787-795.

[13] Zheng J, Li J, Zhang Z, et al. Clinical Data based XGBoost Algorithm for infection risk prediction of patients with decompensated cirrhosis: A 10-year (2012–2021) Multicenter Retrospective Case-control study [J]. BMC gastroenterology, 2023, 23 (1): 310.

[14] Gupta R, Bhandari M, Grover A, et al. Correction: Predictive modeling of ALS progression: an XGBoost approach using clinical features [J]. BioData Mining, 2025, 18: 5.

[15] Zhu M, Xia J, Jin X, et al. Class weights random forest algorithm for processing class imbalanced medical data [J]. IEEE access, 2018, 6: 4641-4652.

[16] Salehpour A, Norouzi M, Balafar M A, et al. A cloud-based hybrid intrusion detection framework using XGBoost and ADASYN-Augmented random forest for IoMT [J]. IET Communications, 2024, 18 (19): 1371-1390.