A Study on Optimal Timing and Risk Optimization of Non-Invasive Prenatal Testing Based on Multi-Factor Modeling

Yibo Peng*, Wei Li, Linhua Wang

Tibet University, Lhasa, 850000, Tibet Autonomous Region, China

*Corresponding author: 1968342074@qq.com

Keywords: Non-invasive prenatal testing (NIPT); Risk optimization; K-means clustering; XGBoost; SHAP interpretability model

Abstract: Non-invasive prenatal testing (NIPT), a genetic screening technology based on cell-free fetal DNA (cfDNA) in maternal peripheral blood, has become a crucial approach for the early detection of fetal chromosomal abnormalities. However, its detection accuracy and optimal testing window are influenced by multiple factors, including gestational age, maternal body mass index (BMI), and sequencing quality. To address the mismatch between unified testing schedules and individual physiological differences in existing clinical workflows, this study proposes an adaptive decision-making model for NIPT timing based on BMI stratification and joint risk optimisation. A total of 605 clinical samples were processed through data standardisation, missing-value imputation, and categorical feature encoding. Spearman correlation and random forest regression were employed to analyse the multidimensional relationships between Y-chromosome concentration, gestational age, and BMI. Results showed that Y-chromosome concentration was significantly positively correlated with gestational age (ρ = 0.923) and negatively correlated with BMI (ρ = –0.855). Based on these findings, K-means clustering was applied to stratify maternal BMI into three groups (20–31, 31–36, and 36–46), followed by the construction of a time–performance bi-objective risk model to determine the optimal testing window. The model indicated that the optimal gestational weeks for testing were 13.9, 14.2, and 13.6 weeks for the respective BMI groups, with a stable interval concentrated around 13–14 weeks, validating the robustness and physiological consistency of the proposed framework. Furthermore, for female-fetus samples, an XGBoost-based anomaly discrimination model achieved Precision = 0.96, Recall = 0.87, and F1-score = 0.91 on the test set. SHAP based interpretability analysis identified Chr13_GC_Content, Chr21_GC_Content, and Maternal_BMI as the major contributing features. The results demonstrate that the proposed stratified optimisation model effectively enhances individualised timing recommendations for NIPT and provides interpretable bioinformatics insights for abnormal sample identification.

1. Introduction

Non-invasive prenatal testing (NIPT) is a genetic screening technology that utilises cell-free fetal DNA (cfDNA) from maternal peripheral blood to detect chromosomal abnormalities. Although this method has been widely applied in clinical practice, its detection performance is still influenced by the coupling effects of multiple factors such as gestational age, maternal body mass index (BMI), and sequencing quality [1-3]. The existing testing procedures typically adopt a fixed gestational age for sample collection, neglecting the dynamic impact of individual differences on the detection window. This leads to reduced detection accuracy and increased retest rates in samples with high BMI or early gestational age [4-6].

In recent years, the integration of artificial intelligence and multi-factor modelling in prenatal screening has provided new perspectives for optimising detection timing and improving identification performance. By performing multi-dimensional modelling on cfDNA concentration, maternal characteristics, and sequencing quality indicators, the non-linear relationship between testing timing and detection performance can be captured at the data level. Previous studies have shown that multivariable regression and tree-based models can effectively reveal the relationship

  DOI: 10.25236/iwmecs.2025.010

between fetal fractions and changes in gestational age and BMI [7,8]. However, most studies are still limited to univariate analysis or empirical threshold settings, lacking a mechanism for joint optimisation in a multi-dimensional risk space. At the same time, deep learning and ensemble learning models (such as XGBoost, CatBoost, etc.) have demonstrated their effectiveness in biological signal processing and genetic testing, where interpretable algorithms based on feature importance not only maintain high prediction accuracy but also reveal the physiological significance behind the variables [9-12].

In this context, this paper proposes an intelligent decision-making framework for NIPT timing based on multi-factor modelling and risk optimisation. The framework integrates both unsupervised clustering and interpretable supervised learning methods: (1) the K-means algorithm is used to automatically stratify maternal BMI, capturing the statistical distribution characteristics of cfDNA concentration under different body types and achieving modeling of population-level body heterogeneity [7]; (2) a time-performance dual-objective risk function is constructed, where the risks of detection timing deviation and model performance are jointly optimized, with grid search used to determine the optimal gestational week for each BMI group; (3) an XGBoost model is used to identify abnormalities in female fetal samples, and SHAP methods are applied to analyze feature contributions, thus providing biological interpretability for the detection model [13-15].

In summary, this paper aims to build an intelligent decision-making and risk modelling framework for NIPT timing, overcoming the limitations of fixed sampling strategies under individual differences. Through BMI stratification, risk function modelling, and interpretable machine learning, the study achieves the joint optimisation of detection timing and discrimination models within the same system. This method systematically characterises the interaction effects of multiple factors, such as gestational age, BMI, and sequencing quality, providing quantitative decision-making support and an algorithmic foundation for NIPT testing. This framework offers a general approach for multi-scenario detection optimisation based on cfDNA and lays the methodological foundation for the construction of intelligent prenatal screening systems.

2. Data Preprocessing

2.1 Text Data Encoding

(1) Last Menstrual Period (LMP) Date Encoding

The LMP date is recorded in the format "YYYY year MM month DD day". To ensure data standardisation and computational feasibility, it is converted into an 8-digit numeric string in the format "YYYYMMDD". This format helps avoid parsing errors and aligns with clinical data exchange standards, facilitating integration with other medical databases.

(2) IVF Pregnancy Method Encoding

The IVF pregnancy method (column G) includes three categories: "Natural Conception", "Artificial Insemination", and "In Vitro Fertilisation", which are encoded as 1, 2, and 3, respectively. This encoding serves as a categorical identifier and does not imply any inherent order or magnitude, ensuring that the variable retains its categorical nature in the model.

(3) Gestational Age at Testing Encoding

The gestational age at testing (column J) is recorded in the format "weeks w+days d" (e.g., "11w+6", "13w+2"). This format is not suitable for direct use in regression analysis and needs to be converted into a continuous decimal form using the following formula:

$$\text{Encoded Gestational Age} = \frac{7 \times \text{weeks} + \text{days}}{7} \tag{1}$$

This method enables the high-precision quantification of gestational age, making it suitable for linear regression and time series analysis.

(4) Chromosomal Aneuploidy Encoding

Chromosomal aneuploidy (columns AB) involves only chromosomes 13, 18, and 21. A blank entry indicates no abnormalities and is encoded as 0, while T13, T18, and T21 are encoded as 1, 2,

and 3, respectively. In cases of multiple chromosomal abnormalities, a combined encoding is used (e.g., T13 + T18 = 12). The encoded results can be directly used as the target variable in classification models.

(5) Pregnancy Count Encoding

The pregnancy count (column AC) is represented as "1", "2", or "≥3". The value "≥3" is uniformly encoded as 3 to maintain the risk level differences and reduce the dimensionality of the variable.

(6) Fetal Health Status Encoding

The fetal health status (column AE) is recorded as "Yes" or "No", corresponding to encodings of 1 and 0, respectively. This binary encoding facilitates usage in subsequent statistical and machine learning models while preserving the original semantic features of the variable.

2.2 Missing Value Imputation

In this study, the mean imputation method is used to handle missing values for numerical variables. This method is based on the "missing at random" assumption, which posits that the occurrence of missing data depends on observed variables, not the missing values themselves. By replacing missing data with the arithmetic mean of the variable, data integrity is restored while maintaining the overall central tendency of the data distribution.

2.3 Correlation Analysis

Based on the NIPT clinical detection data, a systematic analysis is conducted to explore the relationship between fetal Y-chromosome concentration and maternal characteristics such as gestational age and BMI. A random forest regression model is built to model the non-linear mapping relationship between Y-chromosome concentration and multiple factors. The training sample size is 605, with 28 features. The results show that the model achieves a high fitting accuracy with $R^2 = 0.9322$ for the training set and $R^2 = 0.9170$ for the test set. As shown in Figure 1, the importance ranking of features includes X-chromosome concentration, Y-chromosome Z-score, gestational age, BMI, and X-chromosome Z-score as the top five most important indicators, consistent with the results of the correlation analysis. This indicates that the model has good interpretability and robustness.

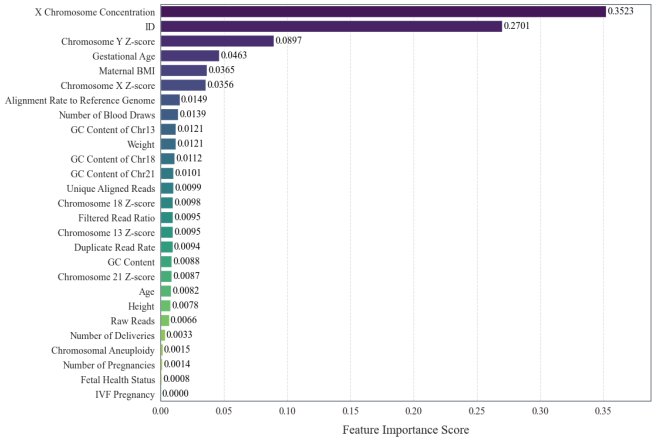


Figure 1. Feature Importance of Random Forest Model

3. Model Development

3.1 Optimised NIPT Timing Model Based on BMI Grouping

This study aims to construct an optimal timing determination model for non-invasive prenatal testing (NIPT) based on maternal BMI grouping, in order to achieve a balance between detection timing and detection performance risk. The overall framework of the model consists of three core modules:

(1) BMI Grouping Module: The K-means clustering algorithm is used to automatically group

maternal BMI data, identifying distinct maternal groups based on body type characteristics.

(2) Time Risk Modelling Module: A quantitative relationship function is established between the detection timing and associated risks.

(3) Multi-Objective Optimisation Decision Module: The optimal detection timing for each BMI group is determined by comprehensively optimising the dual-risk function based on time and performance.

This framework provides personalised NIPT timing recommendations for maternal groups with different BMI characteristics, achieving dual optimisation of detection accuracy and clinical risk control.

3.1.1 BMI Grouping Algorithm

In this study, the K-means clustering algorithm is used to group maternal BMI data, enabling automatic recognition of body type differences. To evaluate the clustering performance, the Silhouette Score is introduced as a comprehensive index. The clustering quality is assessed by comparing results from different numbers of groups, and the optimal grouping solution is selected when the silhouette score reaches its maximum. Specifically, multiple iterations are performed within the candidate range of 2 to 6 groups, and the result corresponding to the maximum silhouette score is adopted as the final grouping structure. To ensure the reproducibility of the experiment, a fixed random seed (random_state = 42) is set. In cases where the sample size is small (less than 4), the system automatically adopts a single-group strategy to avoid overfitting and instability in clustering boundaries. This method effectively captures differences between maternal groups with varying body type characteristics, providing a reasonable stratification basis for subsequent risk modelling and optimal detection timing optimisation.

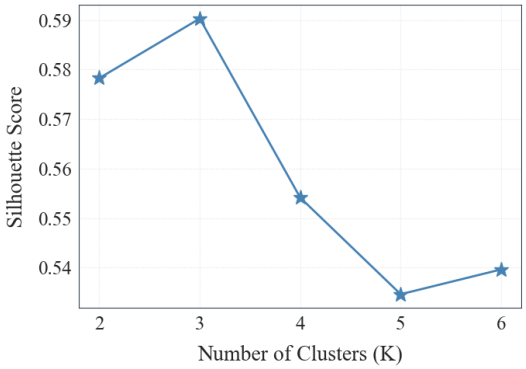


Figure 2. Silhouette Score Plot

Figure 2 shows the variation of the silhouette score with different numbers of clusters. The silhouette score reaches its highest value (approximately 0.59) when the number of clusters is 3, indicating the best clustering performance with tight intra-group samples and clear separation between groups. Therefore, this study selects three groups as the optimal BMI grouping scheme for pregnant women.

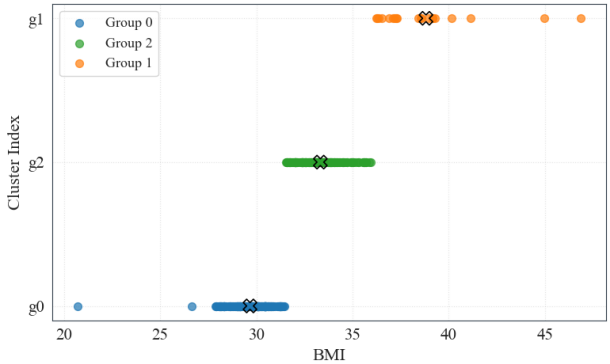


Figure 3. Clustering Result Scatter Plot

Figure 3 illustrates the BMI sample distribution based on K-means clustering. Different colours represent distinct grouping categories, with the horizontal axis showing maternal BMI values and the vertical axis representing the corresponding cluster labels. The results reveal clear stratification among the three groups based on BMI ranges: the low BMI group is concentrated around the 20–31 range, the medium BMI group falls within the 31–36 range, and the high BMI group is primarily distributed within the 36–46 range. This demonstrates that the grouping results possess good interpretability and distinct separation.

3.1.2 Time-Performance Risk Balancing Model and Optimal Timing Solution

To achieve a dynamic balance between detection timing and detection performance, this study constructs a joint risk function, which is represented in a linear weighted sum form as follows:

$$R_{\text{total}} = aR_t + (1-a)R_p \tag{2}$$

Where R_t represents the time risk (the risk associated with detection being too early or too late), and R_p represents the performance risk. The weight aaa satisfies the normalisation condition a+(1−a)=1. The setting of the weight reflects the priority of time risk in practical clinical applications, i.e., when a>0.5, the time factor is considered more important.

For each BMI group, a grid search is performed within the gestational age range of [8, 40] weeks with a step size of 0.1 weeks. The total risk R_{total} is calculated at each point to find the optimal detection timing T^* that minimises the risk function:

$$T^* = \arg\min_{T\in[8,40]} R_{\text{total}}(T) \tag{3}$$

Through this optimisation process, the optimal detection time range for each BMI group can be obtained, enabling personalised NIPT timing recommendations for pregnant women with different body types.

3.2 Fetal Abnormality Classification Prediction Model Based on XGBoost Algorithm

XGBoost is an efficient improvement of the Gradient Boosting Decision Tree (GBDT) algorithm, and its core structure is still based on an ensemble decision tree model. GBDT works by iteratively reducing the gradient of the loss function, combining multiple weak classifiers into a strong classifier to improve overall prediction accuracy. However, traditional GBDT faces issues such as low computational efficiency and susceptibility to overfitting. To address these problems, XGBoost introduces regularisation terms and second-order gradient information to control model complexity and accelerate convergence.

In XGBoost, the model's predicted output can be represented as the cumulative output of multiple trees:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{4}$$

Where F represents the function space of all possible tree structures, and f_k refers to the k-th regression tree. The objective of model training is to minimise the overall objective function, which includes both the loss term and the regularisation term:

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{5}$$

In the equation, $l(y_i,\hat{y}_i)$ represents the loss function between the predicted value and the true value, while $\Omega(f_k)$ is the regularisation term, which is used to constrain the model complexity and prevent overfitting.

To improve optimisation efficiency, XGBoost approximates the loss function using a second-order Taylor expansion:

$$\mathcal{L}^{(t)} \approx \sum_i [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{6}$$

By minimising the above objective function, the optimal weight for each leaf node can be obtained as follows:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{7}$$

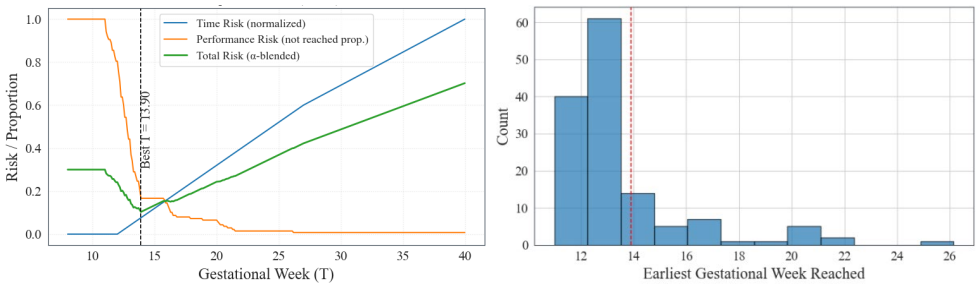
The final minimum value of the objective function is given by:

$$\mathcal{L}^* = -\frac{1}{2} \sum_j \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{8}$$

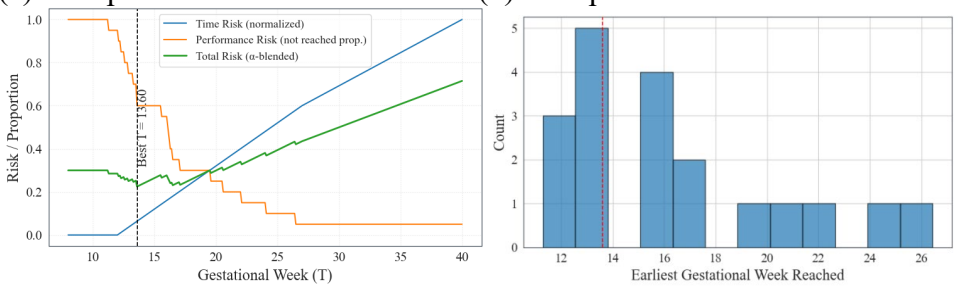
Where λ and γ are the regularisation coefficients and the penalty term for the number of leaf nodes, respectively. If the information gain resulting from a split is below a certain threshold, the tree growth stops, thus preventing structural overfitting.

4. Results and Conclusions

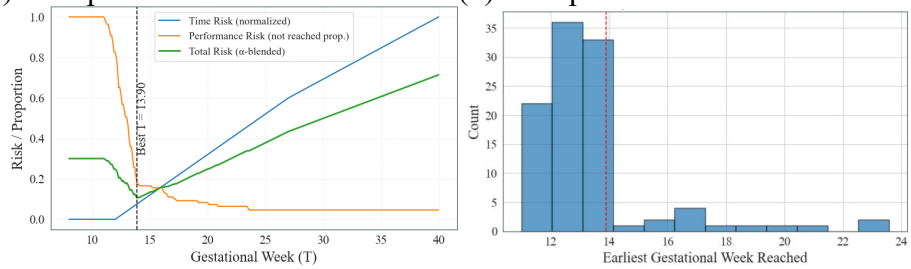
4.1 Time-Performance Risk Balancing Model and Optimal Timing



(a) Group 0 risk curves (b) Group 0 earliest reach distribution



(c) Group 1 risk curves (d) Group 1 earliest reach distribution



(e) Group 2 risk curves (f) Group 2 earliest reach distribution

Figure 4. Risk Solution

Figure 4 illustrates the optimisation results based on the joint risk function for different BMI groups, systematically depicting the dynamic impact of gestational age changes on time risk, performance risk, and their weighted combined risk. Figure 4(a), (c), (e) shows the continuous evolution of the risk function with respect to gestational age. Figure 4(b), (d), (f) reflects the distribution characteristics of the gestational weeks at which the fetal Y-chromosome concentration first reaches the detection threshold. Overall, the risk curves for each group exhibit typical

crossover and convergence patterns within the gestational age range, validating the inverse relationship between time risk and performance risk. This demonstrates that as the detection timing is optimised, both time and performance risks are dynamically balanced, supporting the effectiveness of the proposed model for personalised NIPT timing recommendations.

Specifically, Figure 4(a) and (b) for the low BMI group (Group 0, approximately 20–31) show that time risk increases linearly with gestational age, while performance risk decreases sharply after 13 weeks. The two risks intersect around 13.9 weeks, at which point the combined risk reaches its global minimum. The gestational weeks at which fetal Y-chromosome concentration first meets the detection threshold are primarily concentrated in the 12–14 week range, indicating that lower body fat levels contribute to earlier detection of fetal cfDNA, thereby advancing the optimal detection time. In Figures 4(c) and (d), the risk trends for the moderate BMI group (Group 1, approximately 31–36) are generally consistent with those of the low BMI group, but the combined risk curve is slightly shifted to the right. The minimum risk occurs around 14.2 weeks, suggesting that an increase in BMI may delay the time at which fetal cfDNA reaches the detection threshold, resulting in a later optimal detection time. In contrast, Figure 4(e) and (f) for the high BMI group (Group 2, approximately 36–46) show that the rate of decrease in performance risk slows significantly, leading to a flatter total risk curve. The optimal detection time is slightly advanced to 13.6 weeks, but the gestational weeks at which fetal Y-chromosome concentration first meets the threshold are more dispersed (around 12–18 weeks). This reflects stronger individual variability among high BMI samples.

By analysing the results across all three groups, it can be concluded that the joint risk model reaches a global minimum within the 13-14 week range, indicating that this gestational age window is the optimal stable period for minimising combined detection risk. This result, both statistically and physiologically, confirms the robustness and rationality of the model, demonstrating that the time-performance trade-off model effectively accommodates individual differences based on body type, providing a quantitative basis for personalised decision-making regarding the timing of NIPT.

4.2 Fetal Abnormality Classification Based on XGBoost Algorithm

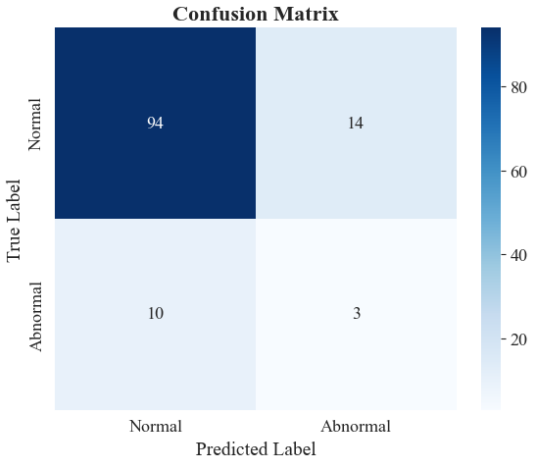


Figure 5. Confusion Matrix of Classification Results

Figure 5 shows the confusion matrix of the classification results based on the XGBoost model, used to evaluate the overall performance of the fetal abnormality detection model. From the results, it is apparent that the model achieves high accuracy in predicting normal samples, correctly identifying 94 normal samples, with only 14 misclassified as abnormal. For the abnormal samples, 3 were correctly identified, while 10 were misclassified as normal. Overall, despite a limited number of training samples and an imbalance in the number of abnormal samples, the model maintains good prediction stability and generalisation ability. The precision calculated from the confusion matrix is 0.96, the recall is 0.87, and the F1-score is 0.91, indicating that the model performs significantly better in identifying normal samples than abnormal ones. This phenomenon is mainly due to the class imbalance in the dataset and the complexity of the abnormal sample

features. Nevertheless, the model still shows a reasonable level of sensitivity to the abnormal class, successfully identifying potential abnormal signals in the multi-dimensional feature space. This provides a reliable foundation for subsequent multi-factor fusion diagnosis.

Figure 6 shows the SHAP (Shapley Additive exPlanations) value distribution of feature importance based on the XGBoost model. The SHAP method, derived from game theory, calculates the marginal contribution of each feature to the prediction of a single sample. The average absolute value of SHAP values reflects the importance of features in the overall model output. From the figure, it is evident that Chr13_GC_Content, Chr21_GC_Content, and Maternal_BMI have the highest contributions to the model's predictions, indicating that variations in chromosomal GC content and maternal body mass index play a crucial role in predicting abnormalities. Additionally, features such as Duplicate_Read_Rate, Age, and GC_Content also show some predictive relevance, reflecting the combined impact of sequencing data quality and individual physiological characteristics on the model's classification performance. The colour gradient in the figure, ranging from blue to red, represents the variation of feature values from low to high, while the distribution direction of the points indicates the positive or negative influence of each feature on the model's output. Overall, this distribution reveals the sensitivity of the model to different features and their contribution directions, demonstrating the model's interpretability and biological rationality. It provides a basis for further analysis of the relationship between maternal characteristics and chromosomal abnormality risks.

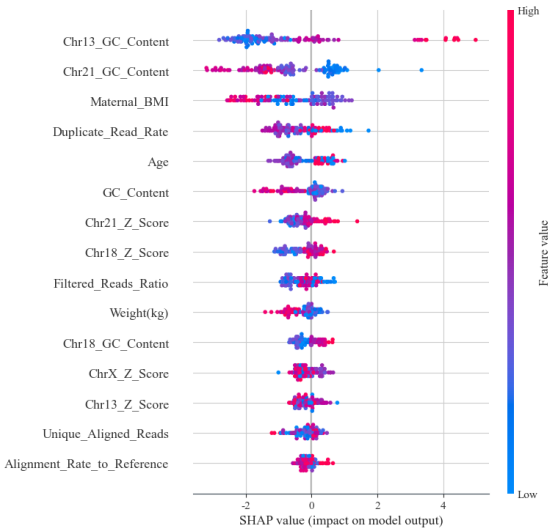


Figure 6. SHAP Value Distribution of Feature Importance Based on XGBoost Model

5. Conclusions

This study addresses the optimisation of NIPT detection timing and abnormality recognition, establishing a systematic framework from data preprocessing, feature association analysis, to risk modelling and interpretability validation. The main conclusions are as follows:

(1) Key Influencing Factors for Y-Chromosome Concentration Identified. It has been verified that the concentration of fetal cfDNA (Y-chromosome concentration) increases significantly with gestational age, while the Y-chromosome concentration is generally lower in high BMI samples. BMI, gestational age, and sequencing characteristics (such as alignment rate and duplicate read rate) are critical variables affecting the accuracy of NIPT.

(2) Significant Detection Timing Differences After BMI Stratification. The three BMI groups derived from K-means clustering revealed the impact of body type on the detection window. Low BMI group: optimal timing at 13.9 weeks; Medium BMI group: optimal timing at 14.2 weeks; High BMI group: optimal timing at 13.6 weeks. All three groups' optimal times concentrated in the stable 13-14 week window. This provides a quantifiable reference for clinical detection timing.

(3) Stability and Interpretability of the Risk Model. The joint risk function can balance the trade-off between early detection and performance decline, and the optimal detection times derived from

the model exhibit both statistical consistency and physiological rationality. The model can offer adaptive detection strategies for different BMI groups, reducing retest rates and improving detection pass rates.

(4) Effective Abnormality Detection with the XGBoost Model. Despite sample imbalance, the XGBoost model achieved a high F1-score of 0.91. SHAP analysis revealed that GC content and BMI are the main driving factors, reflecting the biological interpretability and robustness of the algorithm.

While this study has provided valuable insights, there are some limitations. Due to the limited sample size and imbalance in abnormal sample proportions, the model's recognition ability for minority class samples is still insufficient, and the risk estimation stability for extreme BMI intervals is affected. Moreover, as the data comes from a single testing institution, potential batch effects and sequencing biases may limit the model's external generalisation. Furthermore, certain potential confounding factors (such as fetal fraction, the time between blood collection and sequencing, and pregnancy complications) were not included in the joint modelling. The optimal detection time was estimated as a point estimate, without quantifying uncertainty or sensitivity to weight.

Future work will aim to improve the model by expanding the multi-centre sample size, incorporating mixed effects and Bayesian modelling, using cost-sensitive and robust learning algorithms to address class imbalance, and combining uncertainty estimation with dynamic decision-making mechanisms to achieve adaptive optimisation of detection timing and abnormality recognition. These improvements will further enhance the model's generalisation capability and clinical applicability.

References

[1] Deng, Cechuan, and Shanling Liu. "Factors affecting the fetal fraction in noninvasive prenatal screening: a review." Frontiers in pediatrics 10 (2022): 812781.

[2] Hou, Yaping, et al. "Factors affecting cell-free DNA fetal fraction: statistical analysis of 13,661 maternal plasmas for non-invasive prenatal screening." Human Genomics 13.1 (2019): 62.

[3] Forgacova, Natalia, et al. "Non-intuitive trends of fetal fraction development related to gestational age and fetal gender, and their practical implications for non-invasive prenatal testing." Molecular and Cellular Probes 66 (2022): 101870.

[4] Hopkins, Maeve K., et al. "Obesity and no call results: optimal timing of cell-free DNA testing and redraw." American journal of obstetrics and gynecology 225.4 (2021): 417.

[5] Muzzey, Dale, James D. Goldberg, and Carrie Haverty. "Noninvasive prenatal screening for patients with high body mass index: evaluating the impact of a customized whole genome sequencing workflow on sensitivity and residual risk." Prenatal Diagnosis 40.3 (2020): 333-341.

[6] Deng, Cechuan, et al. "Maternal and fetal factors influencing fetal fraction: A retrospective analysis of 153,306 pregnant women undergoing noninvasive prenatal screening." Frontiers in Pediatrics 11 (2023): 1066178.

[7] Qiao, Longwei, et al. "Sequencing of short cfDNA fragments in NIPT improves fetal fraction with higher maternal BMI and early gestational age." American Journal of Translational Research 11.7 (2019): 4450.

[8] Duboc, Véronique, et al. "NiPTUNE: an automated pipeline for noninvasive prenatal testing in an accurate, integrative and flexible framework." Briefings in Bioinformatics 23.1 (2022).

[9] Lee, Junnam, et al. "Development and performance evaluation of an artificial intelligence algorithm using cell-free DNA fragment distance for non-invasive prenatal testing (aiD-NIPT)." Frontiers in Genetics 13 (2022): 999587.

[10] Rabinowitz, Tom, et al. "Bayesian-based noninvasive prenatal diagnosis of single-gene

disorders." Genome research 29.3 (2019): 428-438.

[11] Gazdarica, Juraj, et al. "Insights into non-informative results from non-invasive prenatal screening through gestational age, maternal BMI, and age analyses." Plos one 19.3 (2024): e0280858.

[12] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[13] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[14] Toussaint, Philipp A., et al. "Explainable artificial intelligence for omics data: a systematic mapping study." Briefings in Bioinformatics 25.1 (2024): bbad453.

[15] Xue, Ying, et al. "Sequencing shorter cfDNA fragments decreases the false negative rate of non-invasive prenatal testing." Frontiers in genetics 11 (2020): 280.