Optimal NIPT Timing Prediction Based on K-Means Clustering and XGBoost Regression

Junnan Yang[1], Hanbing Liu[1], Yichen Ma[1], Wanwan Wang[2]
[1]*School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China*
[2]*iFLYTEK Co., Ltd., Hefei, Anhui, China*

Abstract: With the deep application of artificial intelligence in smart healthcare, utilizing data mining techniques to solve complex clinical decision-making problems has become a research hotspot. Addressing the challenge of high failure rates in Non-Invasive Prenatal Testing (NIPT) caused by maternal physical heterogeneity, this study proposes a data-driven stratified augmented prediction framework. Firstly, the K-Means clustering algorithm is introduced to perform unsupervised stratification on multi-dimensional clinical data, effectively resolving distribution differences among samples. Subsequently, a feature-augmented XGBoost ensemble learning model is constructed to accurately fit the detection baselines of different populations by capturing non-linear interactions among physiological features. Finally, combined with Monte Carlo simulation technology, the uncertainty of prediction results is quantified, and historical data biases are corrected. Experimental results demonstrate that the proposed framework performs exceptionally well in high-risk populations with severe obesity. Compared with traditional linear models, the prediction error is significantly reduced, and the interpretability is greatly improved. The differentiated sampling recommendation table generated by this study provides a scientific basis for realizing "precision triage" and optimizing the allocation of medical resources in clinical practice.

Keywords: Clustering Stratification; Ensemble Learning; Stochastic Simulation; Intelligent Triage

1. Introduction

With the expanding application of non-invasive prenatal testing (NIPT) within antenatal screening systems, enhancing the success rate of single tests and ensuring the accuracy of results have been established as key priorities for clinical research. The implementation pathway for risk assessment involves measuring the concentration of cfDNA in maternal peripheral blood. In male fetal detection scenarios, a Y-chromosome concentration of 4% or higher is established as the fundamental threshold for result credibility. Variations in maternal constitution, particularly regarding gestational age and BMI, have been repeatedly observed to significantly influence fetal DNA concentration distribution. Setting sampling timepoints too early frequently results in insufficient concentration leading to test failure; conversely, arranging sampling too late increases the risk of missing the optimal intervention window. The necessity of clarifying the appropriate gestational weeks for testing across different maternal constitutions is particularly evident within the context of personalised screening requirements. Consequently, the incorporation of NIPT sample data from a specific region enabled quantitative analysis of how BMI and gestational age influence Y-chromosome concentration. The development of predictive models supports the formulation of personalised optimal testing timing recommendations for pregnant women across different BMI groups, demonstrating its reference value in optimising clinical sampling strategies and reducing repeat testing costs.

Research accumulating around the effects of gestational age and BMI on fetal cell-free DNA proportion has reached a substantial scale. Qin et al.[1] modelled multi-chromosomal detection data from 118,969 pregnant women, enabling the demonstration of correlations between maternal physiological indicators and testing stability. Deng et al.[2] and Qiao et al.[3] integrated BMI and gestational age variables within a unified analytical framework, revealing the opposing effects of these factors on fetal DNA proportion.

They also demonstrated enhanced detection sensitivity of short-fragment cfDNA sequencing technology under conditions of high BMI and early pregnancy. A systematic review by Zaki-Dizaji et al.[4] systematic review of maternal age and assisted reproductive factors demonstrated their pervasive influence on cfDNA release, exhibiting interactive dynamics. While such studies provide clear foundational support for understanding influencing factors[5], their analytical focus remains predominantly descriptive.

The complexity of predictive challenges manifests as a primary technical hurdle. Most models are built upon linear assumptions, implicitly requiring the stable superposition of different factors' influences. This premise demonstrates insufficient adaptability in clinical settings. In severely obese populations, the phenomenon of cfDNA concentration plateaus has been documented; when multiple risk factors coexist, the non-linear superposition of combined effects has also been observed. The inadequate learning capacity of existing models regarding nonlinear relationships and variable interactions[6, 7] increases the likelihood of repeated sampling and prolonged diagnostic timelines in high-risk populations. This trend is accompanied by heightened psychological stress and expanded consumption of healthcare resources. Addressing these issues, a stratified analysis method for evaluating NIPT sampling timepoints is introduced. Population subgrouping was achieved through K-Means clustering analysis. The formation of subgroups with similar physical characteristics demonstrated reduced inter-group variability. Within the mixed linear model framework, incorporating nonlinear terms for gestational age and BMI alongside other clinical covariates enabled more comprehensive consideration of background factors.

Practical results demonstrate that the modified XGBoost model possesses the capability to identify the nonlinear dilution effect corresponding to elevated BMI. The interaction between maternal age and other factors is clearly displayed, achieving a population baseline that approximates true physiological states. Situations involving substantial individual prediction uncertainty were incorporated into the analysis. The introduction of uncertainty during sampling, combined with Monte Carlo simulation methods, enabled the integration of actual sampling behaviour into a discrete sample value system. A recommended testing window table stratified by BMI was compiled, meeting the conditions for direct clinical application. This methodology demonstrates significant reference value for severely obese or multiple high-risk pregnant women. It provides a practical foundation for achieving outcomes such as increased first-test pass rates, reduced repeat sampling, shorter waiting times, and conservation of healthcare resources. Furthermore, it offers a reference technical pathway for similar triage processes and sampling timing decisions.

2. Preliminaries

2.1 Data Description

This experimental dataset comprises NIPT data from high-BMI pregnant women in a specific region, publicly released during the 2025 National College Students Mathematical Modelling Competition. It includes 267 pregnant women with a total of 1,082 NIPT data points. Data attributes comprise 31 columns including: - Participant ID - Gestational week at testing - Maternal BMI - Y chromosome concentration - Fetal health status Each participant has 1–7 NIPT data points. Fetal health status serves as the target feature. Selected original dataset information is presented in Table 1.

Table 1. NIPT Data for Pregnant Women

Feature Category	Feature Name
Basic Maternal Information (Partial)	Maternal ID
	Gestational Age at Testing
	Maternal BMI
	Age
	IVF Pregnancy
Target Feature	Y Chromosome Concentration

2.2 Data Preprocessing

Given that raw clinical data may contain human or systematic errors during collection, organization, and entry, systematic preprocessing of the dataset was undertaken to ensure the accuracy and robustness of subsequent modelling results. This included removing missing values, verifying duplicate entries, and detecting anomalous numerical features. Checks revealed 12 missing records in the 'Last Menstrual Period' column. As this variable is closely linked to gestational age calculation, interpolation to fill missing values could introduce structural bias. Consequently, samples

containing missing values were directly excluded. Duplicate entries were confirmed absent through repeated comparisons.

To further enhance data quality, the Z-score method was applied to identify and filter out anomalous numerical features. Detection revealed 144 records significantly deviating from the overall distribution. After deletion, 926 high-quality valid samples were retained. Subsequent secondary verification via box plots and probability distribution visualisations demonstrated stable distribution patterns, sound structural integrity, and no conspicuous outliers.

Furthermore, to satisfy the continuity requirement for regression model inputs, the "gestational week at examination" field underwent standardised numerical conversion. Textual expressions such as "12W+3" were transformed into directly computable continuous floating-point values (e.g., 12.43), structurally unifying the data into quantifiable analytical variables. The aforementioned preprocessing steps collectively form the foundational data assurance for model construction, providing reliable data dependencies for subsequent statistical inference and machine learning model training.

2.3 Non-linear Feature Analysis

This study employed a multi-level mixed linear model to conduct significance tests on feature variables including gestational age, the quadratic term of gestational age, and maternal BMI. This approach assessed the extent and direction of influence exerted by each factor on Y chromosome concentration. The model formula is as follows:

$$y_{log} = \beta_0 + \beta_1 J + \beta_2 J^2 + \beta_3 K + u_i + \varepsilon_{ij} \quad (1)$$

In the equation, y_log denotes the logarithmically transformed dependent variable representing the original Y chromosome concentration. β_0 is the fixed-effect intercept term reflecting the baseline level of the model. β_1, β_2, and β_3 represent the fixed-effect coefficients for gestational week J, the quadratic term J^2 for gestational week, and maternal BMI (denoted as K), respectively, characterising the average influence of these variables on y_log. u_i denotes the random effect term corresponding to the i^(th) pregnant woman, following a normal distribution with mean 0 and variance σ_{u^2}, accounting for individual differences between women; ε_{ij} represents the random error term, following a normal distribution with mean 0 and variance σ^2, reflecting random fluctuations within a single test.

The parameter results obtained via maximum likelihood estimation are presented in Table 2.

Table 2. Fixed Effects Parameter Estimation Results

Variable	Estimated coefficient	Standard error	Z-value	P-value	95% confidence interval
Intercept(β_0)	-1.877	0.322	-5.832	<0.001	[-2.508,-1.246]
Gestational Age(β_1)	-0.046	0.019	-2.465	0.014	[-0.082,-0.009]
Quadratic Gestational Age(β_2)	0.002	0.001	4.706	<0.001	[0.001,0.003]
Maternal BMI(β_3)	-0.022	0.009	-2.447	0.014	[-0.039,-0.004]

Table 2 analysis indicates that all fixed-effect parameters yielded P-values below 0.05, confirming statistically significant effects of gestational age's linear term, nonlinear term (quadratic term), and maternal BMI on y_log. Specifically: the quadratic term coefficient for gestational age is positive, which, combined with the linear term coefficient, reflects a non-linear acceleration in Y chromosome concentration growth. The negative coefficient for maternal BMI indicates that at the same gestational age, higher maternal BMI correlates with a tendency towards lower fetal Y chromosome concentrations.

This study employed the coefficient of determination [8] and root mean square error to assess the model's explanatory power. The model's $R^2 = 0.8158$ indicates it explains 81.58% of the variation in ylog, demonstrating good fitting performance. The RMSE = 0.1885 falls within a reasonable range for the research context, with predictive accuracy meeting analytical requirements.

2.4 Analysis of Population Heterogeneity

Addressing the quantitative requirements for traditional statistical methods in processing such heterogeneous data, this study established the construction of an ordinary least squares baseline linear regression model as the primary step. Empirical data feedback indicates that within an unstratified global dataset environment, the extremely poor performance metrics of single linear model fit and the disproportionately high root mean square error values have been confirmed, failing to meet the requirements for

clinically precise prediction.

Regarding this prediction failure phenomenon, its fundamental cause is attributed to the presence of multimodal data heterogeneity. Differences in maternal physical characteristics resulted in a complete fragmentation of physiological distribution patterns: within the low BMI group, a pronounced sensitivity to gestational age changes was evident, alongside an accelerated trend in cfDNA concentration elevation. Conversely, in the high BMI group, the concentration increase was observed to be sluggish or even plateauing, suppressed by the fat dilution effect. Attempting to fit two fundamentally distinct curves using the same global linear equation inevitably induces systematic generation of substantial biases at both extremes—namely, overly conservative predictions for low-risk individuals and overly optimistic predictions for high-risk individuals occurring simultaneously. As illustrated in Figure 1, the distribution of marginal residuals reveals distinct non-random patterns, visually corroborating the inadequacy of the global model in capturing these heterogeneous physiological dynamics.

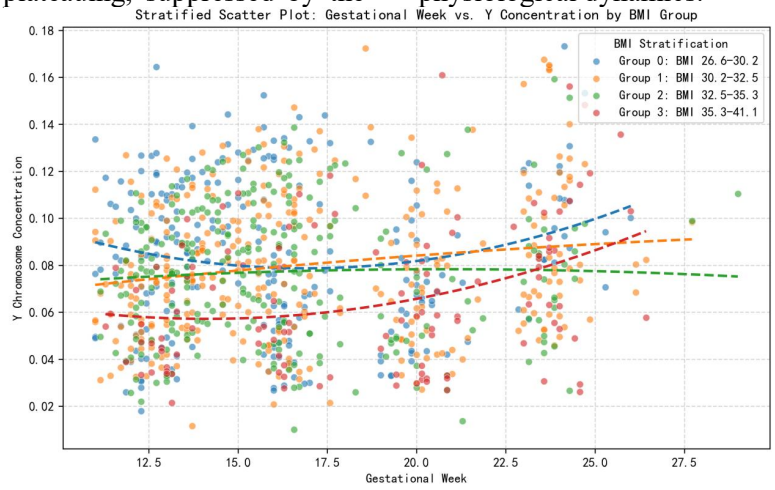

Figure 1. Distribution of Marginal Residuals after Removing Fixed Effects

Based on this premise, this study identifies the principle of divide and conquer alongside the establishment of non-linear enhancement strategies as key avenues for improvement. The primary step involves introducing unsupervised clustering algorithms to achieve objective stratification of populations. By clustering samples with similar constitutions into sub-groups, inter-group heterogeneity noise is effectively eliminated. Secondly, addressing the bottleneck of insufficient explanatory power from single variables, the adoption of machine learning algorithms and explicit incorporation of nonlinear feature terms are deemed essential measures. This strategy not only captures the marginal effects of individual indicators but also enables effective learning of higher-order interactions between maternal age, IVF, and BMI. Consequently, it lays the algorithmic foundation for subsequent high-precision predictive model construction tasks.

3. Research Methods

3.1 Overall Computational Framework

Given the presence of internal multimodal heterogeneity and non-linear physiological characteristics within NIPT clinical data, this paper proposes a data-driven, three-stage hybrid computational framework comprising "stratification-augmentation-simulation". The framework is designed to decouple the complex non-linear point-in-time prediction problem into solvable sub-problems. The overall processing workflow is illustrated in Figure 2.

Regarding specific details, the framework incorporates the following three core computational modules:

Implementation of heterogeneity resolution via unsupervised clustering: Given the significant differences in cfDNA release mechanisms among pregnant women of varying constitutions, the K-Means clustering algorithm is first applied to perform coarse-grained partitioning of the multidimensional sample space. Through mapping high-dimensional features to low-variance physiological subgroups (e.g., low-risk group, high-risk obesity group), global noise interference in predictive models is effectively mitigated, confirming the achievement of the "divide and conquer" strategy.

Feature-enhanced nonlinear fitting construction: Within each heterogeneous subgroup, feature-enhanced XGBoost regression models were constructed to capture physiological threshold effects unattainable by traditional linear models. Non-linear terms (e.g., quadratic terms for gestational age and BMI) and clinical covariates (age, IVF status) were explicitly incorporated. Leveraging the gradient boosting mechanism of ensemble tree models, precise fitting of cfDNA concentration baselines across subpopulations was achieved.

Robust decision-making based on Monte Carlo simulation: To address the uncertainty risks associated with single-point predictions in clinical applications, this study introduced Monte Carlo stochastic processes. Through 500 random perturbations of the prediction results, the population safety threshold was calculated at a 95% confidence level. This enabled the generation of a visualised, stratified optimal detection window lookup table, thereby providing direct guidance for clinical triage.

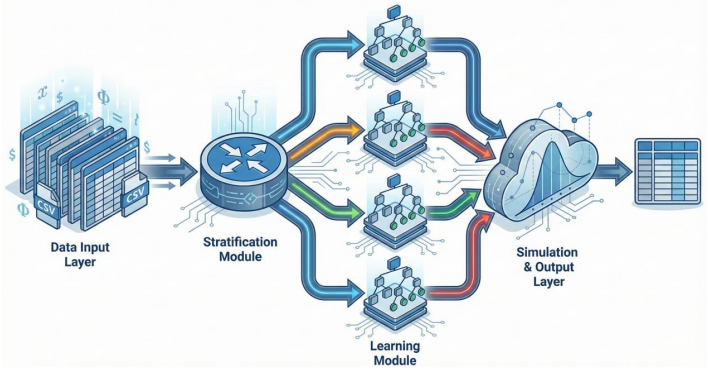

Figure 2. Three-stage Computational Workflow Architecture Diagram for "Stratification-Enhancement-Simulation"

3.2 Feature Engineering

To overcome the limitations of traditional linear models when handling heterogeneous biological data, this study first reconstructs the feature space of the raw clinical data. Based on the statistical analysis conclusions from Section 2, we constructed an expanded feature set X incorporating nonlinear and interaction terms. Specifically, beyond fundamental physiological indicators, this study introduced the following key features: Nonlinear Enhancement Terms: Constructed quadratic terms for BMI and gestational age to explicitly capture BMI's nonlinear inhibitory effect on concentration and the U-shaped dynamic variation of gestational age. Clinical covariate encoding: Assisted reproductive status (IVF) was binary-encoded (natural conception = 0, assisted reproduction/IUI = 1), while maternal age underwent cleaning and median imputation. The final feature vector input to the model is defined as follows:

$$x = [BMI, Y_Conc, Age, IVF\_Flag, BMI^2, Week^2] \quad (2)$$

The construction of this feature set transforms the original linear regression problem into a non-linear manifold learning problem in high-dimensional space, thereby establishing the data foundation for subsequent high-precision modelling.

3.3 K-Means-Based Clustering Architecture

Due to the pronounced multimodal heterogeneity observed in cfDNA release mechanisms across pregnant women with different body compositions, the adoption of a single global predictive model is insufficient for simultaneously accommodating both low-risk and high-risk populations. To address this limitation, a hierarchical routing module, as illustrated in Fig. 2, was designed, in which a K-Means clustering algorithm [9,10] was employed to implement a divide-and-conquer strategy for sample stratification.

Given that maternal body mass index (BMI) has been consistently identified as the dominant determinant affecting cfDNA concentration, BMI was selected as the sole feature for clustering. Accordingly, the entire sample space was partitioned into four constitution-based subgroups. The clustering process aims to determine an optimal partition that minimizes within-cluster variance, as defined by the following objective function:

$$\min_{\mu} \sum_{j=1}^{4} \sum_{i \in C_j} ||BMI_i - \mu_j||^2 \quad (3)$$

where μj denotes the centroid of the j-th

subgroup. After iterative convergence, samples were automatically routed into four statistically distinct BMI intervals. This stratification strategy effectively reduced intra-group variance and enabled downstream models to focus on learning physiological patterns specific to each body composition subgroup.

3.4 XGBoost Regression Modeling

XGBoost is a high-performance ensemble learning algorithm based on gradient-boosted decision trees[11]. For the i-th sample in the g-th subgroup, the model approximates the true earliest qualifying gestational week yi by aggregating the outputs of T regression trees, expressed as:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i), \quad f_t \in F \tag{4}$$

Compared with conventional gradient boosting decision tree (GBDT) methods, XGBoost incorporates a second-order Taylor expansion of the objective function along with explicit regularization terms, which substantially improves convergence efficiency and mitigates overfitting. The objective function at the t-th iteration is formulated as:

$$L^{(t)} \approx \sum_{i=1}^{n} [\, l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{5}$$

where l denotes the mean squared error loss; gi and hi represent the first-order gradient and second-order Hessian, respectively; and Omega denotes the regularization term.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{6}$$

Given the relatively limited sample size in Group 3, the regularization coefficient was intentionally increased to impose stronger constraints on leaf node weights. This adjustment prevents the model from memorizing noise and ensures robust generalization performance in small-sample, high-risk subpopulations.[12, 13]

3.5 Monte Carlo Simulation Framework

The NIPT workflow is inherently affected by uncertainty arising from sampling operations, instrument precision, and individual physiological fluctuations. Single-point predictions lack confidence interval information and are therefore insufficient for direct clinical decision support. To address this limitation, a Monte Carlo simulation framework was introduced[14].

Unlike conventional approaches that inject noise only at the output stage, the proposed framework explicitly simulates uncertainty throughout the entire pipeline, ranging from data acquisition to model inference. The simulation procedure was repeated 500 times and consisted of the following steps:

Input Noise Injection: Observed Y-chromosome concentration values were assumed to follow a Gaussian distribution. Random white noise was added to the original measurements to reflect the average level of clinical measurement error.

$$C_{sim} = \max(0, C_{obs} + \epsilon) \tag{7}$$

Dynamic Label Reconstruction: Based on the perturbed concentration trajectories, the theoretical earliest qualifying gestational week for each sample was recalculated, thereby constructing a synthetic training dataset.

Model Re-adaptation: The XGBoost model was retrained using the synthetic dataset, and predictions were generated to obtain an empirical distribution of the predicted gestational weeks.

Finally, the 95th percentile of the resulting distribution was selected as the recommended sampling time point. This conservative strategy ensures that even under unfavorable measurement error conditions, the recommended gestational week achieves a detection success probability of at least 95%, thereby maximizing the likelihood of successful one-time sampling.

3.6. Experimental Setup

To objectively evaluate the performance of the proposed framework, a rigorous experimental protocol was established.

Dataset Partitioning Strategy: Considering the class imbalance across different body composition groups, a stratified sampling strategy based on clustering was adopted. Within each of the four BMI-based subgroups identified by K-Means, samples were randomly divided into training and testing sets at a ratio of 7:3.

Hyperparameter Configuration: Hyperparameters were optimized using grid search. All XGBoost models across the four subgroups shared a unified optimal configuration, with the number of base learners set to n_estimators= 50, maximum tree depth max_depth= 3, and learning rate learning_rate= 0.05. This configuration achieved an optimal balance between bias and variance.

Evaluation Metrics: Model performance was

evaluated using root mean squared error (RMSE) and the coefficient of determination R^2, providing a comprehensive assessment of prediction accuracy and goodness of fit across different body composition groups.

4. Experimental Results and Analysis

4.1. Baseline Models and Performance Comparison

To validate the superiority of the XGBoost model in handling nonlinear and heterogeneous data, three widely used regression algorithms were selected as baseline models for comparative evaluation: linear regression (LR), support vector regression (SVR), and random forest (RF). Table 3 summarizes the performance comparison of all models across the four BMI-based subgroups on the test sets. Evaluation metrics include RMSE and R^2.

Based on the stratified sampling strategy defined in Section 3.6, Table 3 and Fig. 3 present the performance comparison of different models across the four BMI subgroups. The experimental results demonstrate clear heterogeneity in model adaptability when handling populations with distinct physiological characteristics.

Table 3. Performance Comparison of Different Models across BMI-Based Subgroups

BMI based subgroups	Model	RMSE	R^2
Group 0 (26.6-30.2)	LR	2.16	0.431
	RF	1.91	0.554
	SVR	2.15	0.433
	XGBoost	**1.89**	**0.562**
Group 1 (30.2-32.5)	**LR**	**3.28**	**0.446**
	RF	3.43	0.393
	SVR	3.54	0.352
	XGBoost	**3.30**	**0.438**
Group 2 (32.5-35.3)	**LR**	**2.41**	**0.611**
	RF	2.55	0.565
	SVR	2.54	0.570
	XGBoost	**2.44**	**0.600**
Group 3 (35.3-41.1)	LR	2.84	0.551
	RF	**2.36**	**0.691**
	SVR	3.34	0.379
	XGBoost	**2.46**	**0.663**

In Group 1 and Group 2, where physiological indicators fall within moderate ranges, linear regression exhibited stable predictive performance, with RMSE values of 3.28 and 2.41 weeks, respectively. The corresponding coefficients of determination were comparable to, or even slightly better than, those of certain nonlinear models. These findings indicate that, in populations with intermediate BMI levels, cfDNA concentration maintains a relatively strong linear relationship with gestational age, allowing traditional linear assumptions to adequately capture the underlying trend in routine prediction tasks.

However, in populations with extreme physiological characteristics, the advantage of nonlinear models became pronounced. In particular, within the clinically critical severely obese subgroup (Group 3), the explanatory power of linear regression deteriorated substantially, with an R^2 value of only 0.551. This suggests that a single linear equation is insufficient to model the concentration plateau caused by severe adiposity–related dilution effects. In contrast, ensemble tree–based models demonstrated superior robustness: both Random Forest and XGBoost achieved R^2 values exceeding 0.66 in this group, while reducing RMSE by approximately 15% compared with linear regression. These results provide strong evidence that, as BMI increases, nonlinear physiological mechanisms dominate cfDNA dynamics, necessitating the introduction of machine learning models to correct the systematic bias inherent in traditional methods for high-risk populations [15, 16].

From the perspective of overall generalization, although Random Forest showed competitive performance in certain subgroups, it exhibited instability in Group 1. In contrast, XGBoost benefited from the regularization term embedded in its objective function, which effectively constrained overfitting and enabled consistently high goodness-of-fit with low error variance across all BMI subgroups. Consequently, XGBoost was selected as the optimal baseline model, achieving a balance between computational efficiency in low-risk populations and predictive accuracy in high-risk groups[16].

4.2 Monte Carlo Simulation Analysis

To evaluate the robustness of the proposed model under realistic clinical measurement uncertainty, 500 Monte Carlo simulations were conducted following the procedure described in Section 3.5. The simulation results, summarized in Table 4 and Fig. 4, reveal pronounced stratification effects in the optimal sampling window across BMI subgroups.

The data indicate that, as BMI increases, the recommended optimal sampling time exhibits a distinctly nonlinear delayed trend. For Group 0, the model-recommended mean sampling time was 15.55 weeks, with a narrow 95% confidence interval width of only 0.61 weeks, suggesting relatively stable cfDNA release dynamics and high predictive certainty within this physiological range. In Group 1 and Group 2, the recommended sampling times clustered around 16 weeks, with noticeable oscillations; notably, the mean recommendation for Group 1 was slightly later than that for Group 2. This non-monotonic pattern highlights the complexity of the underlying biological mechanisms, indicating that interactions between BMI and covariates such as maternal age and IVF status may outweigh the effect of BMI alone in moderately obese populations. These findings further justify the necessity of introducing XGBoost to model nonlinear interactions.

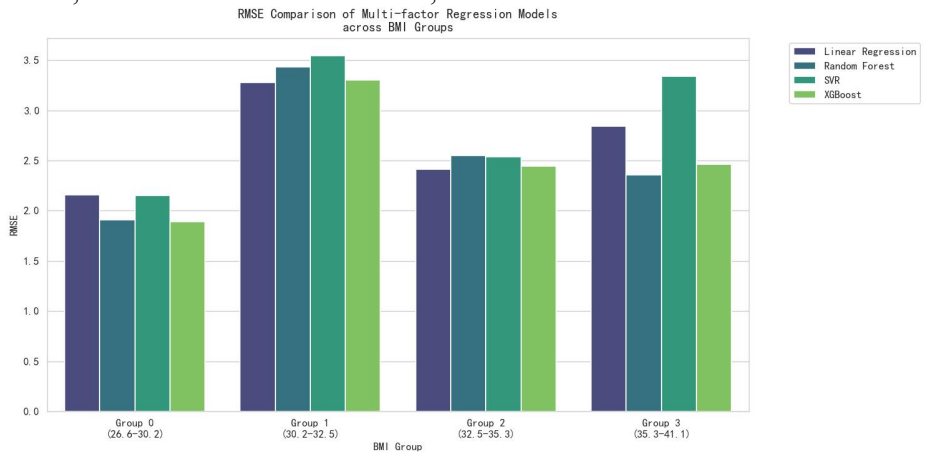

Figure 3. Comparison of RMSE Values for Different Models across BMI Groups

Table 4. Optimal NIPT Timing for Different BMI Groups

Group ID	BMI Range	Sample Size	Mean Predicted Gestational Week	Mean Observed Gestational Week	95% Confidence Interval
0	26.6-30.2	234	15.55	13.64	[15.26-15.87]
1	30.2-32.5	318	16.27	15.89	[15.88-16.72]
2	32.5-35.3	256	15.88	14.93	[15.57-16.25]
3	35.3-41.1	118	18.16	16.71	[17.55-18.85]

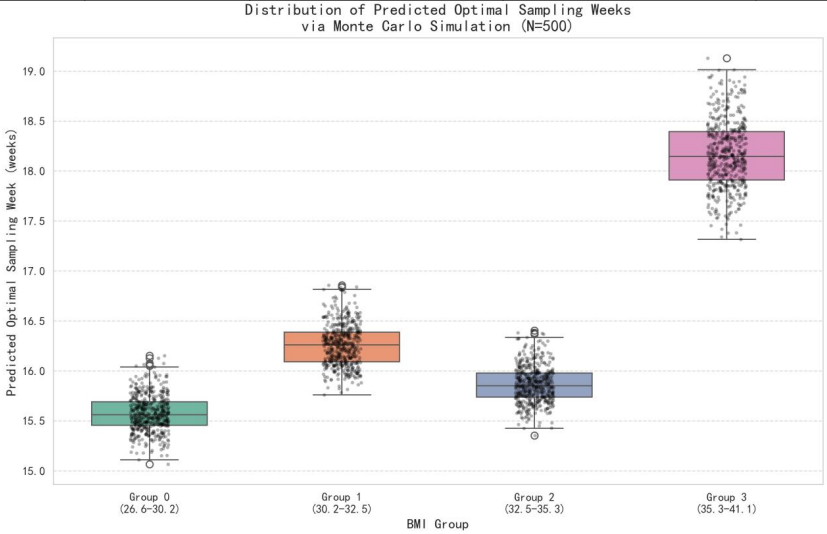

Figure 4. Boxplot of Monte Carlo Simulation–based Prediction Result Distributions

A notable qualitative shift was observed in Group 3. The recommended mean sampling time increased sharply to 18.16 weeks, representing a delay of nearly 2.3 weeks compared with Group 2. This pronounced lag provides strong evidence for the existence of a physiological "threshold effect" associated with severe obesity: once this threshold is exceeded, the time required for cfDNA concentration to reach the detection criterion increases exponentially. Meanwhile, the

width of the 95% confidence interval expanded to 1.30 weeks—more than twice that of Group 0—objectively reflecting substantial inter-individual metabolic heterogeneity within the severely obese population and indicating an elevated risk of failure for single-point predictions.

Based on the distributional characteristics revealed by the simulations, a conservative clinical decision strategy is proposed. For pregnant women with BMI > 35.3, routine adherence to the standard 12-week testing guideline is not recommended, as it carries a high risk of insufficient cfDNA concentration and false-negative results. According to the lower bound of the 95% confidence interval estimated by the model, it is advisable to postpone scheduled sampling until at least 18 weeks of gestation, or, when clinically feasible, to directly consider invasive diagnostic procedures. This strategy proactively avoids high-risk windows, thereby minimizing repeat sampling, reducing medical costs, and alleviating psychological burden for patients.

Furthermore, a cross-sectional comparison between the model-recommended sampling times and statistical averages derived from historical clinical data revealed a commonly overlooked statistical bias. Taking Group 0 as an example, the model-recommended sampling time (~15.6 weeks) was significantly later than the "empirical mean" (~13.6 weeks) calculated solely from samples that successfully reached the detection threshold. This discrepancy reflects a classic manifestation of survivor bias in retrospective analyses.

Traditional clinical statistics often include only early one-time successful cases while excluding samples that failed due to insufficient concentration, required repeated testing, or were ultimately abandoned. Such selection bias systematically underestimates the true average time required to reach the detection threshold. By introducing a conservative labeling strategy during data preprocessing—using the maximum observed gestational age among threshold-reaching samples as the lower bound—combined with Monte Carlo simulation, the proposed framework effectively corrects this bias. Although this adjustment renders the recommended sampling window numerically "later," it explicitly incorporates previously neglected difficult-to-detect cases, thereby providing decision support with broader population coverage and improved clinical safety compared with reliance on empirical means alone.

5. Conclusion

This study addresses the high failure rate of non-invasive prenatal testing (NIPT) associated with maternal heterogeneity by developing a data-driven prediction framework integrating K-Means stratification and feature-enhanced XGBoost modeling. Based on real-world clinical data and Monte Carlo simulation, several conclusions can be drawn.

First, nonlinear modeling proves effective in mitigating population heterogeneity. A clear nonlinear negative association between maternal BMI and the gestational time required to reach an adequate cfDNA fraction is observed, with a pronounced saturation effect in severely obese populations. Compared with conventional linear approaches, the proposed XGBoost-based model achieves an approximately 15% reduction in RMSE within high-risk subgroups and improves the coefficient of determination to above 0.66, indicating superior generalization capability for complex physiological data.

Second, the incorporation of Monte Carlo simulation enables correction of clinical statistical bias. The simulation results demonstrate a systematic upward shift in the recommended testing time with increasing BMI. More importantly, uncertainty quantification allows effective correction of survivorship bias inherent in historical datasets that exclude failed tests, resulting in a more conservative and clinically safer decision boundary than empirical averages.

Third, the stratified decision-support strategy facilitates more precise clinical management. The BMI-specific recommendation table generated by the model supports a "defensive postponement" strategy for extremely high-risk individuals (BMI > 35), thereby reducing the likelihood of false-negative results caused by premature sampling.

Despite its robustness, this study has limitations. The cohort was derived from a single geographic region, which may constrain external generalizability, and certain potentially relevant covariates, such as lipid metabolism indicators, were not included. Future work will focus on multi-center validation and the integration of additional biological features to further promote the transition of prenatal screening from

standardized workflows toward intelligent and precision-oriented decision support.

References

[1] S. Qin, H. Wang, Y. Liu, et al., "Performance evaluation of NIPT on 24 chromosomes in 118,969 pregnant women in Sichuan, China," J. Int. Med. Res., vol. 52, no. 9, p. 3000605241274584, 2024.

[2] C. Deng and S. Liu, "Factors Affecting the Fetal Fraction in Noninvasive Prenatal Screening," Front. Pediatr., vol. 10, p. 812781, 2022.

[3] L. Qiao, Q. Zhang, Y. Liang, et al., "Sequencing of short cfDNA fragments in NIPT improves fetal fraction with higher maternal BMI and early gestational age," Am. J. Transl. Res., vol. 11, no. 9, pp. 4450–4459, 2019.

[4] M. Zaki-Dizaji, M. Akbari, K. Kamali, et al., "Maternal and fetal factors affecting cfDNA fraction in prenatal screening: a systematic review," J. Reprod. Immunol., vol. 160, p. 103533, 2023.

[5] Y. Hou, D. Lv, Y. Lai, et al., "Factors affecting cell-free DNA fetal fraction: statistical analysis of 13,661 maternal plasmas for NIPT," Hum. Genom., vol. 13, no. 1, p. 11, 2019.

[6] H. W. Loh, C. P. Ooi, S. Seoni, et al., "Application of explainable artificial intelligence in healthcare: A systematic review of the last decade (2011–2022)," Comput. Methods Programs Biomed., vol. 226, p. 107161, 2022.

[7] H. B. Lee, H. S. Lee, S. Kim, et al., "Development and performance evaluation of an artificial intelligence algorithm for non-invasive prenatal testing using cell-free DNA fragment distance," Front. Genet., vol. 13, p. 999587, 2022.

[8] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genom., vol. 21, Art. no. 6, 2020.

[9] S. P. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[10] B. Zhou, B. Lu, and S. Saeidlou, "A hybrid clustering method based on the several diverse basic clustering and meta-clustering aggregation technique," Cybern. Syst., vol. 54, no. 3, pp. 1–27, 2022.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[12] F. Rahmayanti, A. Pradana, B. M. W. Budiman, et al., "Comparison of machine learning algorithms for classification of fetal health using cardiotocogram data," J. Inf. Syst. Eng. Bus. Intell., vol. 8, no. 1, pp. 22–32, 2022.

[13] A. Ogunleye and Q. G. Wang, "XGBoost model for chronic kidney disease diagnosis," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 17, no. 6, pp. 2131–2140, 2020.

[14] J. H. Jones and N. Fleming, "Simulation with Monte Carlo methods to focus quality improvement efforts on interventions with the greatest potential for reducing PACU length of stay: a cross-sectional observational study," BMJ Open Qual., vol. 13, no. 4, p. e002933, 2024.

[15] Y. Cao, M. P. Forssten, B. Sarani, et al., "Development and Validation of an XGBoost-Algorithm-Powered Survival Model for Predicting In-Hospital Mortality Based on 545,388 Isolated Severe Traumatic Brain Injury Patients from the TQIP Database," J. Pers. Med., vol. 13, no. 9, p. 1401, 2023.

[16] S. M. Lundberg, B. Nair, M. S. Vavilala, et al., "Explainable machine learning predictions for the prevention of hypoxaemia during surgery," Nat. Biomed. Eng., vol. 2, no. 10, pp. 749–760, 2018.

