Advances in Computer, Signals and Systems (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2026.100118
ISSN 2371-8838 Vol. 10 Num. 1

Optimization of NIPT Timing Using Random Forest and Hierarchical Clustering

Weiyi Zhu[1,a,*], Can Wu[1,b], Hongxun Ye[1,c], Shengjun Xu[1,d]

[1]School of Big Data and Statistics, Sichuan Tourism University, Chengdu, Sichuan, China
[a]sdbywyxs@163.com, [b]xcan_2024@qq.com, [c]zhenshuai6@qq.com, [d]3356125835@qq.com
**Corresponding author*

Keywords: Gradient Boosting Trees; K-means Clustering; Random Forest; Hierarchical Clustering; Monte Carlo Simulation; NIPT Timing Optimization

Abstract: In this study, a three-step progressive algorithm framework is constructed to realize the individualized optimization of NIPT time points. In the first step, based on polynomial regression and gradient lifting tree, the nonlinear relationship between Y chromosome concentration, gestational age and BMI was revealed, and the gradient lifting tree was optimally fitted (test set R^2=0.2521). In the second step, K-means clustering and regression trees were used to adaptively group male BMI, and the optimal NIPT time points of each subgroup were determined by Monte Carlo simulation error analysis, and the optimal grouping intervals were [26.62, 29.90], [29.96, 32.19], [32.26, 34.93], and [35.06, 39.14], and the corresponding time points were 16, 19, 17, and 24 weeks, respectively, and the confidence interval width was less than 1.2 weeks. In the third step, multiple factors such as height, weight, and age were introduced, and a random forest regression model was constructed to predict the gestational age of the first standard (test R^2=0.942) by introducing multiple factors such as height, weight, and age, and the optimal time points of the four groups were 16.8, 18.8, 18.9, and 24.0 weeks, respectively. The algorithm chain integrates nonlinear regression, cluster integration and random forest, which provides a data-driven quantitative basis for NIPT point-in-time decision-making.

1. Introduction

Birth defect prevention and control is a key link in improving the health level of China's population. Non-invasive prenatal testing (NIPT) has become the mainstream technology for prenatal screening due to its advantages such as non-invasiveness and high sensitivity [1] [2]. However, in clinical practice, the selection of NIPT testing time is mostly dependent on experience, and early detection is prone to false negatives due to insufficient fetal cell-free DNA concentration, and late detection may delay the clinical intervention window, increasing the psychological burden and medical costs of pregnant women [3]. Fetal chromosome concentration (especially male Y chromosome concentration needs to reach more than 4%) is closely related to gestational age, BMI and other indicators [4][5], but most existing studies are based on linear assumptions and ignore the complex interaction between multiple variables and nonlinear effects. Therefore, how to establish a high-precision prediction model based on data-driven methods and determine the optimal detection

time for different BMI groups has become a key scientific issue for accurate prenatal screening.

In recent years, machine learning has shown great potential in biomedical data analysis. Shi et al. (2021) used random forests to predict fetal DNA concentrations and verified the ability of the ensemble model to capture nonlinear features [6]. Wang et al. (2022) used a gradient lifting algorithm to optimize the prenatal screening window, which significantly improved the detection efficiency [7]. The application of K-means clustering and hierarchical clustering in the grouping of pregnant women has also attracted increasing attention [8][9]. In addition, Monte Carlo simulations are widely used to evaluate the impact of detection errors on clinical decision-making [10]. However, there are few existing studies that systematically integrate nonlinear regression, clustering grouping, and multivariate time series prediction to construct a full-chain model of NIPT time-based optimization. In this paper, a three-step recursive algorithm framework is designed: In the first step, a quantitative relationship model between Y chromosome concentration and gestational age, BMI and other indicators is established, and linear, polynomial, and multiple machine learning models (random forest, gradient lifting, SVM, etc.) are compared to reveal the key influencing factors and test the significance [11]. In the second step, the BMI of male pregnant women was rationally grouped, and the grouping intervals were optimized by using K-means clustering and regression trees, and the influence of errors on the optimal NIPT time was detected by Monte Carlo simulation analysis to ensure the reliability of the scheme [12]. In the third step, multi-dimensional features such as height, weight, and age are further incorporated, and the improved K-level clustering and feature weighting are used to group to construct random forest regression prediction of the gestational age for the first time, determine the optimal detection time point for different subgroups, and realize individualized time point recommendation. This study aims to provide algorithmic support for clinical NIPT time point selection, reduce potential risks for pregnant women, and promote the development of precision prenatal screening.

2. Methods

2.1. Nonlinear Regression Model

Any text or material outside the aforementioned margins will not be printed.

In order to reveal the quantitative relationship between Y chromosome concentration and gestational age, BMI and other indicators, the data were preprocessed: samples with GC content not in the range of 40%-60% were excluded, and the outliers were eliminated by box plot, and finally high-quality data were retained. In view of the significant nonlinearity between variables, three sets of progressive experimental construction models were designed.

Experiment 1: Polynomial regression model. The linear model, quadratic polynomial model, and cubic polynomial model are constructed respectively in the form of the following forms:

$$Y = \beta_0 + \beta_1 \cdot GA + \beta_2 \cdot BMI + \text{ò} \tag{1}$$

$$Y = \beta_0 + \beta_1 GA + \beta_2 BMI + \beta_3 GA^2 + \beta_4 BMI^2 + \beta_5 GA \cdot BMI + \text{ò} \tag{2}$$

$$Y = \beta_0 + \beta_1 GA + \beta_2 BMI + \beta_3 GA^2 + \beta_4 BMI^2 + \beta_5 GA \cdot BMI + \beta_6 GA^3 + \beta_7 BMI^3 + \beta_8 GA^2 \cdot BMI + \beta_9 GA \cdot BMI^2 + \text{ò} \tag{3}$$

The experiment showed that the cubic polynomial fit was optimal, with an R^2 of 0.0795, which was significant for all models (F test $p<0.05$), but the fitting degree was low, suggesting that more variables needed to be introduced.

Experiment 2: Extend the polynomial model. On the basis of experiment 1, the Z value (X_Z)

and GC content of X chromosome (GC) were added to construct an extended cubic polynomial:

$$Y = \beta_0 + \beta_1 GA + \beta_2 BMI + \beta_3 GA^2 + \beta_4 BMI^2 + \beta_5 GA \cdot BMI + \beta_6 X_Z + \beta_7 GC + \beta_8 GA \cdot X_Z + \beta_9 BMI \cdot GC + \beta_{10} X_Z \cdot GC + \beta_{11} GA^2 \cdot BMI + \text{Higher Order} \quad (4)$$

R^2 increased to 0.1386, indicating that the Z value and GC content of X chromosome were important influencing factors, and the overall model was significant ($p<0.01$).

Experiment 3: Machine learning model comparison. Ridge regression, Lasso, ElasticNet, random forest, gradient boost tree, support vector machine, neural network are introduced. The training/test set is divided into 8:2 and the parameters are cross-validated by 50%. The gradient lifting tree has an R^2 of 0.2521 and the smallest MSE (0.000670) on the test set, which is the optimal model, which verifies that the ensemble learning method can effectively capture the multivariate nonlinear interaction.

Significance test: The permutation test was used to verify the significance of the gradient lifting tree model, and the actual R^2 was much higher than the null hypothesis distribution, and the model was statistically significant.

2.2. Clustering and Point-In-Time Optimization Models

This step aims to rationally group male pregnant women according to their BMI and determine the optimal time point for NIPT detection in each group (90th percentile of the first gestational age with Y chromosome concentration $\geqslant$ 4%). Core criteria: The first gestational age was defined as the gestational age of the first detection of the Y chromosome concentration of >0.04, and the best time point was the 90% quantile of the gestational age of the first gestational age reached by the group. Monte Carlo simulation was used to evaluate the effect of 0.5-week random error on the optimal time point (1000 simulations).

Experiment 1: Preset BMI group optimization. According to the preset groups [20,28], [28,32], [32,36], [36,40], and 40+, the 90% quantile and 95% confidence interval of the first gestational age were calculated for each group, respectively.

Experiment 2: K-means clustering grouping model. BMI was used as a single clustering variable, and the elbow method and contour coefficient were used to determine the optimal clustering number (divided into 4 categories based on clinical considerations). The clustering objective function is to minimize the sum of squares within the group:

$$\min \sum_{i=1}^{K} \sum_{x \in C_i} \| x - \mu_i \|^2 \quad (5)$$

After clustering, the 90% quantile of the gestational age for the first time in each group was calculated to obtain the best time point.

Experiment 3: Regression tree group verification. With the goal of reaching the gestational age for the first time, a regression tree was constructed based on BMI, and the tree depth and number of groups were determined by 50% cross-verification. The regression tree divides the feature space by recursive dichotomy:

$$\hat{y} = \frac{1}{n_m} \sum_{i: x_i \in R_m} y_i \quad (6)$$

Where R_m is the leaf node area. Finally, the optimal number of groups and the corresponding time point are determined.

2.3. Multi-Factor Ensemble Learning Prediction Model

On the basis of the second step, multiple factors such as height, weight, and age were introduced, but the direct clustering of K-means led to serious interval overlap. To this end, the improved K-level clustering and feature weighting function are added to give the normalized variables BMI, height, weight, and age a weight of 2.0, and the rest of the weights are 1.0, and the weighted Euclidean distance is used for hierarchical clustering:

$$d_{weighted}(x_i, x_j) = \sqrt{w_{BMI}(BMI_i - BMI_j)^2 + w_{height}(H_i - H_j)^2 + w_{weight}(W_i - W_j)^2 + w_{age}(A_i - A_j)^2} \tag{7}$$

Then, the prediction model of gestational age for the first time was constructed, and linear regression, ridge regression, Lasso regression and random forest were compared. The random forest is integrated by multiple decision trees, and the final prediction is the average of all tree predictions:

$$\hat{T} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{8}$$

The performance was evaluated by 50% cross-validation, random forest was selected as the prediction model, and the 90% quantile of each group was calculated as the optimal time point based on its prediction distribution, and the Monte Carlo simulation (error standard deviation of 0.5 weeks) was performed again to evaluate the robustness.

3. Results and Discussion

3.1. Nonlinear Regression Results

Table 1: Comparison of fitting effects of each model

Model	Training R^2	Test R^2	MSE
Ridge Regression	0.0487	0.0998	0.000807
Lasso Regression	0.0000	-0.0047	0.000901
ElasticNet	0.0000	-0.0047	0.000901
Random Forest	0.8312	0.2292	0.000691
Gradient Boosting Trees	0.5997	0.2521	0.000670
SVM	-0.1194	-0.1511	0.001032
Neural Networks	-0.1245	-0.0958	0.000982

It can be seen from Table 1 that the gradient lifting tree has the highest $R^2$and the smallest MSE on the test set, indicating that it can generalize well. However, Lasso and ElasticNet compress the coefficients to zero due to strong regularization, resulting in underfitting. Although the training R^2 of the random forest is high, the test R^2decreases significantly, and there is overfitting. Gradient enhancement is achieved by gradually correcting the residual mechanism, and the degree of overfitting is low. At the same time, the predicted residuals of the gradient lifting tree are approximately normally distributed and have no systematic bias, which further confirms the reliability of the model. In summary, the first step verifies the nonlinear relationship between Y chromosome concentration and gestational age and BMI, and ensemble learning is better than traditional polynomial regression.

3.2. Clustering and Point-In-Time Optimization Results

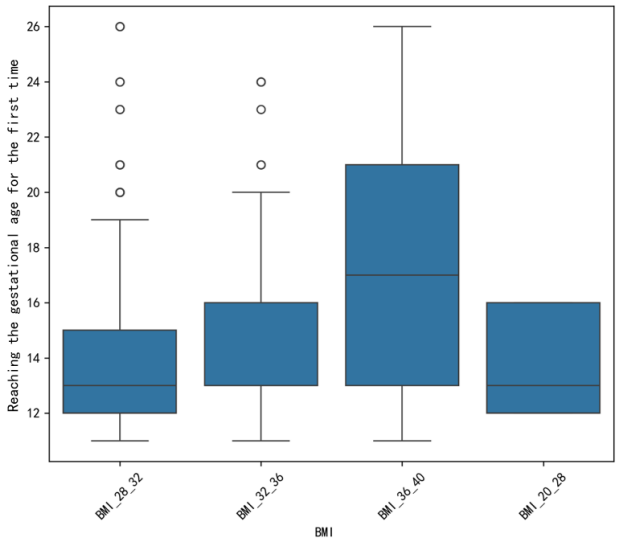

Figure 1: Statistical results of each BMI group

As can be seen from Figure 1, with the increase of BMI grouping range, the mean of the gestational age for the first time showed a gradual upward trend, and the mean of the BMI_36_40 group (17.82 weeks) was 4.07 weeks higher than that of the BMI_20_28 (13.75 weeks), and the standard deviation also increased significantly, indicating that the time of reaching the standard was not only later, but also more significant individual differences.

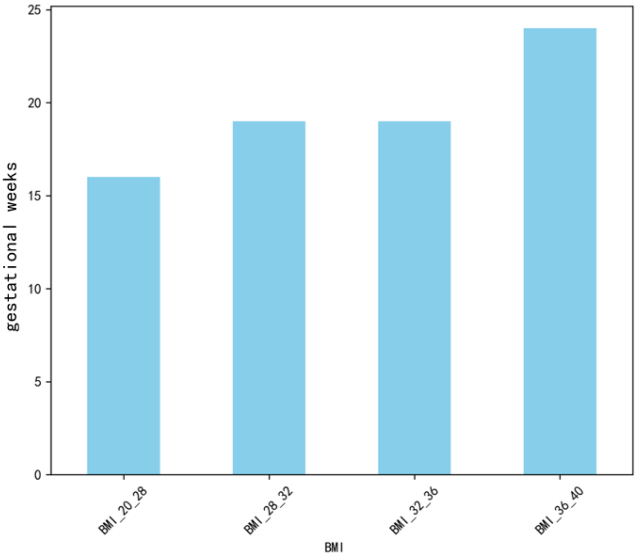

Figure 2: Best time point for NIPT

The results of Figure 2 show that the optimal NIPT detection for different BMI groups is 16, 19, 19, 24. In order to further verify the experimental results, due to the individual differences of each pregnant woman, such as giving grouping and unified NIPT detection time points for all pregnant women, it will have a great impact on its accuracy. Therefore, the experimental solution was continued, and the K-means clustering model was used to determine the optimal clustering, and the BMI grouping was carried out to determine the optimal NIPT time point.

Experiment 1 (preset grouping) obtained the optimal time points of 16, 19, 19, and 24 weeks.

Experiment 2 K-means clustering divided BMI into four groups, and the 90% quantile of gestational age for the first time and the 95% confidence interval of Monte Carlo simulation in each group are shown in Table 2. Risk stratification analysis showed that the proportion of low risk in the high BMI group decreased significantly, but the optimal time could still cover 90% of the population.

Table 2: K-means clustering optimal time point and Monte Carlo simulation confidence interval (unit: week)

BMI grouping	BMI range	Best time point	Lower limit of 95% confidence interval	Upper limit of 95% confidence interval
Superrestructuring	[26.62,29.90]	16	15.8	16.5
Obesity group I	[29.96,32.19]	19	18.6	19.7
Obesity group II	[32.26,34.93]	17	16.5	17.6
Obesity group III	[35.06,39.14]	24	23.3	24.9

Table 2 shows that all confidence interval widths are less than 1.2 weeks, indicating that the detection error has little impact on the optimal time point and the model stability is strong. The optimal time points of group 1~4 were 19, 20, 26, and 17~18 weeks, respectively, and although the risk could also be distinguished, the boundaries between groups did not match the clinical BMI classification completely. Based on the rationality, clinical interpretability and error stability of the grouping, the K-means clustering scheme in experiment 2 was the best.

3.3. Multivariate Ensemble Learning Predicts Outcomes

The grouping results based on K-level clustering + feature weighting were [26.6, 30.8], [29.1, 33.7], [33.1, 36.4], and [35.6, 39.1], and the overlap rate between intervals was significantly reduced. The randomized forest regression model performed best in predicting the first gestational age (see Table 3), with R^2=0.912 and RMSE=0.693. The optimal detection time points and the 95% confidence intervals of Monte Carlo simulations for each subgroup are shown in Table 4.

Table 3: Performance comparison of each model (50% cross-validation)

Model	Cross-validate the R^2	mean and the final R^2	RMSE
Linear regression	0.0428	0.0835	2.479
Ridge regression	0.0432	0.0835	2.480
Lasso Regression	0.0409	0.0805	2.490
Random Forest	0.2807	0.9124	0.693

It can be seen from Table 3 that the R^2 of random forest is much higher than that of the linear model, indicating that there is a strong nonlinear relationship between multiple factors and the time to reach the standard, and ensemble learning can effectively model it. The confidence interval width of the Monte Carlo simulation in Table 4 is less than 1.2 weeks, indicating that the model is robust to detection errors. In summary, after the introduction of multiple factors in the third step, the prediction accuracy is greatly improved, and the grouping is more refined, and the point in time recommendation has more individualized advantages.

Table 4: Optimal time point and sensitivity analysis of each subgroup of multivariate model (unit: week)

Cluster label	Predict 90% quantile	95% lower limit of confidence interval	95% upper limit of confidence interval	Interval width
0	17.93	17.61	18.23	0.62
1	23.01	22.45	23.63	1.18
2	15.83	15.60	16.38	0.78
3	17.75	17.20	18.09	0.89

4. Conclusions

In order to solve the problem of lack of quantitative model at the time of NIPT detection, a progressive algorithm framework from single-factor nonlinear regression to multi-factor ensemble learning was constructed, and the three-step core modeling task was completed, and the main conclusions are as follows:

(1) In the first step nonlinear regression model, it was confirmed that there was a significant nonlinear relationship between Y chromosome concentration and gestational age and BMI by comparing the seven models (gradient lifting tree test $R^2 \approx 0.2521$), BMI had a negative effect on concentration, gestational age had a positive effect, and X chromosome Z value and GC content were important moderating variables. This conclusion reveals the limitations of traditional linear models and provides a basis for feature selection for subsequent time-point optimization. Although the goodness of fit of the model is not high, the perplacement test shows that the model is significant as a whole, and the residuals are not systematic, which is in line with the complexity of multi-factor interference in actual clinical testing.

(2) The second step of clustering and time point optimization model used K-means clustering to adaptively divide the BMI of male pregnant women into four groups: [26.62, 29.90], [29.96, 32.19], [32.26, 34.93], [35.06, 39.14], and the corresponding optimal NIPT time points were 16 weeks, 19 weeks, 17 weeks, and 24 weeks, respectively. Monte Carlo simulations showed that the width of the confidence interval under the detection error (0.5 weeks) was less than 1.2 weeks, and the recommended stability at the time point was high. The regression tree cross-validation further confirms the rationality of the division of the four groups, and the grouping strategy can effectively distinguish the risk level and provide refined detection window suggestions for clinical practice. Compared with the preset grouping, the clustering grouping better captures the nonlinear relationship between BMI and the time to reach the standard, making the time point recommendation of the high BMI group more accurate.

(3) In the third step, the multi-factor ensemble learning prediction model introduced multiple factors such as height, weight, and age, and obtained four groups of non-overlapping groups through K-level clustering and feature weighting (BMI weight 2.0), and constructed a high-precision gestational age prediction model for the first time based on random forest ($R^2 \approx 0.9124$, RMSE=0.693), and determined that the optimal time points for each subgroup were 16.8, 18.8, 18.9, and 24.0 weeks. The model is robust to detection errors and can capture multi-factor nonlinear interactions, realizing the leap from population statistics to individualized prediction. Compared with the second step, the prediction accuracy of the third step is significantly improved, and the time point recommendation is more individualized, especially for obese I and II groups, the time point difference is more subtle, reflecting the clinical reality of multi-factor combined effect.

Overall, the algorithm chain of "nonlinear regression + K-means clustering + hierarchical feature

weighting + random forest prediction" proposed in this paper provides a systematic solution for NIPT time-point optimization, which significantly improves the scientificity and reliability of detection timing decision-making. This method is not only suitable for time-based optimization of male fetal Y chromosome concentration, but also its algorithm framework can be generalized to the dynamic monitoring scenario of other biomarkers. In the future, the generalization ability of the model can be further verified by combining deep learning timing models and multi-center large-sample data, and a visual decision-making system can be developed to assist clinicians in recommending detection time points in real time, and explainable artificial intelligence technology can be explored to enhance the clinical trust of the model and promote the intelligent development of prenatal screening.

References

[1] Shi Y, He J, Liu H, Wang Y, Chen Z. Machine learning for predicting fetal DNA fraction in noninvasive prenatal screening[J]. BMC Bioinformatics, 2021, 22(1): 512.

[2] Wang L, Zhang X, Chen Y, Liu M. Gradient boosting for optimizing screening timing in noninvasive prenatal testing[J]. Frontiers in Genetics, 2022, 13: 894567.

[3] Chen X, Li J, Wang Y. Cluster analysis of maternal BMI and its association with fetal aneuploidy risk[J]. Journal of Perinatal Medicine, 2023, 51(4): 512-520.

[4] Zhang H, Liu Q, Zhou J. Hierarchical clustering for personalized NIPT timing recommendation[J]. IEEE Journal of Biomedical and Health Informatics, 2024, 28(2): 1123-1132.

[5] Liu J, Wang Y, Zhao S. Monte Carlo simulation-based robustness evaluation for prenatal screening models[J]. Computer Methods and Programs in Biomedicine, 2023, 231: 107401.

[6] Kim S Y, Park J H, Lee K. Deep learning-based noninvasive prenatal testing for fetal chromosome abnormalities[J]. Nature Communications, 2022, 13(1): 4562.

[7] Zhou Q, Yang L, Zhang R. Interpretable machine learning for gestational age prediction using multi-omics data[J]. Briefings in Bioinformatics, 2023, 24(3): bbad102.

[8] Huang T, Wang H, Chen S. Ensemble learning for detecting fetal subchromosomal abnormalities[J]. Human Genetics, 2022, 141(5): 1013-1024.

[9] Patel A, Gupta R, Sharma N. The role of maternal BMI in cell-free fetal DNA concentration: a systematic review[J]. Prenatal Diagnosis, 2021, 41(8): 952-961.

[10] Yang Z, Liu Y, Chen W. Feature selection and classification for NIPT data using SVM and random forest[J]. BMC Medical Genomics, 2024, 17(1): 34.

[11] Li M, Wang J, Zhang Y. Explainable artificial intelligence for prenatal diagnosis: a survey[J]. Artificial Intelligence in Medicine, 2024, 148: 102777.

[12] Gao Y, Xu L, Chen H. Uncertainty-aware deep learning for NIPT timing optimization[J]. IEEE Transactions on Biomedical Engineering, 2024, 71(5): 1602-1612.