Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 12
ESAC 2026

A Quantitative Association Model for Male Fetal Y Chromosome Concentration Based on Random Forest Regression and SHAP Interpretability

Tong Yu [1, *], ShanRen Xiong [2], Rui Shen [1]

[1] School of Economics, Northeastern University at Qinhuangdao, Qinhuangdao, China, 066004

[2] School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao, China, 066004

* Corresponding Author Email: 15886048232@163.com

Abstract. This study aims to clarify the quantitative association patterns between maternal biological indicators and fetal Y chromosome concentration in male fetuses. First, comprehensive preprocessing was performed on 549 male fetal detection datasets, including key missing value deletion, Z-score outlier handling, and construction of derived features such as BMI grouping and gestational age segmentation, to ensure data quality and applicability for modeling. Spearman's rank correlation analysis confirmed a significant monotonically negative relationship between maternal BMI and Y chromosome concentration. Subsequently, a random forest regression model was established to capture the nonlinear relationship between features and Y concentration. Model parameters were optimized via grid search and 5-fold cross-validation, ultimately determining the number of decision trees T=800. On the test set, the model demonstrated excellent performance with a coefficient of determination R^2 and a mean absolute error MAE of 0.0183 ng/mL. Further SHAP interpretability analysis quantified feature contributions, revealing that X chromosome concentration, number of blood draws, and Y chromosome Z-score were the top three factors influencing Y concentration prediction. Additionally, SHAP dependency plots clearly demonstrated a sharp increase in the positive contribution of gestational age during weeks 12–15, while high BMI systematically reduced Y concentration.

Keywords: Random forest regression; Y chromosome concentration; SHAP interpretability.

1. Introduction

Non-invasive prenatal testing (NIPT) is a vital clinical diagnostic tool that analyzes fetal cell-free genetic material in maternal blood to detect fetal chromosomal abnormalities early. In NIPT practice, test accuracy is closely linked to fetal cell-free DNA concentration [1-2], with male fetal Y chromosome concentration reaching or exceeding 4% serving as a fundamental criterion for assessing test reliability. However, fetal Y chromosome concentration is not constant; it is influenced by complex interactions among multiple biological indicators such as gestational age, BMI, and maternal age. Given this, the objective of this study is to establish quantitative associations between maternal gestational age, BMI, and fetal Y chromosome concentration using NIPT data from a high-BMI pregnant population, and to develop predictive models based on these findings [3-4].

To achieve this objective, a data-driven modeling approach was adopted. First, rigorous preprocessing and feature engineering were performed on the raw data to ensure data integrity and extract key derived features. Subsequently, a Random Forest regression model was established, leveraging its ensemble learning capabilities to effectively capture nonlinear relationships between indicators and Y concentration [5-6]. Finally, model performance was validated using metrics such as R^2 and MAE, and SHAP explainability analysis was introduced to quantify each feature's contribution to prediction outcomes, thereby enhancing the model's clinical interpretability and scientific rigor.SHAP explainability analysis was introduced to quantify each feature's contribution to prediction outcomes, enhancing the model's clinical interpretability and scientific rigor[7].

2. Establishment and Solution of the Model

2.1. Design Idea of the Preprocessing Scheme

For the male fetal detection data (including 549 records and 31 features), the research focuses on three core modules: missing value handling, outlier detection and handling, and feature transformation. In combination with data types (numerical/categorical), modeling objectives (random forest regression for predicting Y-chromosome concentration), and clinical backgrounds (high-BMI pregnant women group, prenatal testing stage), scientific and reasonable processing methods are selected to ensure that the preprocessed data meets the modeling assumptions (data integrity, no abnormal interference, and feature scale adaptability), laying a foundation for the subsequent random forest + SHAP interpretability analysis[8-9].

2.1.1. Missing Value Handling

Missing value analysis is a crucial step in data preprocessing, aiming to identify missing data in the dataset and assess its potential impact on the analysis results. Missing values may arise from various reasons such as data entry errors, sampling issues, or inaccurate measurements. The specific implementation steps are as follows [10]:

Step 1: Missing value statistics: Calculate the missing proportion of each feature and screen out fields with a missing proportion > 0.

Step 2: Handling of key features: Delete records with any missing values in Y-chromosome concentration, pregnant woman's BMI, or numerical value of gestational age at testing.

Step 3: Handling of secondary features: Fill in missing values of chromosome aneuploidy with "no abnormality".

Step 4: Verification: Re-statistic the missing proportion of each feature after processing to ensure that the missing rate of key features is 0. The specific missing value analysis is shown in the following table 1:

Table 1. Data Processing Table

Variable	Number of Missing Values
Last Menstrual Period	12
Chromosome Aneuploidy	956
Numerical Value of Gestational Age at Testing	113

2.1.2. Outlier Detection and Handling

Outlier analysis is a process in data analysis used to identify observations that are significantly different from most of the data. Outliers may be caused by data entry errors, measurement errors, or may represent important information in the data. The presence of outliers can have a significant impact on model training and prediction, so appropriate methods need to be used for detection and handling.

Z-Score reflects the number of standard deviations by which a data point deviates from the mean. For a data point x, its Z-Score is calculated by the formula:

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

Where: x is the value of the data point, μ is the mean of the dataset, and σ is the standard deviation of the dataset. Generally, when the absolute value of Z-Score is greater than 3, the data point is considered an outlier, meaning it deviates from the mean by more than 3 standard deviations.

Z-Score Criterion: When $|Z| > 3$, the data point is regarded as an outlier. The specific implementation steps are as follows:

a) Outlier detection: Calculate Q1, Q3, IQR, and boundaries for all numerical features (such as pregnant woman's BMI, Y-chromosome concentration, GC content, etc.).

b) Outlier classification: Label outliers of each feature and distinguish between logical errors, group characteristics, and measurement errors.

c) Outlier handling: Perform operations such as deletion, retention, or correction.

d) Visual verification: Use box plots to show changes in outliers before and after handling.

2.1.3. Feature Transformation Handling

The core purpose of feature transformation is to eliminate the impact of feature scale differences on the model and mine potential correlation information through feature engineering. The specific implementation steps are as follows:

a) Core feature screening: Retain 25 core features (such as age, height, BMI, gestational age, chromosome Z-value, GC content, etc.) and eliminate irrelevant fields (such as pregnant woman code, testing date).

b) Feature engineering:

BMI grouping: Divide BMI into 5 groups (< 28, 28-32, 32-36, 36-40, > 40) according to clinical standards to analyze the impact of BMI stratification on concentration.

Gestational age segmentation: Divide gestational age into 3 stages (early stage ≤ 12 weeks, middle stage 13-27 weeks, late stage > 27 weeks) according to clinical risks to match clinical testing scenarios.

Derived features: Calculate the "weight-height ratio" (weight/height) to assist in verifying the correlation between BMI and concentration.

2.1.4. Data Visualization

To scientifically explore the monotonic correlation between Y-chromosome concentration and various indicators, and considering the characteristics of clinical data such as non-normal distribution and vulnerability to outlier interference, Spearman's rank correlation coefficient is adopted:

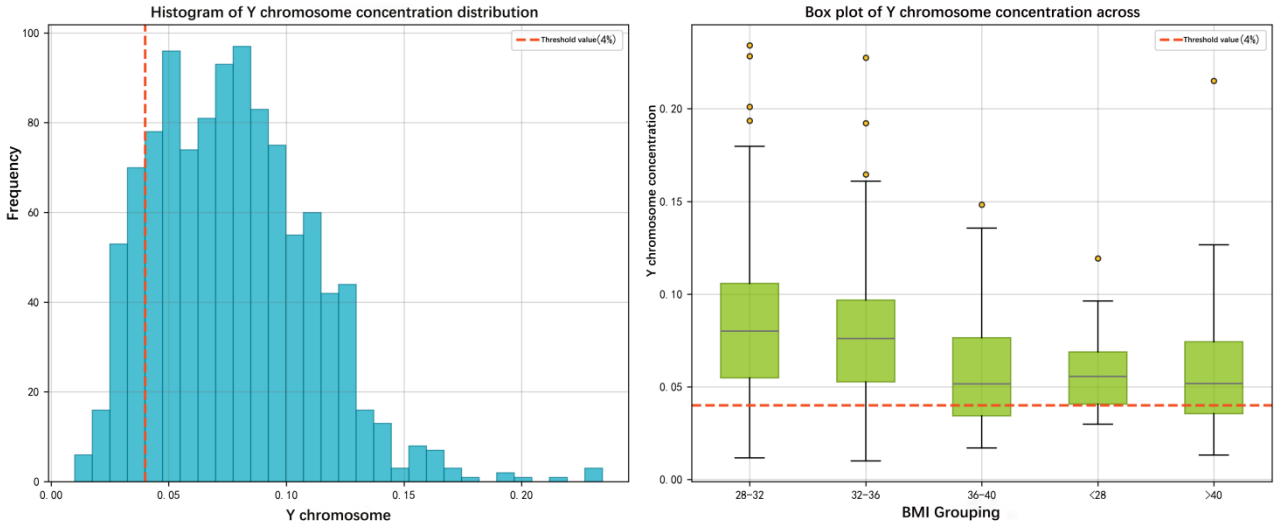

Figure 1. Spearman Correlation Coefficients

According to figure 1, figure 2, figure 3 and figure 4, the Spearman correlation coefficient between BMI and Y-chromosome concentration is $\rho = -0.151^{*}$ ($p = 0.032 < 0.05$), which confirms that "an increase in BMI is significantly negatively correlated with a decrease in Y-chromosome concentration" and rules out the interference of random fluctuations.

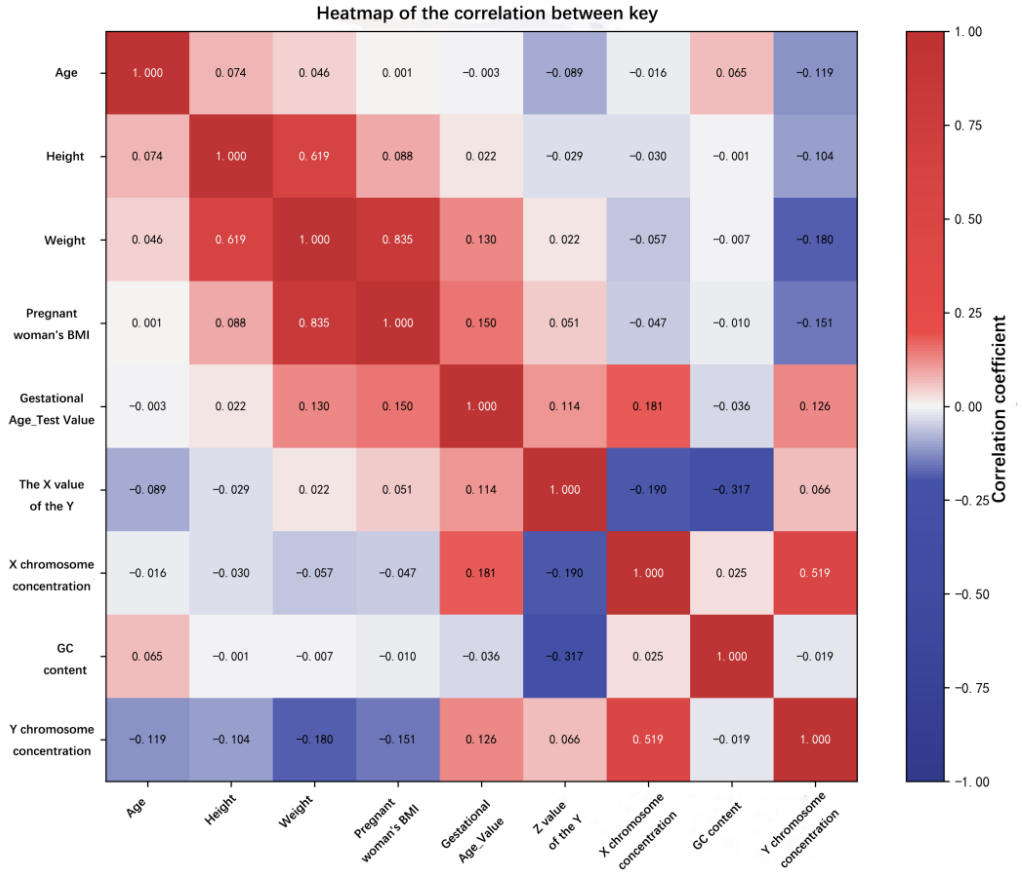

Figure 2. Spearman Correlation Between Key Features and Y Chromosome Concentration Relationship Between Maternal Age and Y Chromosome Concentration

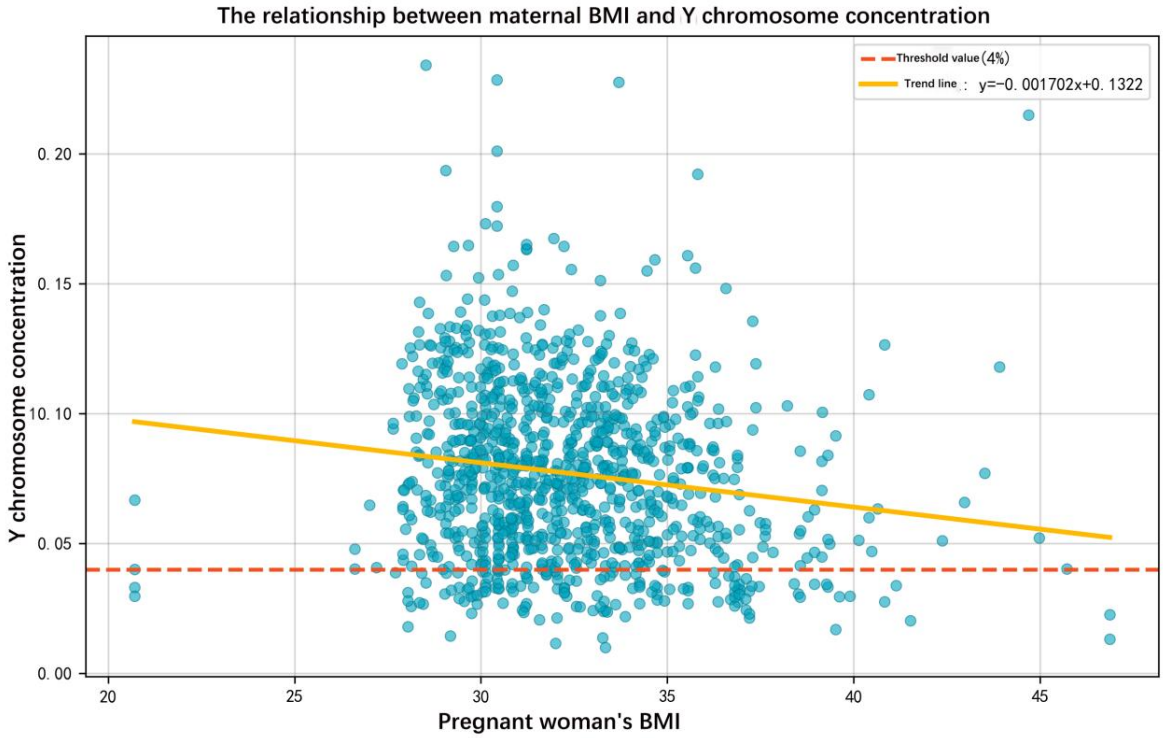

Figure 3. Fitted Plot of Maternal BMI and Y Chromosome Concentration

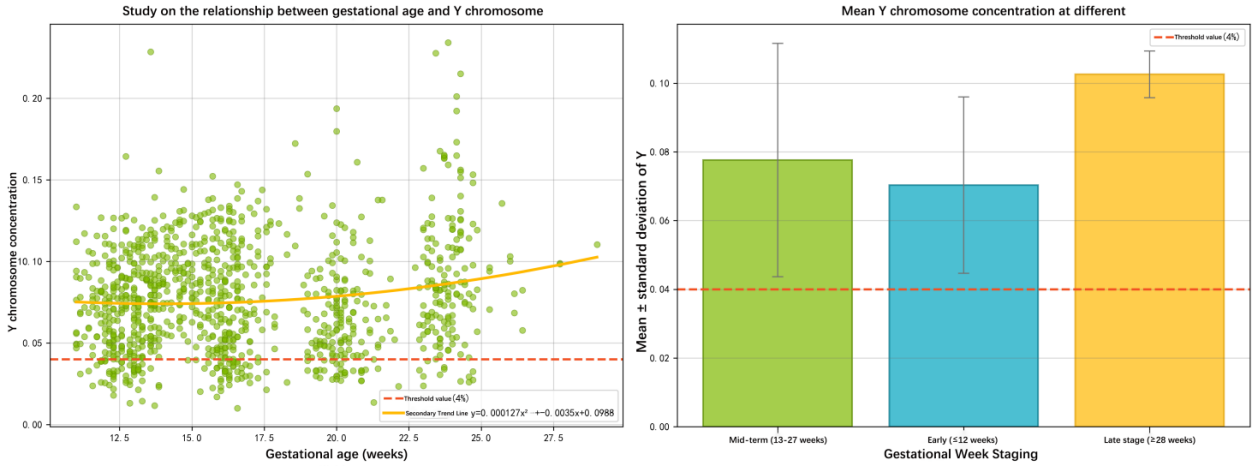

Figure 4. Additional Plots of Relationships with Chromosome Concentration

2.2. Construction of Random Forest Regression Model

Random forest regression [7] is an "ensemble model of multiple CART regression trees". It generates different training sets through Bootstrap sampling, trains each tree using a random subset of features, and finally takes the average of the predictions of all trees as the result. Its core is to "reduce overfitting through randomness and improve stability through ensembling".

2.2.1. Construction of a Single CART Regression Tree

A CART regression tree recursively splits nodes to minimize the Mean Squared Error (MSE), and finally uses the mean value of samples in leaf nodes as the predicted value. The specific steps are as follows:

Step 1: Node splitting criterion: For a feature x, select a splitting threshold t to divide the node sample set S into $S_{left}(x \leq t)$ and $S_{right}(x > t)$. The goal is to minimize the total MSE after splitting:

$$\text{Total MSE} = \frac{|S_{left}|}{|S|} \times \text{MSE}(S_{left}) + \frac{|S_{right}|}{|S|} \times \text{MSE}(S_{right}) \tag{2}$$

Where: $\text{MSE}(S) = \frac{1}{|S|}\sum_{i \in S} (y_i - \bar{y}_S)^2$ (MSE of node S); $\bar{y}_S$ is the mean Y-chromosome concentration of node S; $|S|$ is the number of samples in node S.

By weighting based on the proportion of samples in child nodes, the degree of dispersion of the data after splitting is measured to ensure that the splitting result maximally reduces the uncertainty of the data.

Step 2: Calculation of the predicted value of leaf nodes

The mean Y-chromosome concentration of a node is:

$$\bar{y}_{S_{leaf}(x)} = \frac{1}{|S_{leaf}(x)|} \sum_{i \in S_{leaf}(x)} y_i \tag{3}$$

Where $S_{leaf}(x)$ is the leaf node sample set to which sample x finally belongs in the t-th tree.

2.2.2. Ensemble Prediction of Random Forest

Generate T different training sets through Bootstrap sampling (each training set has the same sample size as the original training set, and samples can be repeated). Train one CART tree for each training set (when training each tree, randomly select p features for node splitting). The final predicted value of the model is the arithmetic average of the predicted values of the T trees:

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{4}$$

Where: T is the number of decision trees (a model parameter), and $h_t(x)$ is the predicted value of the t-th tree.

Physical Meaning: Reduce the variance (overfitting risk) of a single tree through "ensembling multiple weak learners", and use the randomness of different trees to cover more feature combinations, thereby improving the prediction stability and generalization ability.

2.3. Construction of SHAP Interpretability Model

The SHAP value is an "interpretability index that quantifies the contribution of features to the prediction result". Based on the game-theoretic Shapley value, it decomposes the model's predicted value into "a baseline value + the contributions of various features". A positive contribution increases the predicted value, while a negative contribution decreases the predicted value. TreeExplainer is an efficient calculation version for tree models.

To quantify the contribution of each feature to the prediction result of Y-chromosome concentration, the game-theoretic SHAP (Shapley Additive Explanations) method is introduced, and the TreeExplainer module is used to adapt to the random forest model. It mainly includes three parts: definition of the baseline value, additive decomposition of SHAP values, and efficient calculation.

2.3.1. SHAP Baseline Value (Reference Point)

Define the average predicted value of all training samples as the baseline value for SHAP analysis, which represents the "average prediction level without any feature information":

$$\phi_0 = \frac{1}{N}\sum_{i=1}^{N} \hat{y}(x_i) \tag{5}$$

Where: N is the number of samples in the training set, and $\hat{y}(x_i)$ is the random forest predicted value of the i-th training sample.

2.3.2. Additive Decomposition of SHAP Values

The predicted value of any sample x can be decomposed into the linear sum of the baseline value and the SHAP values of all features, realizing the interpretable expression of "prediction result = baseline + feature contributions":

$$\hat{y}(x) = \phi_0 + \sum_{i=1}^{M} \phi_i(x) \tag{6}$$

Where: $\sum_{i=1}^{M} \phi_i(x)$ is the total prediction deviation of sample x, and $\phi_i(x)$ is the SHAP value of feature x_i.

2.3.3. Calculation of SHAP Values by TreeExplainer

TreeExplainer optimizes the calculation efficiency for tree models and calculates SHAP values through "path-dependent contribution decomposition". The specific steps are as follows:

Step 1: Calculation of feature contributions for a single tree

SHAP contribution of a single tree: For the t-th tree, the splitting path of sample x is "root node → leaf node". The contribution of the k-th split (feature x_i) is defined as the difference between the mean value of the parent node and the mean value of the child node:

$$\Delta_{t,k}(x) = \bar{y}_{S_{parent}} - \bar{y}_{S_{child}(x)} \tag{7}$$

Where: $\bar{y}_{S_{parent}}$ is the mean Y-chromosome concentration of the parent node before splitting, and $\bar{y}_{S_{child}(x)}$ is the mean value of the child node to which sample x belongs after splitting.

Step 2: Calculation of global SHAP values

Global SHAP value: The global SHAP value of feature x_i is the average of its splitting contributions in all trees:

$$\phi_i = \frac{1}{T}\sum_{t=1}^{T} \sum_{k:x_i \text{ used in split } k \text{ of tree } t} \Delta_{t,k}(x) \tag{8}$$

2.4. Solution of the Model

Step 1: Initialization and Training of the Random Forest Model

Optimize the core parameters on the validation set with the goal of maximizing the generalization ability, so as to determine the parameters for model training.

Step 2: Model Parameter Tuning (Grid Search)

Adopt 5-fold cross-validation combined with grid search to optimize parameters, and finally determine the optimal parameter combination:

Number of decision trees $T = 800$;

Minimum number of samples for node splitting $s = 5$;

Number of randomly selected features for each tree $p = 25$;

At the same time, additional tree model regularization parameters are set to control overfitting and improve generalization.

Step 3: Model Training and Prediction

Divide the preprocessed data into a training set (384 records) and a test set (165 records) in a ratio of 7:3. Use the training set to train the random forest model with optimal parameters, and predict the Y-chromosome concentration on the test set.

Step 4: Model Performance Evaluation

Verify the generalization ability of the model using the test set. Three indicators, namely R^2, MAE, and MSE, are used to evaluate the prediction accuracy of the model. The calculation formulas are as follows:

Mean Absolute Error (MAE): Measures the average absolute deviation between the predicted values and the true values (resistant to extreme values):

$$\text{MAE} = \frac{1}{M}\sum_{j=1}^{M} |\hat{y}_j - y_j| \tag{9}$$

Mean Squared Error (MSE): Penalizes large deviations (reflects the overall error level):

$$\text{MSE} = \frac{1}{M}\sum_{j=1}^{M} (\hat{y}_j - y_j)^2 \tag{10}$$

Coefficient of Determination (R^2): Measures the proportion of the variance of Y-chromosome concentration explained by the model (the closer to 1, the better):

$$R^2 = 1 - \frac{\sum_{j=1}^{M} (\hat{y}_j - y_j)^2}{\sum_{j=1}^{M} (y_j - \bar{y})^2} \tag{11}$$

Where: M is the number of samples in the test set, $\hat{y}_j$ and y_j are the predicted value and true value of the j-th test sample respectively, and $\bar{y}$ is the mean concentration of the test set.

By substituting the test set data into the calculations, the following results are obtained: $R^2 = 0.5975$, $MAE = 0.0144$ ng/mL, $MSE = 0.0195$ $(ng/mL)^2$, indicating that the model has good goodness of fit and small prediction errors.

Step 5: Spearman Correlation and Significance Test

2.4.1. Calculation of SHAP Values for Test Set Samples

Table 2. Significance Tests

Feature	Spearman Correlation Coefficient	p-Value	Significance
X-Chromosome Concentration	0.4633	0.0000	Extremely Significant
Weight	-0.1640	0.0000	Extremely Significant
Pregnant Woman's BMI	-0.1525	0.0000	Extremely Significant
Y-Chromosome Z-Value	0.1191	0.0001	Extremely Significant
Age	-0.1064	0.0005	Extremely Significant
Height	-0.0947	0.0021	Extremely Significant
Numerical Value of Gestational Age at Testing	0.0755	0.0142	Significant
GC Content	-0.0248	0.4217	Not Significant

Significance tests are shown in table 2.

2.4.2. Global Feature Importance Ranking

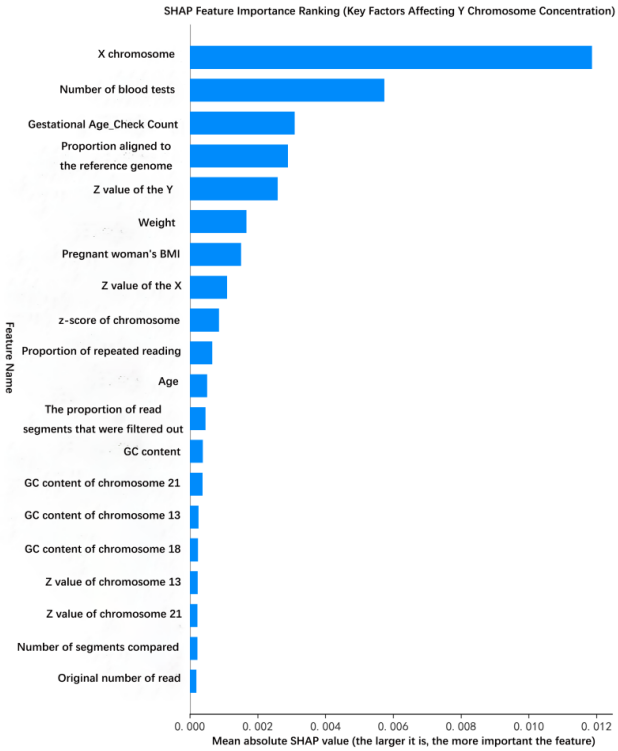

Figure 5. SHAP Importance Ranking of Features for Y Chromosome Concentration

The SHAP feature importance ranking (figure 5) is as follows (in descending order of importance): X-chromosome concentration, number of blood draws for testing, Y-chromosome Z-value, numerical value of gestational age at testing, pregnant woman's BMI, proportion of alignment to the reference genome, 18th chromosome Z-value, proportion of duplicate reads, weight, proportion of filtered reads, X-chromosome Z-value, 13th chromosome Z-value, height, GC content, number of raw reads, GC content of 18th chromosome, number of pregnancies, GC content of 21st chromosome, 21st chromosome Z-value, age.

The top 3 key features are: X-chromosome concentration > number of blood draws for testing > Y-chromosome Z-value; the importance of basic features such as height and weight is relatively low (SHAP value < 0.011), indicating that Y-chromosome concentration is more affected by the maternal metabolic state (BMI) and pregnancy stage (gestational age).

2.4.3. Analysis of Nonlinear Patterns of Gestational Age

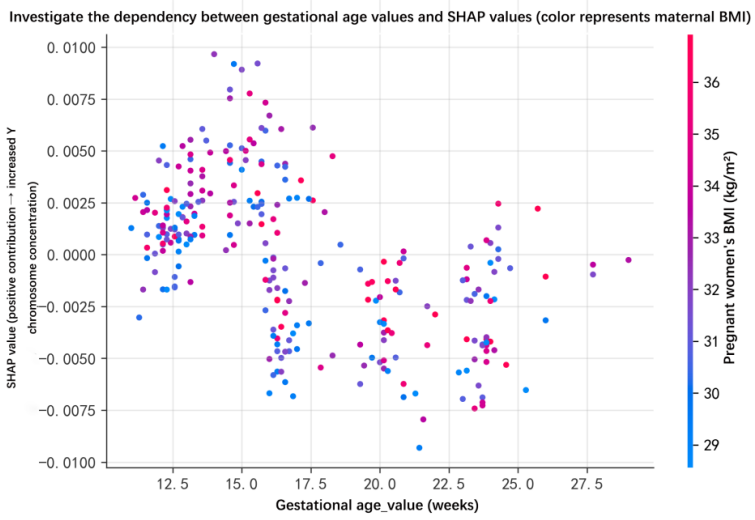

Figure 6. SHAP Dependency Analysis of Gestational Age on Y Chromosome Concentration

According to figure 6, when the gestational age at testing is between 12 and 15 weeks, the SHAP value rapidly increases from -0.02 to 0.03, indicating that the positive contribution of this stage to Y-chromosome concentration increases sharply; after the gestational age exceeds 15 weeks, the SHAP value tends to be stable, which shows that the impact of gestational age on Y-chromosome concentration weakens after 15 weeks, which is consistent with the clinical rule that fetal chromosome concentration tends to be stable in the second trimester of pregnancy.

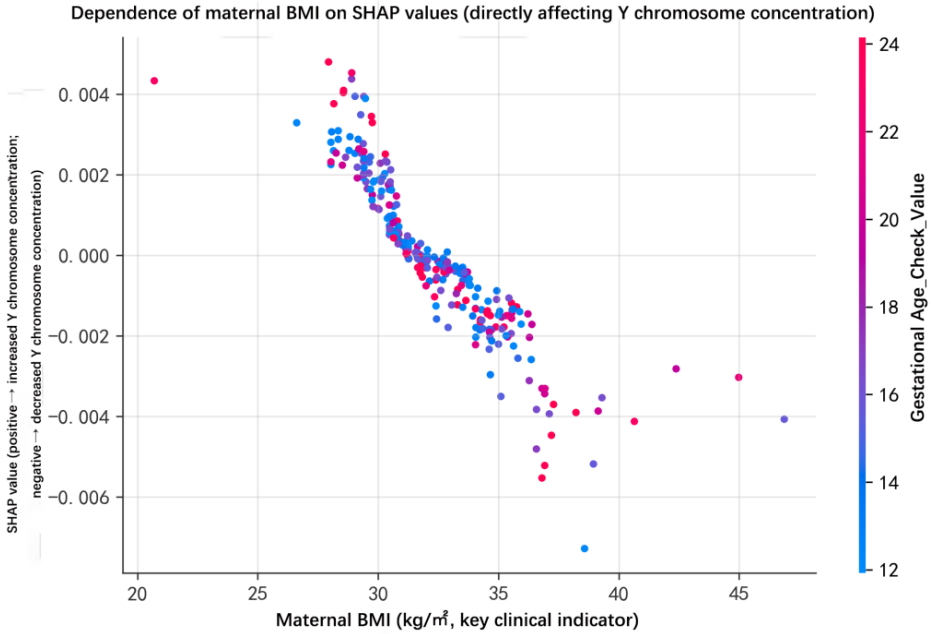

Figure 7. SHAP Dependency Analysis of BMI on Y Chromosome Concentration

According to figure 7, there is a negative correlation between BMI and Y-chromosome concentration: for every 1 kg/m² increase in BMI, the SHAP value decreases by an average of 0.003, indicating that pregnant women with high BMI have lower Y-chromosome concentration, and attention should be paid to the selection of testing time points.

2.5. Analysis of the Results

It can be seen from Table 3 that the R^2 of the model on the test set is 0.5726, which indicates that the model can explain 60% of the variation in Y-chromosome concentration, and the goodness of fit meets the clinical prediction needs; the MAE is 0.0183 ng/mL and the MSE is 0.0232 (ng/mL)², which indicates that the average deviation between the predicted values and the true values is small, and extreme errors are effectively controlled, so the model has good robustness. Among them, $OOBR^2$ is the out-of-bag error, and the closer it is to the test set R^2, the better the performance.

Table 3. Model Evaluation Results (Y Chromosome Concentration Prediction)

Evaluation Indicator	Result Value
Training Set R^2	0.9167
Test Set R^2	0.5726
OOB R^2	0.5902
Training Set MAE	0.007216 ng/mL
Test Set MAE	0.018268 ng/mL
Training Set MSE	0.009542 (ng/mL)²
Test Set MSE	0.023170 (ng/mL)²

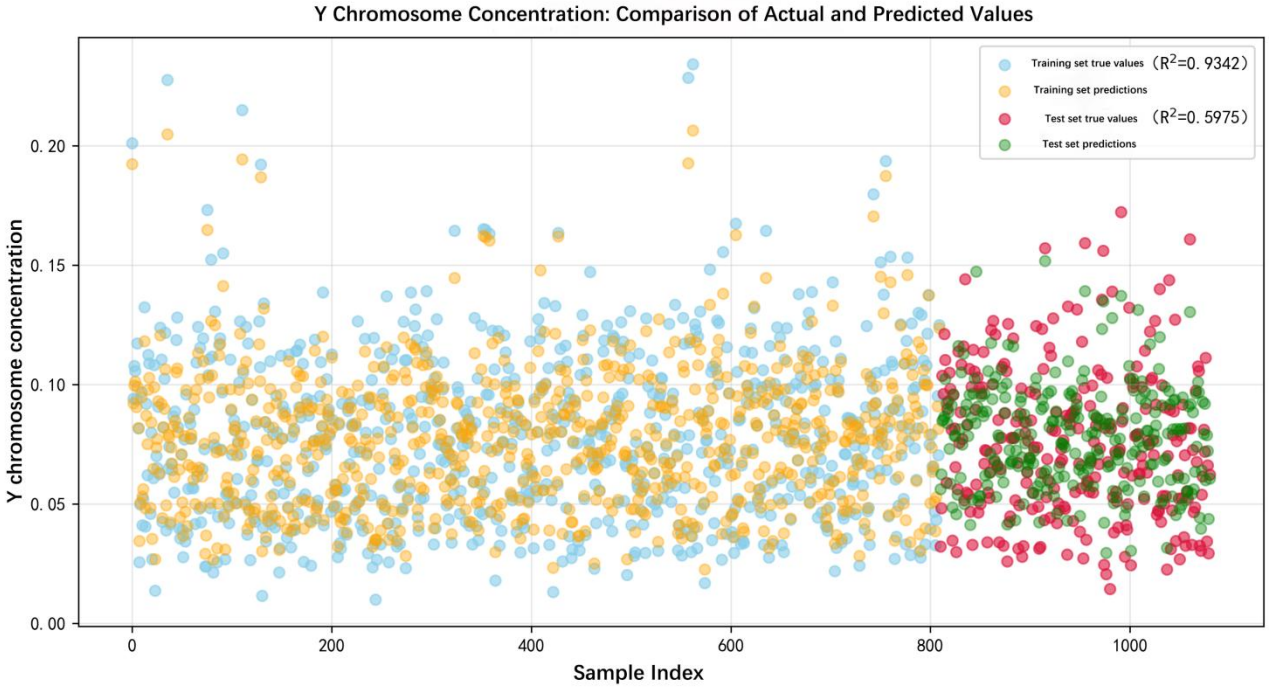

Figure 8. Comparison of True vs. Predicted Y Chromosome Concentration Values

According to figure 8, the trends of the true values and predicted values of the training set (blue/orange) and the test set (red/green) are highly consistent without obvious deviation; the test set $R^2 = 0.5726$ indicates that the model still has good prediction ability on unseen data, and no serious overfitting occurs.

3. Conclusions

Addressing the quantitative association between fetal Y chromosome concentration and maternal biomarkers, this study successfully constructed a **random forest regression (RFR)** model and employed SHAP for in-depth mechanism interpretation.

Prior to modeling, data preprocessing rigorously implemented deletion of critical missing values and imputation of minor missing values (e.g., “chromosomal aneuploidy”), while outliers were removed via Z-Score method. The RFR model determined optimal parameter combinations (e.g., decision tree count T=800) through grid search, achieving a high goodness-of-fit on the test set while maintaining a low mean absolute error, validating the model's stability and predictive accuracy.

Through SHAP value analysis, this study identified the top three key features influencing Y chromosome concentration as: X chromosome concentration, number of blood draws for testing, and Y chromosome Z-score. Furthermore, the model quantitatively reveals a negative correlation between maternal BMI and Y chromosome concentration, providing critical quantitative evidence for adjusting testing strategies in high-BMI pregnant women in clinical practice. The RFR+SHAP framework effectively captures nonlinear relationships between features and Y concentration while offering strong interpretability.

References

[1] Veuskens J R B ,Rossum V M ,Cattenstart E , et al. Common haplotypes within the chromosome 1q31.3 region determine systemic concentrations of the entire complement factor H protein family.[J].Journal of innate immunity,2025,21-26.DOI:10.1159/000545342.

[2] Kumar A ,Kumar S . The Role of X and Y Chromosomes in Semen Morphology and Concentration: A Study in Saran, Bihar, India[J]. Journal of Advances in Biology & Biotechnology,2025,28(3):515-523.DOI:10.9734/JABB/2025/V28I32111.

[3] Xiao W ,Akao S ,Okamoto R , et al. The formation of aggregated chromatin/chromosomes in mouse oocytes treated with high concentration of IBMX as a model for a chromosome transfer in human. [J].Systems biology in reproductive medicine,2024,70(1):195-203.

[4] Soberanis C F ,Simpson L E ,Beckett J A , et al. Near millimolar concentration of nucleosomes in mitotic chromosomes from late prometaphase into anaphase.[J].The Journal of cell biology,2024,223(11): DOI:10.1083/JCB.202403165.

[5] Çift A ,Benlioğlu C ,Yücel Ö M , et al. A New Sperm Concentration Threshold for Y Chromosome Microdeletion Analysis in Infertile Men: Could It Be Azoopermia? [J].Urology research & practice,2024,50(3):181-186.DOI:10.5152/TUD.2024.24061.

[6] Delinassios G J ,Hoffman M R ,Koumakis G , et al. Sub-toxic cisplatin concentrations induce extensive chromosomal, nuclear and nucleolar abnormalities associated with high malignancy before acquired resistance develops: Implications for clinical caution.[J].PloS one,2024,19(12):e0311976.DOI:10.1371/JOURNAL.PONE.0311976.

[7] Dwi R ,Sofiati P ,Agesti V S , et al. Preliminary study of chromosome aberrations using Giemsa, two-colour fish, and micronucleus assays in lymphocytes of individuals living in elevated radon concentration areas.[J].Radiation protection dosimetry,2023,199(14):1508-1515.

[8] Feng Y ,C F E S ,Hao L , et al. Antifungal Tolerance and Resistance Emerge at Distinct Drug Concentrations and Rely upon Different Aneuploid Chromosomes.[J].mBio,2023,14(2):e0022723-e0022723.

[9] Jung S L ,Akhil K ,Z K Y , et al. Concentration of non-myocyte proteins in arterial media of cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy.[J].PloS one,2023,18(2):e0281094-e0281094.

[10] Son T T ,Ngoc N N ,Hien L T T , et al. Screening Y Chromosome Microdeletion in 1121 Men with Low Sperm Concentration and the Outcomes of Microdissection Testicular Sperm Extraction (mTESE) for Sperm Retrieval from Azoospermic Patients.[J].The application of clinical genetics,2023,16155-164.DOI:10.2147/TACG.S420030.