Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 5, No. 2, 2025

A Model for Detecting Fetal Chromosomal Abnormalities Based on Quality-Weighted Regression and K-Means Clustering

Yanxiao Wang, Lihao Zhang and Weihao Dong

Brunel London School, North China University of Technology, Beijing, China

Abstract: Non-invasive prenatal testing (NIPT) screens for abnormalities in chromosomes 21, 18, and 13 by analyzing fetal cell-free DNA in maternal blood. Conducted between 10 and 25 weeks of gestation, it aids in early assessment of fetal health, with accuracy contingent upon male fetuses exhibiting Y chromosome concentrations ≥4% and female fetuses displaying normal X chromosome concentrations. This paper focuses on constructing a data model for NIPT data preprocessing. A quality-weighted linear regression model was developed to establish associations between fetal Y concentration and gestational age/BMI, with statistical significance verified. Dual-breakpoint K-means clustering was used to initialize centers for BMI classification in male fetus pregnancies. Risk minimization was applied to determine the optimal NIPT timing while analyzing errors. Feature reconstruction fused LASSO regression with decision trees to establish a method for detecting female fetal chromosomal abnormalities. Final results demonstrate that this model can accurately quantify correlations between indicators, effectively capture the nonlinear relationship between gestational age and Y concentration as well as sequencing quality interference. The optimal timing for BMI groups is earlier, with high BMI groups requiring consideration of height and age. Simultaneously, it enables scientifically accurate detection of female fetal chromosomal abnormalities.

Keywords: K-means clustering, BMI grouping, dual-dependent variable regression, chromosomal abnormality assessment

1. Introduction

Non-invasive Prenatal Testing (NIPT) [1] is a diagnostic technique that aims to assess fetal health by collecting maternal blood to detect fetal cell-free DNA fragments, subsequently analyzing them for chromosomal abnormalities. Examples include Down syndrome, Edwards syndrome, and Patau syndrome caused by abnormal concentrations of chromosomes 21, 18, and 13, respectively. The accuracy of NIPT is influenced by fetal sex chromosome concentrations. Testing for sex chromosome concentrations is recommended between 10 and 25 weeks of gestation. Results are generally reliable when the male fetus's Y chromosome concentration is ≥4% and the female fetus's X chromosome concentration is within normal range. Furthermore, the earlier fetal abnormalities are detected, the lower the associated risks. The concentration of the male Y chromosome is closely related to the gestational age and BMI of the pregnant woman. In practice, NIPT timing is often determined by grouping based on BMI. However, due to individual differences in maternal age, BMI, and pregnancy status, simple empirical grouping and uniform testing timepoints can affect accuracy. Additionally, sequencing failures may occur during testing. Some pregnant women undergo multiple blood draws or repeated testing from a single sample to enhance result reliability. This study investigates issues arising from these processes [2].

2. Model Assumption

Assumption 1: Detection data exhibits no systematic error. It is assumed that all samples undergo testing procedures adhering to uniform standards with consistent equipment precision, including processes such as blood sampling, gene sequencing, and read alignment. Experimental data—including duplicate read proportions, raw read counts, and alignment rates on reference genomes—contains no batch-related biases or errors stemming from operational mistakes.

Assumption 2: Filtered reads represent random error. It is assumed that the proportion of filtered reads is solely attributable to random technical errors such as low-quality reads or non-specific binding. This does not systematically affect chromosome copy number calculations or aneuploidy detection, being independent of chromosome characteristics or sample contamination.

Assumption 3: Short-term stability of maternal physiological indicators. It is assumed that when maternal control variables (e.g., height, age) remain constant, fluctuations in weight/BMI across different gestational weeks reflect normal pregnancy growth. The impact of such fluctuations on chromosomal test results can be quantified by the gestational week variable, with no unmeasured nonlinear interactions (e.g., Z-score, Y-chromosome concentration).

Assumption 4: Minimal impact of unmeasured confounding factors. It is assumed that other unmeasured factors have negligible influence or are randomly distributed across samples. These include maternal lifestyle habits and underlying medical conditions affecting fetal chromosomal status and health, which do not cause systematic bias. This excludes factors such as age, height, weight, BMI, gestational week at testing, number of pregnancies, and number of births.

Assumption 5: Sensitivity remains stable within the tested gestational age range. It is assumed that within the gestational age range covered by the study, increasing gestational age does not cause a systematic increase/decrease in the rate of missed aneuploidies or false positives. That is, the sensitivity and specificity of the chromosomal testing technology do not change significantly, and the relationship between gestational age and testing accuracy can be quantified using a linear or piecewise model.

3. Condition Tests of Rougher Flotation

3.1. Data Preprocessing

Based on the clinical rationale and data characteristics of NIPT testing, reliable and interpretable datasets were selected for modeling while mitigating sequencing errors and individual variations that could confound results. Therefore, the provided data underwent preprocessing. First, sample-level filtering retained only male fetal samples, as female fetuses lack Y chromosome data and thus require no inclusion in analysis. Simultaneously, adhering to NIPT clinical testing requirements, samples were filtered to meet the gestational age range of 10-25 weeks, specifically retaining those with gestational age more than 10 weeks. Samples with missing key indicators-including gestational age, Y chromosome concentration, BMI, GC content, and filtered read proportion-were excluded to ensure data integrity.

To eliminate the impact of data format differences on analysis results, core indicators undergo standardization processing. The "gestational week + days" format is uniformly converted to the "gestational week" numerical format; Y chromosome concentration data in the 0-1 ratio range is converted to percentage form; GC content is converted to percentage format; and the filtered read proportion remains in the 0-1 format. Simultaneously, differentiated quality control was implemented based on sequencing quality requirements. For samples with gestational age ≤24 weeks, GC content within the normal range of 40%-60% was selected, as abnormal GC content may lead to sequencing quality issues. For samples with a filtered read proportion < 0.20, no additional quality filtering conditions were set for samples with gestational age > 24 weeks, considering the limited number of later-stage samples. This avoids excessive data reduction that could compromise model reliability. Z-scores outside the normal range (±3) were flagged as outliers. For example, Z=3.616 at position 13 in A029 was labeled "Abnormal at position 13," with the original value retained to prevent information loss. BMI anomalies were grouped per medical standards: normal (18.5–23.9), overweight (24–27.9), obese (≥28). Original values were retained with added classification labels, and extreme values were not deleted. IVF pregnancy mode converted to binary: 1 = IVF, 0 = natural conception. Additionally, male fetus group added gestational week × BMI interaction term and Y chromosome concentration compliance status; female fetus group added Z-score deviation correction term and chromosomal abnormality flag.

Next, to address sequencing failures, we quantified and graded sequencing quality. A sample quality score ranging from 0 to 1 was constructed based on three core metrics: GC content, proportion of uniquely mapped reads, and proportion of duplicate reads. The normal range for GC content is 40%-60%, which was used to label gc-status (1 = abnormal, 0 = normal). A proportion of uniquely mapped reads below 70% was considered inefficient mapping, labeled as map-status (1 = inefficient, 0 = efficient). A repeat read proportion exceeding 20% was deemed excessive, labeled as repeat-status (1 = excessive, 0 = normal).

Finally, regarding missing value handling: when either height or weight is missing, samples must be directly excluded since the core grouping metric BMI = weight / height² cannot be calculated. For minor age missingness in pregnant women (missing rate <5%), the missing values are imputed with the mean age of the corresponding BMI group. GC content is a mandatory quality metric, with missing values primarily caused by sequencing failures. Thus, missing GC content is flagged as "GC Abnormal." Samples with a single missing chromosome Z-score are discarded, as the Z-score is the core determinant for detecting chromosomal aneuploidy; missing this core metric renders the sample invalid. Missing fetal health results are marked as "Pending Verification" and excluded from model validation. For multiple test data, deduplication and integration are required to avoid redundancy. Direct integration retains the highest-quality sample from duplicate batches; if quality scores are identical, the sample with Y or X concentration closest to the batch mean is retained. Cross-batch duplicate samples are not merged directly; the "first test + last test" data points are retained. The results of the data validation after preprocessing are shown in Figure 1.

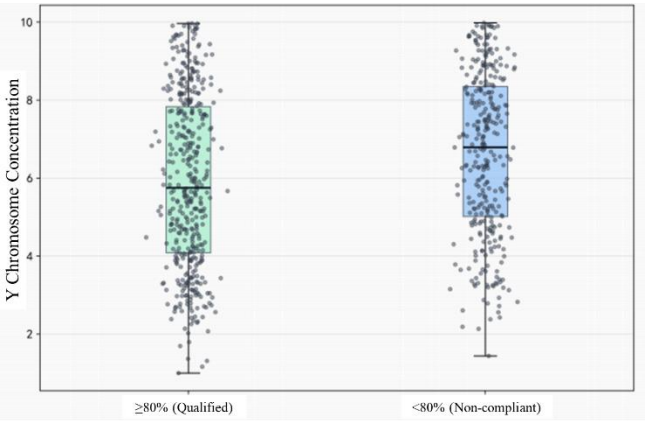

Figure 1. Y Concentration Box-and-Whisker Plot

Figure 1 shows the Y chromosome concentration distributions for the two sample groups with sequencing quality meeting standards (≥80%) and failing to meet standards (<80%). The qualified group exhibits a more concentrated Y-chromosome concentration distribution with higher overall levels, while the unqualified group shows a more dispersed distribution with a higher prevalence of low concentration values. This clearly demonstrates the correlation between sequencing quality and Y-chromosome concentration: achieving qualified sequencing quality facilitates obtaining higher and more stable Y-chromosome concentrations.

3.2. Dual-Dependent Variable Regression Model Based on K-Means Clustering

Achieving the required Y chromosome concentration (≥4%) in male fetuses is critical for accurate NIPT results. This serves as the core prerequisite for assessing how individual maternal factors—such as age, BMI, and height—impact test accuracy. Standardized height (H_{std}) must be derived:

$$H_{std} = \frac{H - \bar{H}}{S_H} \tag{1}$$

Where H represents the mean height of all samples, and S_H denotes the standard deviation of height, eliminating the influence of units on the model. Additionally, calculate characteristics such as the proportion of Y chromosome concentrations meeting standards to establish an initial comprehensive risk indicator. This indicator combines gestational week risk levels—such as low risk in early weeks

(10–12 weeks) and higher risk in mid-pregnancy (13–25 weeks)—with detection error values corresponding to sequencing quality to quantify individual baseline risk.

In practice, applying simple empirical grouping and uniform testing timepoints to all pregnant women significantly impacts accuracy, and BMI is confirmed as the primary factor influencing the timing of achieving Y chromosome concentration targets. Therefore, in the BMI grouping phase, this study employs a "dual-breakpoint + K-means [3] clustering" approach. First, male fetal samples were sorted in ascending order by BMI. Dual inflection points—defined as a ≥10% decrease in Y concentration compliance rate and a ≥0.2 increase in comprehensive risk value—served as initial K-means clustering centers. Post-clustering validation required stable accuracy and risk levels within groups alongside significant intergroup differences. This approach enabled preliminary BMI interval delineation, with samples in different intervals exhibiting distinct core characteristics in compliance rates and risk levels.

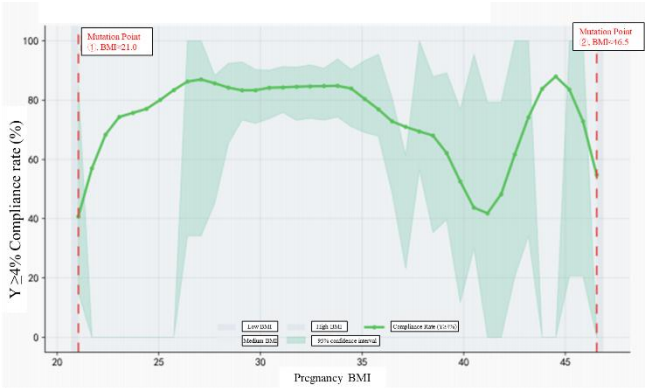

Figure 2. BMI Grouping Explanation: Double Mutation Site + Confidence Band

As shown in Figure 2, two key mutation points are clearly marked at BMI≈21.0 and BMI≈46.5. These points divide pregnant women's BMI into low, medium, and high BMI ranges, with each range distinguished by a different background color. The curve visually illustrates the trend of compliance rates as BMI varies. The "95% confidence interval" band demonstrates the statistical reliability of the data, eliminating interference from random factors.

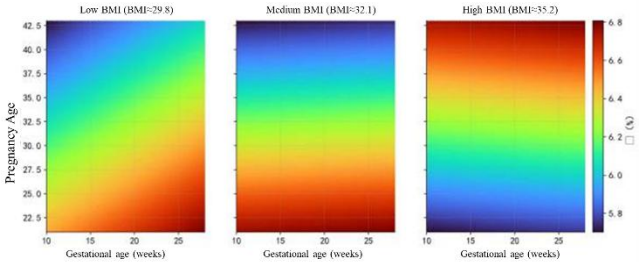

Figure 3. Each BMI group: Heatmap with = 4% contour lines

As shown in Figure 3, the low BMI group achieved target levels within the isocline, covering a broader age range and reaching targets 10 to 12 weeks earlier in gestational age. The moderate BMI group achieved compliance slightly later, starting at 12–13 weeks. The high BMI group only gradually met the target after 14 weeks, with the latest compliance gestational age. Higher age ranges predominantly show non-compliant colors, indicating that compliance becomes increasingly difficult with advancing maternal age.

The time required for male fetuses to achieve adequate Y chromosome concentration is influenced by multiple factors such as height, weight, and age. A single-variable model cannot capture the interactions between these factors. Therefore, a dual-dependent variable regression model incorporating BMI grouping and individual characteristic interaction terms was constructed. Simultaneously predicting both Y chromosome concentration and composite risk scores captures the differential effects of age and height within different BMI groups. The model structure, exemplified by the g-th group, is as follows:

$$\begin{cases} \hat{C} = \beta_{g0} + \beta_{g1} BMI + \beta_{g2} A + \beta_{g3} H_{std} + \beta_{g4} \left(BMI \cdot A \right) + \beta_{g5} \left(BMI \cdot H_{std} \right) + \varepsilon_1 \\ \hat{R} = \gamma_{g0} + \gamma_{g1} BMI + \gamma_{g2} A + \gamma_{g3} H_{std} + \gamma_{g4} \left(BMI \cdot A \right) + \gamma_{g5} \left(BMI \cdot H_{std} \right) + \varepsilon_2 \end{cases} \tag{2}$$

Where $\hat{C}$ represents the predicted Y chromosome concentration (target ≥ 4%), $\hat{R}$ denotes the overall risk prediction (lower values are better), A represents maternal age, $\beta_{g0} - \beta_{g5}$ and $\gamma_{g0} - \gamma_{g5}$ are regression coefficients for group g, and ε_1 and ε_2 are error terms. The model must also pass the t-test, specifically the goodness-of-fit test where individual coefficients have $p < 0.05$ and adjusted $R^2 \geq 0.6$, ensuring that BMI, age, height, and interaction terms are significantly associated with the prediction results and thatthe model has sufficient explanatory power.

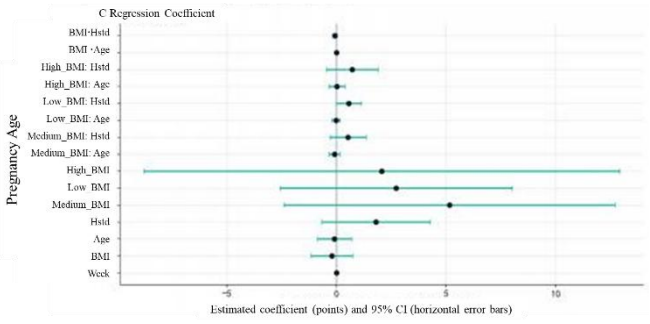

Figure 4. Multivariate Regression: Key Factors and Interaction Effects

As shown in Figure 4, among the key factors, gestational age exhibits the highest coefficient of approximately 8, indicating it is the primary positive factor influencing concentration levels. The BMI coefficient is -2, suggesting that higher BMI correlates with lower concentrations. Among the interaction terms, the interaction coefficient between BMI and age is -1.5, meaning that pregnant women with higher BMI and older age experience slower concentration increases. The interaction coefficient between the high BMI group and

height is positive 3, indicating that among pregnant women with high BMI, taller individuals exhibit a mitigated decline in concentration.

The design of personalized timing screening rules strictly adheres to the clinical risk classification: detectable between 10-25 weeks gestation, with lower risk in the early period (within 12 weeks) and higher risk in the mid-period (13-27 weeks). Optimal timing is selected based on the "personalized risk minimization" principle. Priority is given to screening during the early period (10-12 weeks) for the earliest gestational week meeting $\hat{C}_t \geq 4\%$ and $\hat{R}_t \leq 0.3$. If no qualifying early-pregnancy time point exists, select the mid-pregnancy week (13–25 weeks) with $\hat{C}_t \geq 4\%$, the lowest $\hat{R}_t$, and ≥85% Y concentration compliance.

3.3. Multi-Model Fusion Based on Lasso, Logistic Regression, and Decision Trees

To address the issue that Z-scores for chromosomes with significantly different lengths, such as chromosomes 21, 18, and 13, are not directly comparable, this paper calculates chromosome-specific correction features. Using the formula:

$$Z_{corr,i} = Z_i \times \frac{L_i}{\bar{L}} \tag{3}$$

Where L_i represents the actual length of chromosome i, and known biological data indicate that chromosome 21 is almost 48 Mb, chromosome 18 ≈ 80 Mb, and chromosome 13 ≈ 114 Mb, with $\bar{L}$ denoting the average length of these three chromosomes at approximately 80.7 Mb. This eliminates Z-score bias caused by chromosomal length differences, enabling comparative analysis of abnormal signals across different chromosomes.

Second, to address the interference caused by directly using abnormal GC values that reduces model accuracy, this paper constructs a sequencing quality-weighted feature. Specifically, samples with poor sequencing quality are assigned higher deviation weights for GC anomalies, highlighting their disruptive impact on classification results. The formula is:

$$GC_{diff,i} = \begin{cases} |GC_i - 0.5| \times 1.5 & GC_i \notin [0.4, 0.6] \\ |GC_i - 0.5| & GC_i \in [0.4, 0.6] \end{cases} \tag{4}$$

Simultaneously, to enhance the utilization rate of valid reads, it is necessary to combine the proportion of filtered reads with the number of uniquely mapped reads to correct the interference of low-quality sequencing data on the determination. Furthermore, Problems 2 and 3 have confirmed that BMI affects the Y chromosome concentration in male fetuses, while in female fetuses, BMI may interfere with X/autosome signals by influencing the proportion of maternal cell-free DNA. This study introduces a BMI-stratified interaction feature, $BMI - Z_i = BMI \times Z_{corr,i}$, calculating correlation coefficients between Z-scores and X chromosome concentration for each chromosome across BMI groups to capture BMI's differential impact on chromosomal signals in female fetuses.

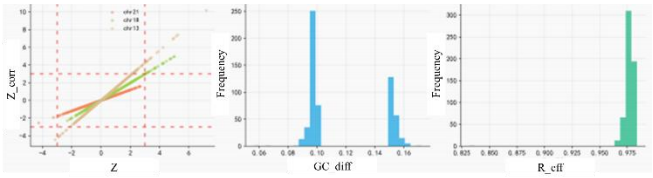

Figure 5. Feature Reconstruction and Quality Characterization

In Figure 5, the left panel shows the changes in chromosome Z-scores before and after correction. It can be observed that after correction, the Z-scores for different chromosomes exhibit greater discriminatory power and a clear range of abnormal detection thresholds. The middle panel displays the GC deviation distribution, with most samples concentrated in a smaller deviation range and a few showing significant deviation, reflecting the overall qualityof sequencing GC content. The right panel depicts the read segment utilization distribution. Most samples exhibit high read segment utilization, indicating sufficient effective sequencing data, though some samples show slightly lower utilization.

First, clinical a priori constraints are set, such as establishing a lower bound for the coefficient of Z-corr on chromosome 21. This is because abnormalities on chromosome 21 represent a clinically critical type of malformation requiring prioritized accuracy in detection. Within the multi-model fusion framework, LASSO regression effectively addresses issues of high feature dimensionality and redundancy, preventing overfitting.

The key mechanism of Lasso regression lies in applying L1 regularization to regression coefficients, causing some coefficients to approach zero and thereby enabling feature selection. This is particularly crucial in scenarios involving high-dimensional data or numerous features, as it helps eliminate irrelevant features, enhancing model interpretability and generalization capability. Its formula is:

$$\min_{\beta_0,\beta} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{5}$$

Among these, y_i is the target variable (response variable), x_{ij} is the input feature, β_j is the regression coefficient for each feature, and λ is the regularization parameter used to control the model's complexity. A larger λ will eliminate more features (setting their corresponding regression coefficients to zero) [4].

This paper introduces lower bound constraints on coefficients for clinically significant features, such as chromosome 21 Z-corr and chromosome 18 Z-corr, within the logistic regression loss function. For instance, setting $|\beta_{21}| \geq 0.8|\beta_{else}|$ ensures the model prioritizes clinically critical chromosomes while preventing feature importance inversion caused by data noise. The core feature set includes 21-Z-corr, 18-Z-corr, 13-Z-corr, X-Z, GC-diff-21, R-eff-21, BMI-Z-21, and the preliminary probability P1. Subsequently, clinical thresholds are embedded into decision trees to enhance the clinical interpretability of results. Z-score analysis serves as the core method for chromosomal aneuploidy detection. By annotating decision bases at leaf nodes, interpretability is enhanced, avoiding the clinical inapplicability of black-box models. Therefore, the root node is restricted to selecting from the Z-corr values of

chromosomes 21, 18, and 13. The split threshold references the standard clinical range of Z-value ±3. Leaf nodes annotate the basis for determination, outputting a refined determination probability P2 and interpretable determination rules.

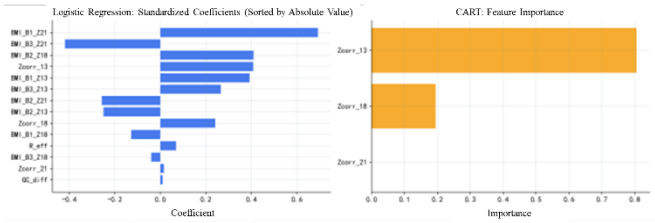

Figure 6. Comparative Analysis of Interpretability in Two-Layer Models

In Figure 6, the left panel displays the standardized coefficients for the logistic regression model. The absolute values of coefficients for different features—such as the interaction term between BMI group and chromosome Z-score, and the corrected Z-scores for each chromosome—exhibit significant differences, reflecting the varying degrees of influence these features exert on anomaly detection. In the CART feature importance plot on the right, the corrected Z-score for chromosome 13 (Zcorr-13) exhibits the highest importance, followed by Zcorr-18 for chromosome 18, while Zcorr-21 for chromosome 21 shows relatively lower importance.

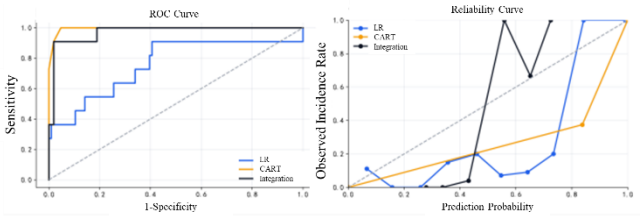

Figure 7. Fusion outperforms single models

In Figure 7, on the left ROC curve, the curve of the ensemble model is closer to the upper left corner, with a higher AUC value than logistic regression and CART. On the right reliability curve, the observed incidence rate of the ensemble model aligns more closely with the predicted probability, whereas logistic regression and CART show significant deviation.

Finally, dynamic weighting enhances robustness. For samples with multiple detections, the coefficient of variation (CV) of core features is calculated. A lower CV indicates more stable detection, resulting in higher weight assignment. Ultimately, the final weighting is determined by

$$P_{final} = w_1 P_1 + w_2 P_2, w_1 + w_2 = 1 \quad (6)$$

Output the determination results, with unstable samples marked as "pending review" to minimize misclassification.

4. Conclusion

The data processing in this model aligns with NIPT testing practices, effectively mitigating the impact of abnormal or unreasonable data. For sequencing quality, it addresses sample quality score calculation, quantifies GC anomaly values, and filters out excessive re-reads in multiple segments. Low-quality samples are filtered using quality-weighted linear regression and stratified error simulation. The model comprehensively captures multifactorial and interaction effects, demonstrating strong adaptability to individual variations. Building upon BMI, this study incorporates standardized height and age to construct a dual-dependent variable regression model with interaction terms, effectively capturing the additive effects of BMI × age and BMI × height. However, limitations exist, such as restricted sample adaptability and room for improvement in generalization capability. The sample size for low-BMI pregnant women (BMI < 20 kg/m²) is extremely small. Models related to BMI may exhibit increased prediction bias in low-BMI populations and cannot be generalized to broader BMI distributions [5].

References

[1] Yu Dandan, Li Fengjin, Yao Xinyu, et al. Efficacy Analysis of NIPT in Detecting Common Fetal Chromosomal Aneuploidy in Different High-Risk Populations [J]. Chinese Journal of Family Planning and Obstetrics & Gynecology, 2025, 17(03): 72-76+82.

[2] Zhang R, Zhang H, Zhang L, et al. Clinical indications and Z-score-assisted NIPT testing: a new perspective in prenatal screening [J]. Gynecology and Obstetrics Clinical Medicine, 2025, 5(01):42-50.

[3] Si Shoukuai, Sun Xijing. Mathematical Modeling Algorithms and Applications [M]. Beijing: National Defense Industry Press, 2011.

[4] Li Jianxin. Research on Optimizing Initial Stress Field Inversion Based on the LASSO-OLS Method [M]. [Place of Publication Not Specified]: School of Water Resources and Hydropower Engineering, Hohai University, 2025.

[5] Zhuo Jinwu. Applications of MATLAB in Mathematical Modeling [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2011.