Optimization of Non-invasive Prenatal Test Based on Ridge Regression and Turnbull NPMLE

Wenbin Liu[1,4], Xinggong Yan[2], Dongbing Liu[3]
[1]Wuhan Technology and Business University, Wuhan, China
[2]Wuhan University of Technology, Wuhan, China
[3]College of Mathematics and Computer, Panzhihua University, Panzhihua, China
[4]Wuchang Institute of Technology, Wuhan, China

Corresponding author: Dongbing Liu (e-mail: cq2206@126.com).

ABSTRACT Non-invasive prenatal test (NIPT) can be used for early screening of fetal chromosomal aneuploidy. However, individual variations among pregnant women make it difficult to align with the traditional standardized testing schedule. Testing too early may fail due to insufficient fetal DNA concentration, while testing too late could result in missing the optimal intervention window. In this paper, ridge regression is used as the main model, supplemented by LASSO sparse screening and random forest nonlinear control, and the performance is evaluated by cross-validation. This potential risk is quantified and optimized to develop NIPT testing timing protocols tailored for different pregnant women.

INDEX TERMS ridge regression, Turnbull NPMLE, Non-invasive prenatal test, optimization.

I. INTRODUCTION

NON-INVASIVE prenatal test (NIPT), as a mature prenatal screening technology, can effectively detect chromosomal aneuploidy abnormalities in fetuses by analyzing fetal free DNA in maternal plasma, providing valuable early warning for subsequent clinical interventions. The core principle is based on the relative concentration of fetal DNA in maternal blood (i.e., "chromosome concentration"), whether the degree is within the normal range to make a judgment [1]–[5]. The accuracy of this key indicator is affected by a dynamic mix of factors, particularly the gestational age and body mass index (BMI) of the pregnant woman [6]–[10]. In clinical practice, traditional empirical grouping and standardized testing time points often fail to accommodate the unique physiological conditions of individual pregnant women, leading to two core clinical challenges. On one hand, premature testing may lead to sequencing failure or unreliable results due to insufficient fetal DNA concentration, thereby wasting valuable testing resources and time. On the other hand, if the testing is delayed, even if abnormalities are ultimately detected, the potential risk to the pregnant woman and family may be significantly increased due to missing the optimal therapeutic window [11]–[15]. Therefore, how to quantify this qualitative potential risk and translate it into actionable optimization targets, thereby providing a scientific and precise NIPT timing selection protocol for pregnant women with diverse physiological characteristics, remains an urgent practical challenge to address [16]–[20].

According to clinical experience, the main congenital anomalies in fetuses are Down syndrome, Edwards syndrome, and Patau syndrome. These conditions are determined by abnormalities in the "proportion of free DNA fragments in chromosomes 21, 18, and 13" (referred to as "chromosome concentration"). The accuracy of NIPT is primarily determined by the concentration of fetal sex chromosomes (XY for male fetuses, XX for female fetuses). Fetal sex chromosome concentration can typically be detected between 10 to 25 weeks of gestation in pregnant women. If the Y chromosome concentration in male fetuses reaches or exceeds 4%, and the X chromosome concentration in female fetuses shows no abnormalities, the NIPT results can be considered essentially accurate. Otherwise, it is difficult to ensure the accuracy of the results. Meanwhile, it is crucial to detect unhealthy fetuses at the earliest stage in practice, as failure to do so may result in a shortened therapeutic window. Early detection (within 12 weeks) carries a lower risk; mid-term detection (13–27 weeks) carries a higher risk; late detection (after 28 weeks) carries an extremely high risk.

Clinical data show that the Y chromosome concentration of male fetus is closely related to gestational age and body mass index (BMI) of pregnant women. The timing of NIPT (relative to gestational age) is typically determined by grouping pregnant women according to their BMI values (e.g., [20,28], [28,32], [32,36], [36,40], $\geq$40).

Given the individual variations in maternal age, BMI, and gestational status among pregnant women, applying

a standardized empirical grouping and a uniform testing time point for NIPT across all cases will significantly compromise its accuracy. Therefore, reasonable grouping of pregnant women based on BMI and determining the optimal NIPT timing for each group can reduce the potential risk of shortened treatment windows due to fetal health issues in certain cases.

To investigate the optimal NIPT timing for various maternal populations and analyze the accuracy of the test, the appendix presents NIPT data from pregnant women in a specific region (predominantly with high BMI). In practical testing, sequencing failures frequently occur. For example, the test time point is too early and the influence of uncertain factors. Meanwhile, to enhance the reliability of test results, multiple blood draws and tests are performed for some pregnant women, or a single blood draw is used for multiple tests.

II. DATA PREPROCESSING

In order to ensure the accuracy and reliability of modeling, the original data are cleaned and processed systematically.

Firstly, the columns in the original data table are preliminarily checked. If key characteristics (such as gestational age, Y chromosome concentration, GC content, X chromosome Z value) are missing, the record is directly excluded. For non-critical variables, mean or median imputation is used if the missing proportion is small, while those with excessively high missing proportions are excluded from modeling features. Meanwhile, outliers that significantly exceed reasonable ranges (e.g., GC content below 20% or above 80%) are excluded.

Secondly, to facilitate unified processing, the original format of gestational age is "week+day" (e.g., 12+3). It is converted into a decimal format measured in weeks through regular expression parsing. The expression is as follows.

$$GA = week + \frac{day}{7} \tag{1}$$

For example, "12+3" is converted to 12.43 weeks.

Thirdly, for proportional variables such as Y chromosome concentration and GC content, considering that the original file may present data in either 0–1 or 0–100% ranges, they are uniformly converted to percentage scales with one decimal place precision.

Fourthly, considering that the distribution of chromosome concentration is usually skewed, the natural logarithmic transformation is further applied to the Y chromosome concentration and X chromosome concentration to weaken the long-tail effect, resulting in two features of $\ln(Y_{\text{conc}})$ and $\ln(X_{\text{conc}})$.

Fifthly, for the label variable "Fetal Health Status", the original data is of text type, containing various expressions such as "Healthy/Normal/Negative" and "Abnormal/Positive". These are binarized. Health/normal/negative is recorded as 1, abnormal/positive as 0, and the final target variable is obtained. Based on this, according to the question and clinical judgment principles, if the sample simultaneously meets the following four conditions, it is considered to be highly credible.

$$\begin{cases} 1, & 10 \le GA \le 25,\ Y_{\text{conc}} \ge 4\%, \\ & |Z_X| \le 3,\ 40\% \le GC \le 60\%, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

GA denotes gestational age, Y_{conc} represents Y chromosome concentration, X_{conc} indicates X chromosome concentration, Z_X refers to the Z value of the X chromosome, and GC stands for GC content. If the sample meets all the above conditions, it is marked as 1; otherwise, it is marked as 0.

The remaining valid samples are then cleaned and screened, with records containing missing values removed to obtain the final dataset. This dataset is subsequently divided into a training set and a test set in a 7:3 ratio. Meanwhile, stratified sampling is adopted to maintain the consistency of category distribution, thus avoiding the bias caused by category imbalance in model training.

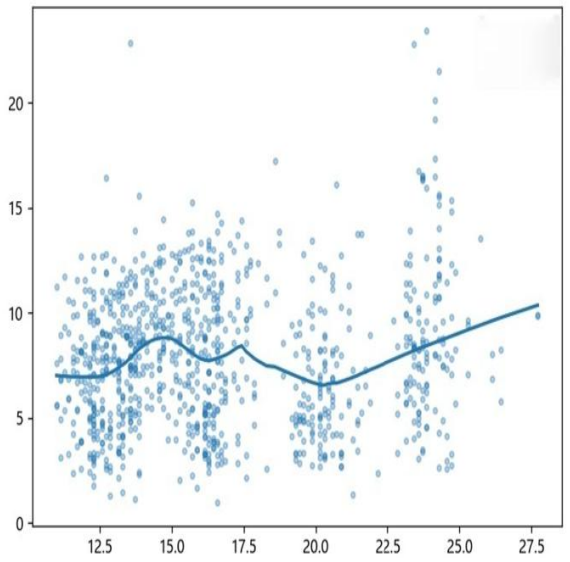

FIGURE 1. the scatter plot of Y chromosome concentration versus gestational age

III. DATA VISUALIZATION

The scatter plot of Y chromosome concentration versus gestational age is shown in Figure 1.

As shown in Figure 1, the Y chromosome concentration exhibits a gradual upward trend throughout the gestational period. There was a slight decline around weeks 19 to 21, and a more pronounced increase after week 22. Local fluctuation may be related to the variation of sample size in individual gestational weeks, individual differences or batch effect.

Box plots of Y chromosome concentration in male fetuses and maternal body mass index (BMI) are shown in Figure 2.

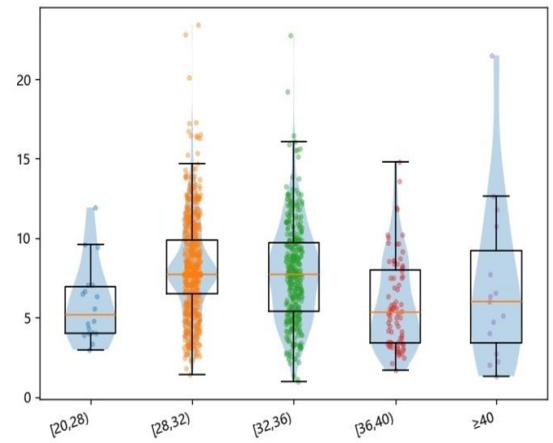

FIGURE 2. Box plots of Y chromosome concentration in male fetuses and maternal body mass index

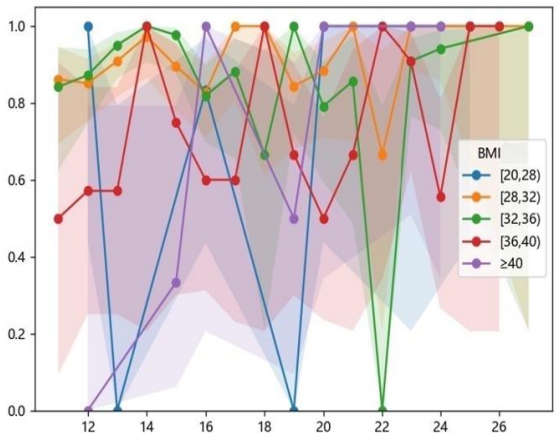

FIGURE 3. The line chart of the compliance rate according to the body mass index (BMI) of male fetuses

The predefined fixed groups were [20,28), [28,32), [32,36), [36,40), and ≥40. From Figure 2, the comparison of different BMI groups can be summarized as follows. Overall, the median values are slightly higher in the middle two groups [28,32) and [32,36); lower in the low BMI group [20,28) and higher in the high BMI group [36,40]; and significantly greater in the ≥40 group. This suggests that BMI may exert a non-monotonic effect on Y concentration: the typical levels of extremely low or high BMI are relatively low, while the intermediate range is relatively higher, and the high BMI group exhibits greater variability.

The line chart of the compliance rate according to the body mass index (BMI) of male fetuses is shown in Figure 3.

Under the achievement criteria established by the research team, the compliance rates across different BMI groups remained consistently high throughout the gestational period, with a gradual increase as gestational age advanced. The curves between BMI groups showed little difference across

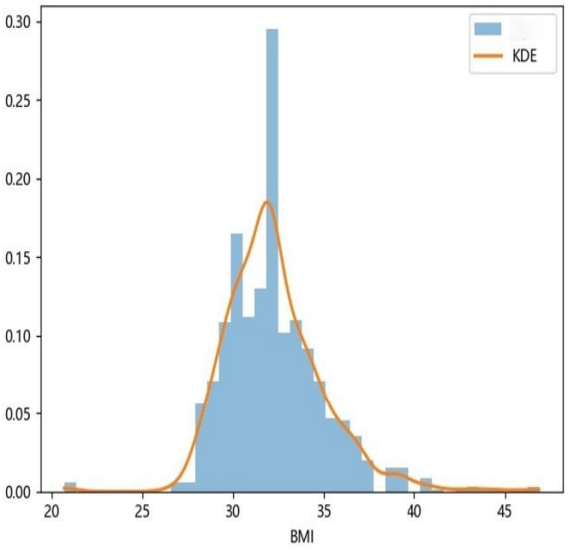

FIGURE 4. the histogram of Y concentration distribution

most gestational weeks, but a brief dip occurred at certain weeks, likely due to small sample sizes or batch effects. Overall, the positive effect of gestational age advancement on the achievement rate is more stable, while the intergroup differences in BMI are relatively minor.

Based on the distribution of maternal body mass index (BMI), the histogram of Y concentration distribution is shown in Figure 4.

As shown in Figure 4, The Y concentration exhibits a right-skewed distribution, with the main peak located in the low-to-moderate range and a long tail on the right side, while a very small number of samples reach higher values. This suggests that although the majority of sample concentrations are concentrated within a relatively stable range, a small number of High-value individuals exert a pull-up effect on the mean, thus subsequent statistical and modeling analyses should prioritize robust measures (e.g., median) while paying attention to outlier handling.

IV. ESTABLISHMENT OF RIDGE REGRESSION MODEL

The research objective is to interpret and predict Y chromosome concentration using observable features, such as one-hot encoded categorical variables including gestational age, BMI, height and weight, sequencing reads/alignment ratio/GC content, chromosome Z value, and IVF [2]. Given the response vector $y \in R^n$ and the design matrix $X \in R^{n\times p}$, the linear model with intercept term is as follows.

$$y = \beta_0 I_n + X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n). \tag{3}$$

Because of the significant multicollinearity in this problem, the ordinary least squares method is unstable in this case. For example, height and weight are strongly correlated, and various sequencing technical indicators show high correlation with chromosomal indicators. Therefore, ridge regression is used for regularization.

A. OBJECTIVE FUNCTION OF RIDGE REGRESSION AND CLOSED FORM SOLUTION

The ridge regression adds an L_2 regularization term to the sum of squared residuals (no penalty for the intercept).

$$\hat{\beta}_\lambda = \arg\min_{\beta} \frac{1}{n} \|y - \beta_0 I_n - X\beta\|_2^2, \quad \lambda \geq 0 \tag{4}$$

After columnar normalization and centralization of the numerical features, β_0 can be recovered from the sample mean. The closed form solution of ridge regression is as follows.

$$\hat{\beta}_\lambda = \left(X^T X + \lambda I_p\right)^{-1} X^T \left(y - \bar{y} I_n\right) \tag{5}$$

The equivalent constraint form is to minimize the sum of the square of the residual and the norm of the constraint coefficient as follows.

$$\min_{\beta} \|y - \beta_0 I_n - X\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_2^2 \leq t \tag{6}$$

B. SVD PERSPECTIVE AND CONTRACTION FACTOR

Let $X = UDV^T$ be the singular value decomposition of the design matrix. $(D = \mathrm{diag}(d_1, \ldots, d_r))$. The ridge regression implements different degrees of shrinkage on the projection of different principal component directions [3]. The fitted values are as follows.

$$\begin{aligned} \hat{y} &= S_\lambda, \\ S_\lambda &= X\left(X^T X + \lambda I_p\right)^{-1} X^T \\ &= U\,\mathrm{diag}\left(\frac{d_1^2}{d_1^2 + n\lambda}, \ldots, \frac{d_k^2}{d_k^2 + n\lambda}, \ldots\right) U^T \end{aligned} \tag{7}$$

It is obvious that the shrinkage factor $\left(d_k^2/(d_k^2 + n\lambda)\right)$ is less than 1 in the direction of small singular value, which can restrain the variance and stabilize the estimation.

C. HAT MATRIX AND EFFECTIVE DEGREES OF FREEDOM

The hat matrix of ridge regression is as follows.

$$S_\lambda = X\left(X^T X + \lambda I_p\right)^{-1} X^T \tag{8}$$

The trace of $df(\lambda) = tr(S_\lambda)$ is often interpreted as the effective degrees of freedom.

The larger the λ, the smaller the $df(\lambda)$, and the simpler the model. When $\lambda \to 0$, it returns to OLS.

D. EVALUATION METRICS AND HYPERPARAMETER SELECTION

The forecast and residual are as follows.

$$\hat{y} = X\hat{\beta}_\lambda + \hat{\beta}_0 I_n, \qquad e = y - \hat{y}. \tag{9}$$

In the paper, the coefficient of determination and RMSE are adopted as follows.

$$\begin{cases} R^2 = 1 - \dfrac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \\ RMSE = \sqrt{\dfrac{1}{n}\sum_{i=1}^n e_i^2}. \end{cases} \tag{10}$$

The penalty coefficient is optimized by K-fold cross-validation. In the paper, K is set to 5. Select the optimal $\hat{\lambda}$ from the candidate set Λ to maximize validation set performance.

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} \frac{1}{K} \sum_{k=1}^{K} R_k^2(\lambda) \tag{11}$$

V. ESTABLISHMENT OF RANDOM FOREST MODEL

In the paper, although ridge regression is used as the main model, random forest is still introduced as a nonlinear control. The objectives are as follows. It can capture the potential nonlinear relationship and high-order interaction without explicitly constructing the polynomial or interaction term. For example, "gestational age × BMI" and "technical index × gestational age". The importance of variables was reviewed from a non-parametric perspective to test the robustness of the conclusion that gestational age is the primary variable, followed by BMI, while technical variables more reflect detection noise. As a sensitivity analysis, it assesses the dependence of the study conclusions on assumptions such as linearity, additivity, and error distribution. It also provides a performance reference upper bound to evaluate the potential gains of more complex models (e.g., tree model integration) over linear baselines. Based on these considerations, the random forest process was constructed prior to (or concurrently with) ridge regression, with specific steps detailed in Table I.

TABLE I. RANDOM FOREST MODELING PROCESS

Step	Task Overview
Step 1	Data preparation: Read the original Excel file, standardize column names and data types, and identify the target variable "Y chromosome concentration".
Step 2	Feature engineering: Analyze 'Pregnancy Week Detection', calculate 'Gestational Days (if available)', convert percentages to numerical values, and delete the ID column.
Step 3	Preconditioning pipeline: Median fill for numeric columns and mode fill with One-Hot for categorical columns in the same pipeline.
Step 4	Model building: The Random Forest Regressor is used as the baseline model, with reasonable starting hyperparameters set.
Step 5	Cross-validation and parameter tuning: Half of the CV, Randomized Search CV optimized by R2.
Step 6	Performance evaluation: The outer half of the CV report R2 mean ± standard deviation, full-sample fitting and calculation of overall R2/RMSE.
Step 7	Overfitting Diagnosis and Constraints: For contrast training or overall and cross-validation (CV) performance, adjust max_depth when necessary, increase min_samples_leaf, and modify max_features.
Step 8	Result output: Export charts such as 'measured vs predicted', 'CV distribution', and 'feature importance', along with the optimal parameter records.

To align with the preprocessing standards of ridge regression, random forest regression applies a standardized pipeline transformation to the data prior to model integration. Let the original explanatory variable be a matrix $X \in R^{n \times p}$, and the response be $y \in R^n$. Firstly, the text representation of 'gestational age' is parsed as a continuous quantity measured in weeks, and the number of gestational days is calculated from the 'last menstrual period' and 'test

date' when available. The percentage feature is converted to a number after removing the "%". The ID column with no information is deleted. Then the preprocessed mapping $\Phi(\cdot)$ is constructed. The value column fills in missing values with the median. The categorical columns are filled with the mode and encoded using One-Hot. Thus, the transformed design matrix $\widetilde{X} = \Phi(x)$ is obtained. Finally, based on the foundation, the random forest learns a non-parametric regression function $\hat{f} : R^{\tilde{p}} \to R$, aiming to minimize the squared error, and enhances robustness and generalization by integrating multiple randomized regression trees.

The core concept of random forest is 'bagging' and 'random subspace'. Given the number of trees T, for each tree $t = 1, \ldots, T$, the bootstrap sample index set I_t is obtained by sampling from $\{1, \ldots, n\}$ with replacement. During the tree training process, the optimal partitioning is identified for each internal node within the randomly selected m_{try} feature subsets [7], thereby introducing inter-model variability to reduce overall variance. The CART regression criterion is used to train the single tree. For the current node sample set $R \subseteq \{1, \ldots, n\}$, the residual sum of squares of the sub-node is minimized on the partition s formed by the candidate features and the threshold.

$$L(s;R) = \sum_{i \in R_{L(s)}} \left(y_i - \bar{y}_{R_{L(s)}}\right)^2 + \sum_{i \in R_{R(s)}} \left(y_i - \bar{y}_{R_{R(s)}}\right)^2 \tag{12}$$

where $R_{L(s)}$ denotes the sample set of left sub-nodes under partition s, $R_{R(s)}$ denotes the sample set of right sub-nodes under partition s. $\bar{y}_A = \frac{1}{|A|}\sum_{i \in A} y_i$. Optimal division is equivalent to minimizing variance (reducing impurity).

$$\Delta i(s;R) = Var(y;R) - \frac{|R_{L(s)}|}{|R|} Var\left(y;R_{L(s)}\right) - \frac{|R_{R(s)}|}{|R|} Var\left(y;R_{R(s)}\right) \tag{13}$$

The tree growth is controlled by several stopping criteria to balance the fit and complexity of the model, including maximum depth(max_depth), minimum partition sample number(min_samples_split), minimum leaf node sample number(min_samples_leaf), etc. After training, the forest prediction is the simple average of the outputs of each sub-model.

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} \hat{f}_t(x), \qquad \hat{y}_i = \hat{f}(\widetilde{x}_i), \quad i = 1, \ldots, n. \tag{14}$$

In the paper, the global importance measure based on impurity reduction is used to quantify the contribution of each feature to the prediction performance. If the j-th feature is involved in the partition of the whole forest, the importance of the j-th feature can be given by the sum of the impurity reduction brought by all the relevant partitions.

$$Imp(j) = \sum_{t=1}^{T} \sum_{s \in S_j^{(t)}} \Delta i\left(s; R^{(t)}(s)\right) \tag{15}$$

Here, $R^{(t)}(s)$ denotes the sample set of node s in the t-th tree, and $\Delta i(\cdot)$ is consistent with Equation (13). In the paper, the hyperparameters are selected by half cross-validation in the model selection and evaluation.

$$\theta = \begin{pmatrix} T, m_{\text{try}}, \text{max\_depth}, \\ \text{min\_samples\_split}, \\ \text{min\_samples\_leaf} \end{pmatrix} \tag{16}$$

When the validation coefficient about Ten percent of k cross-validation is $R^2_{(k)}(\theta)$, the cross-validation objective is to maximize the mean performance.

$$\hat{\theta} = \arg\max_{\theta \in \ddot{\Theta}} \frac{1}{5}\sum_{k=1}^{5} R^2_{(k)}(\theta),$$
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

To prevent data leakage, the $\Phi(\cdot)$ missing value imputation and encoding are only performed during training and applied to the corresponding validation. The optimization adopts random search to cover a wider hyperparameter combination space within the given computational budget. Finally, the model is retrained on the full dataset using the optimal $\hat{\theta}$, with the overall R^2 and root mean square error (RMSE) reported.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{18}$$

The results of the model are presented in the form of graphs and tables, such as the "actual and predicted values", "cross-validation distribution", and "feature importance", which are used to visually assess the fit, stability, and interpretability of the model.

VI. INTERPRETATION OF RESULT

A. INTERPRETATION OF RIDGE REGRESSION RESULT

Figure 5 shows that the predicted values are generally distributed along the 45° line of $y = x$, with the highest density of points in the medium concentration region, and the overall R^2 value is approximately 0.549. The high-value end shows a few points slightly below the dashed line, indicating mild underestimation and weak heteroscedasticity, but this does not alter the alignment trend or the good calibration. This indicates that ridge regression has practical explanatory and predictive ability in the real world of sequencing noise and individual differences.

In Figure 6, the calibration curve superimposed on the scatter points generally follows a smooth upward trend along the $y = x$ axis, exhibiting only a mild S-shaped curve at

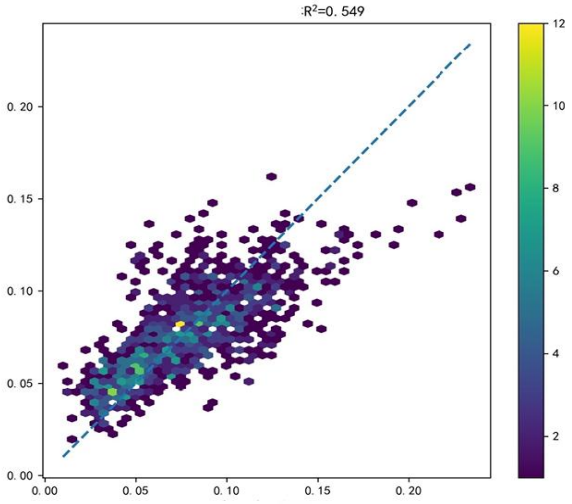

FIGURE 5. Regression of Ridge: Measurement and Prediction

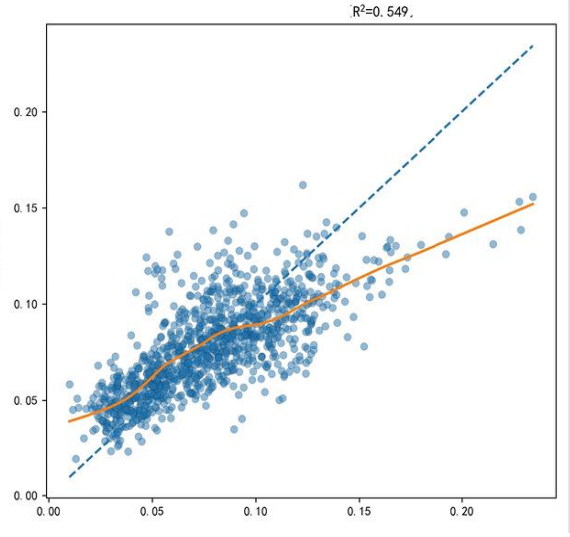

FIGURE 6. Regression of Ridge: Calibrated Curve (LOWESS)

medium-to-high concentration levels and slightly deviating from the diagonal. This indicates that the overall deviation is small and the structural relationship is effectively captured. The slight nonlinearity/heteroscedasticity is mainly concentrated in the high value end, and has limited influence on the overall prediction, which further confirms the reliability of ridge regression as a first-order baseline model.

In Figure 7, the target-related numerical feature maps show distinct clusters of physical characteristics (height-weight-BMI), technical features (read count/alignment ratio/GC/filtration ratio, etc.), and chromosomal metrics, reflecting strong multicollinearity. The ridge regression, which implements coefficient shrinkage and smoothing by adding λI to $X^T X$, is specifically designed for this structure, thus enabling stable estimation and better generalization on this data.

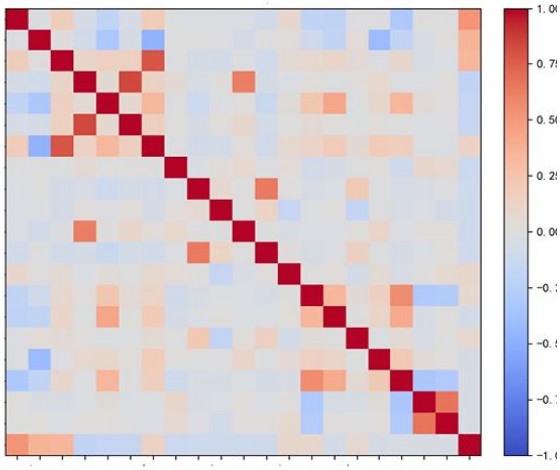

FIGURE 7. Correlation diagram

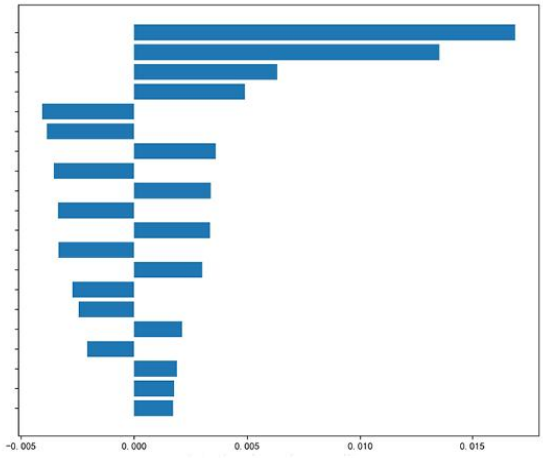

FIGURE 8. standardized coefficient

In Figure 8, Gestational weeks/day of pregnancy, Y chromosome Z value, and X chromosome concentration contributions rank highest, with directions consistent with physiological mechanisms (progression of pregnancy, increased fetal-derived cfDNA). The BMI showed a negative trend, consistent with the dilution effect of body fluid volume. Technical variables (read counts, alignment ratio, GC, etc.) contribute to some extent but are secondary, with the primary emphasis being on detection noise control. Overall, ridge regression provides a biological and interpretable conclusion while maintaining predictive power.

In Figure 9, Most points are close to the diagonal and slightly deviate from the tail, indicating that the residuals are generally approximately normal and the systematic bias is not significant. This means that the linear-Gaussian approximation is basically valid, the statistical inference and the confidence conclusion are reliable, and there is no significant structural omission in the ridge regression fitting.

In Figure 10, The kernel density curve is approximately

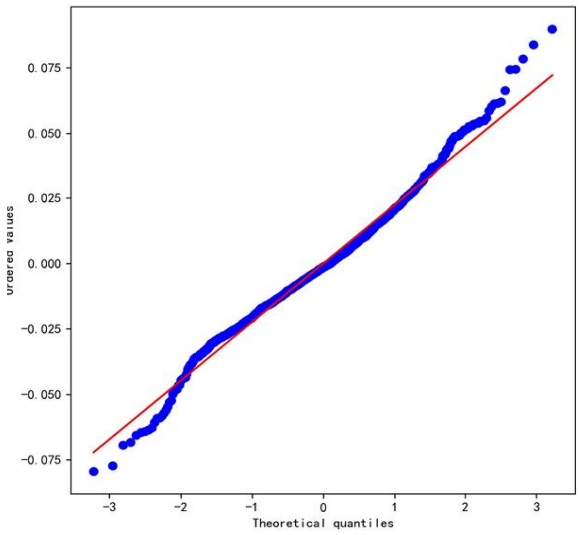

FIGURE 9. Residual Q–Q plot

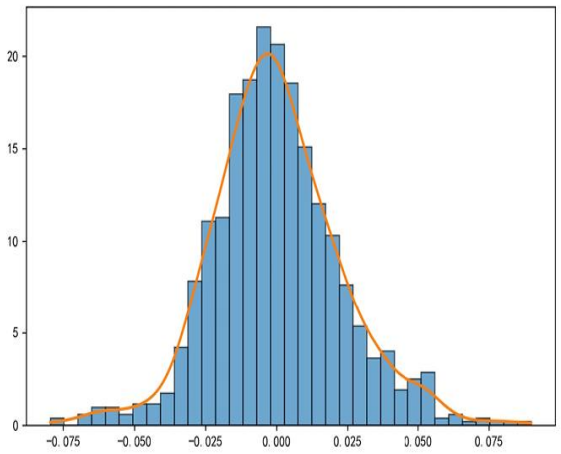

FIGURE 10. Residual distribution (histogram + KDE)

bell-shaped and symmetrical, which indicates that the model has no significant deviation and the error is mainly random fluctuation. The slightly thicker right tail is consistent with the slight underestimation of the high value end, but the scale is limited. It is proved that ridge regression achieves a good balance between accuracy, robustness and interpretability, and is suitable for the main model. The same features and preprocessing caliber as ridge regression are used to ensure that the two models can give the conclusion under the condition of strict comparability.

B. LASSO REGRESSION ANALYSIS

As a linear benchmark alternative to nonlinear controls, LASSO regression is also estimated. In Figure 11, The point cloud generated by LASSO also follows the 45-degree line, with an overall R^2 value of approximately 0.503, comparable to the ridge regression's R^2 of around 0.549.

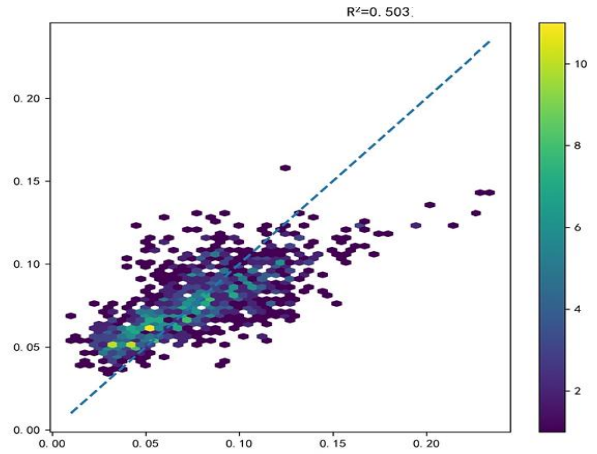

FIGURE 11. LASSO analysis plot

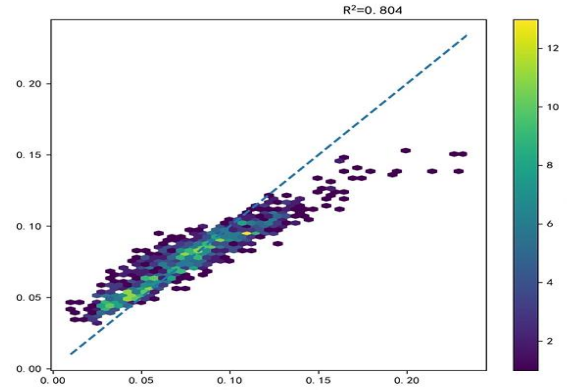

FIGURE 12. Random Forest: Measurement and Prediction

The error structure and calibration trend are also consistent. The main effect variables and coefficients of the two models are highly consistent. The gestational age is the most significant positive factor, while BMI/weight shows an overall negative correlation. Additional explanations are provided by X chromosome concentration and Y chromosome Z value, with technical quality indicators serving as secondary corrections. The key difference is that LASSO employs l_1 regularization to generate sparse solutions, automatically reducing several strongly correlated secondary features to zero, which facilitates variable screening; whereas ridge regression demonstrates slightly better predictive stability. LASSO and ridge regression reached consistent conclusions, with LASSO serving as the screening tool and ridge regression as the final reported model.

C. ANALYSIS OF RANDOM FOREST RESULTS

In Figure 12, the point cloud height in the middle and low range (approximately 0.03 to 0.12) is close to the 45° reference line, with an excellent overall fit, and the coefficient of determination $R^2 = 0.804$. However, at the high-value end (above 0.12), systematic underestimation occurs, with the point cloud as a whole lying below the

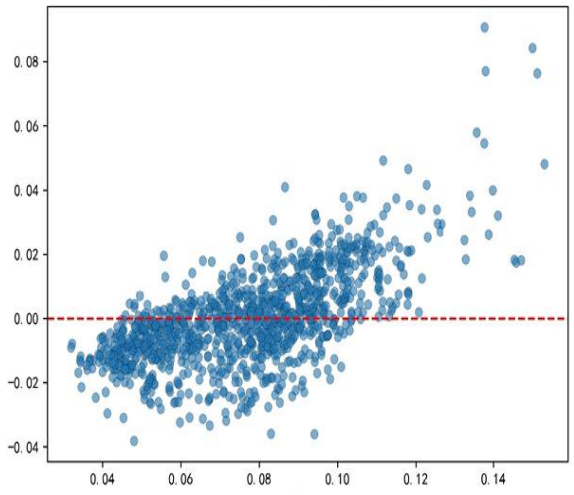

FIGURE 13. Random Forest: Residual and Fitting

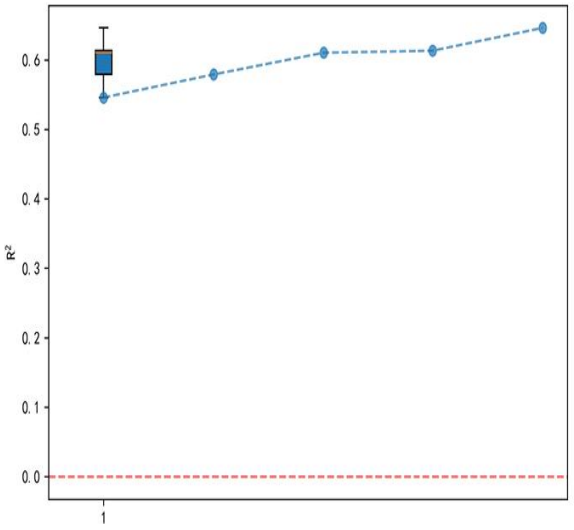

FIGURE 14. cross validation R^2 distribution

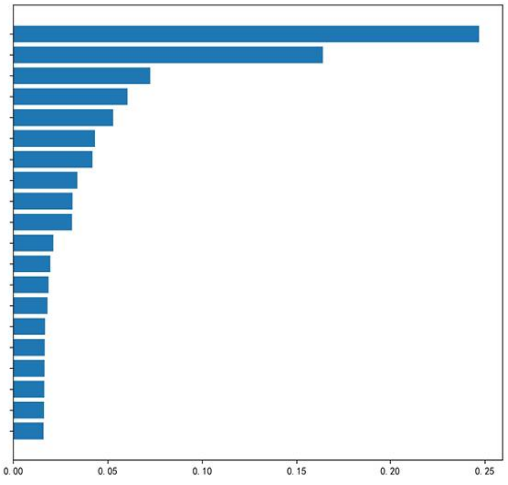

FIGURE 15. characteristic importance

diagonal, indicating insufficient calibration of the model for tail samples and a tendency of "regression to the mean" smoothing. This phenomenon usually means that the model is fit deeply on the local segmentation of the training data, but the ability of extrapolation to the structure of the extreme region is insufficient, which is one of the morphological evidence of overfitting.

In Figure 13, the residual (actual − predicted) increases with the increase of the fitted value, and the variance of the residual increases significantly with the increase of the fitted value, which is a typical "funnel-shaped" dispersion. The residual of low fitted value area is negative, while the residual of middle and high fitted value area is positive, which reflects the coexistence of systematic underestimation and heteroscedasticity of high concentration samples. The figure shows that the model learned relatively fine segmentation boundaries in the middle interval of the training samples, but these boundaries are difficult to reproduce for new samples, which further supports the overfitting judgment from the error structure.

In Figure 14, the median R^2 value of 50% cross-validation ranges from 0.58 to 0.62, significantly lower than the overall fit's 0.804, with a nearly 0.20 generalization gap between them. The small fluctuation between the percentages indicates that the performance degradation is not caused by random sampling, but by the stability problem of the model complexity relative to the sample size and noise level. The graph provides key evidence at the quantitative level. The high explanatory power of training and ensemble fitting did not achieve the same magnitude improvement on the validation fold, and the random forest exhibited significant generalization error.

In Figure 15, the reduction in impurity measurement showed that a few variables (such as X chromosome concentration and gestational days calculated by date) contributed significantly, and a group of related technical and biological characteristics formed a 'related cluster'. In this structure, the tree model tends to repeatedly split within these clusters, thereby memorizing the local boundaries of training samples and increasing variance. Meanwhile, the importance of impurity also amplifies the tendency to use these variables for fine-grained segmentation of continuous high-base variables. The result is consistent with the aforementioned findings. The phenomena of "high overall R^2–low validation R^2" and "funnel-shaped residuals" corroborate each other, revealing the mechanistic basis of overfitting.

VII. MODEL RESULT COMPARISON

In the paper, a model is established with "Y chromosome concentration" as the response variable, considering gestational age, BMI, and chromosomal indicators. Visualization and correlation analysis revealed that the Y concentration exhibited an overall right skew and increased significantly with gestational weeks, while the rise in BMI was ac-

companied by a decline in the median level. There is strong multicollinearity among height, weight, BMI and multiple technical variables. Based on this, ridge regression is adopted as the main model. With unified pipeline (missing value imputation, One-Hot encoding, and normalization) and half cross-validation, the overall R^2 value was approximately 0.55, consistent with the cross-validation performance. The residuals are approximately normal and only slightly heteroscedastic at the high end, which indicates robust generalization and good calibration. The direction of the standardized coefficient is consistent with the physiological mechanism. Pregnancy weeks are the most significant positive factor, with the Y/Z ratio and X concentration providing additional information. BMI shows an overall negative effect, while technical variables play a secondary but stable corrective role.

To test the linear assumption and structural robustness, random forest and LASSO are set as controls. The random forest has a high R^2 on the whole sample, but the cross-validation is significantly reduced, and the systematic underestimation and funnel-shaped residual appear at the high value end, indicating overfitting, so it is only used for the sensitivity verification of nonlinear morphology and importance ranking. LASSO identifies dominant variables consistent with ridge regression, though its predictive performance is slightly inferior, it effectively aids in variable screening. In conclusion, ridge regression demonstrates optimal balance between interpretability and generalization capability under this data structure, and is recommended as the final model. The results indicate a positive correlation between gestational age and Y chromosome concentration, while BMI exhibits a negative correlation with Y chromosome concentration.

ACKNOWLEDGMENT

This work is supported by the Scientific Research Team Plan of Wuhan Technology and Business University (Team number: WPT2023013).

REFERENCES

[1] Q. Jiang, J. Deng, L. Liu, *et al.*, "Genetic Diagnosis of Fetal Heart Abnormalities in Prenatal Ultrasound Examination," *Journal of Jinan University (Natural Sciences and Medical Sciences Edition)*, vol. 46, no. 5, pp. 640–650, 2025.

[2] N. Wu, L. Liao, X. Xu, *et al.*, "Application of echocardiography in the evaluation of fetal cardiac function during the second trimester of gestational diabetes mellitus," *China Maternal and Child Health Care*, vol. 40, no. 24, pp. 4649–4653, 2025.

[3] Y. Zhang, B. Wang, Q. Zhang, *et al.*, "Research on the Diagnostic Value of MRI Combined with Ultrasonography for Fetal Thoracic Abnormalities," *Research on CT Theory and Application*, vol. 34, no. S1, pp. 119–122, 2025.

[4] Y. Lv, X. Qin, and W. Zhu, "Diagnostic efficacy of prenatal ultrasound screening for fetal developmental abnormalities," *Imaging Research and Medical Applications*, vol. 9, no. 24, pp. 104–106, 2025.

[5] M. Huang, Q. Chen, and L. Zeng, "Genetic Diagnosis and Pregnancy Outcomes Analysis of 165 Cases of Fetal Cervical Translucency Thickening," *Journal of Qiqihar Medical University*, vol. 46, no. 22, pp. 2126–2130, 2025.

[6] J. Li, Y. Wang, Q. Gong, *et al.*, "Application of scalable non-invasive prenatal testing technology in prenatal screening for fetal chromosomal copy number variations," *Hainan Medical Journal*, vol. 36, no. 22, pp. 3301–3305, 2025.

[7] Z. Wei, Y. Y. Hong, and X. Huan, "Analysis of the Prenatal Ultrasonography Diagnostic Value of the Abnormality of the Inferior Vena Cava in Fetus," *Medical Information*, vol. 38, no. 22, pp. 161–164, 2025.

[8] C. Lizhen, W. Yi, and W. Jing, "MRI imaging features and classification diagnostic value of fetal corpus callosum anomalies," *China CT and MRI Journal*, vol. 23, no. 11, pp. 9–11, 2025.

[9] M. Wang, "Correlation between Nuchal Translucency Thickness in Early Pregnancy and Prenatal Diagnosis Results and Pregnancy Outcomes," *Image Research and Medical Applications*, vol. 9, no. 22, pp. 56–58+61, 2025.

[10] F. Zeng, C. Xiao, and S. Xiang, "The Application Value of Continuous Segmental Scan of Fetal Abdominal Veins in the Diagnosis of Abnormalities of Fetal Abdominal Veins," *Journal of Medical Imaging*, vol. 35, no. 10, pp. 141–144, 2025.

[11] Q. Yao, "Diagnostic efficacy of prenatal four-dimensional color Doppler ultrasound (HD-flow) combined with two-dimensional ultrasound in detecting congenital cardiac anomalies in fetuses," *Harbin Pharmaceutical*, vol. 45, no. 5, pp. 102–104, 2025.

[12] R. Guo, "Ultrasonographic diagnosis and prognosis analysis of nasal bone absence or dysplasia in early pregnancy," *China Medical Device Information*, vol. 31, no. 20, pp. 19–22, 2025.

[13] Z. Lin, "Prenatal ultrasound diagnostic imaging features and significance of fetal lateral ventricular anomaly and related diseases with abnormal neuronal migration," *Heilongjiang Medical Journal*, vol. 38, no. 5, pp. 1200–1202, 2025.

[14] C. Zhang and X. Hu, "Karyotype analysis of amniotic fluid chromosomes in 2357 pregnant women in Wuhan Maternal and Child Health Hospital," *Journal of Molecular Diagnostics and Therapeutics*, vol. 17, no. 10, pp. 2024–2026, 2025.

[15] D. Qi, D. Zhang, N. Ye, *et al.*, "Clinical Value of Prenatal Ultrasound Measurement of Fetal Lateral Ventricular Width and Posterior Cranial Fossa Pool Width in the Diagnosis of Neurological Abnormalities," *Journal of Clinical Ultrasound Medicine*, vol. 27, no. 9, pp. 741–747, 2025.

[16] J. Li, "Diagnostic value of ultrasound soft indicators combined with serological indicators for fetal chromosomal abnormalities in early and mid-pregnancy," *Modern Medical Imaging*, vol. 34, no. 9, pp. 1740–1743, 2025.

[17] X. Liu, J. Zhang, M. Cheng, *et al.*, "Genetic Analysis and Ultrasonographic Characteristics of Fetal Congenital Malformations of the Umbilical Vein-Portal Vein System," *Journal of Hebei Medical University*, vol. 46, no. 9, pp. 1064–1069, 2025.

[18] H. Dong, S. Liu, and Z. Yin, "Comparative Analysis of the Application Value of CMA and CNV-seq in Prenatal Diagnosis of Fetal Chromosomal Abnormalities," *China Journal of Eugenics and Genetics*, vol. 33, no. 9, pp. 2023–2027, 2025.

[19] Z. Zeng, W. Huang, Z. Gan, *et al.*, "The Application Value of CNV-seq Sequencing Technology in Prenatal Diagnosis Indications," *China Journal of Prenatal Diagnosis (Electronic Edition)*, vol. 17, no. 3, pp. 9–12, 2025.

[20] Z. Wei, L. Wang, and J. Mao, "Analysis of 333 Cases of NT Thickening and Prenatal Diagnosis Results," *China Journal of Prenatal Diagnosis (Electronic Edition)*, vol. 17, no. 3, pp. 21–26, 2025.