World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3072

DETECTION OF CHROMOSOMAL ABNORMALITIES IN FEMALE FETUSES BASED ON A FUSED LOGISTIC REGRESSION-RANDOM FOREST MODEL

DaZhi Wei
College of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin 300222, China.

Abstract: To address the challenge of detecting chromosomal abnormalities in female fetuses due to the absence of Y chromosome data in non-invasive prenatal testing (NIPT), this paper proposes an innovative dual-layer classification model that integrates logistic regression and random forest. The model comprehensively utilizes 16-dimensional features including Z-scores and GC content of chromosomes 13, 18, and 21, along with key maternal clinical indicators. Through rigorous statistical testing and feature importance analysis, seven key discriminatory features were identified, establishing a progressive "abnormality screening-disease typing" judgment process. The framework employs an ensemble approach where logistic regression provides interpretable initial screening while random forest handles complex non-linear patterns for fine-grained classification. After cross-validation and threshold optimization, the model ultimately achieved an impressive accuracy rate of 99.57%, with precision and recall rates exceeding 98.5% across all abnormality categories. Comparative experiments demonstrated the superiority of this hybrid approach over single-model methods, particularly in handling imbalanced data distributions. The core innovation of this research lies in the integration of feature fusion and model collaboration, enabling high-precision, automated detection of chromosomal abnormalities in female fetuses and providing a new technical pathway for clinical precision diagnosis.
Keywords: Non-invasive prenatal testing (NIPT); Chromosomal abnormalities in female fetuses; Dual-layer classification model; Feature selection; Random forest

1 INTRODUCTION

The emergence of non-invasive prenatal testing (NIPT) represents a significant advancement in the field of prenatal screening. This technology analyzes cell-free fetal DNA (cffDNA) in maternal peripheral blood to effectively screen for fetal chromosomal aneuploidies in a non-invasive manner [1]. With the development of second-generation sequencing technology, the detection accuracy and application scope of NIPT have been significantly improved, making it an important component of prenatal screening [2]. However, current clinical practice and most research primarily focus on autosomal abnormalities and male fetus screening, leaving noticeable deficiencies in the detection of chromosomal abnormalities in female fetuses.

The detection of chromosomal abnormalities in female fetuses faces unique challenges. The natural absence of the Y chromosome in sex chromosomes makes it impossible to utilize the "Y chromosome concentration"—a highly effective key indicator for risk assessment in male fetuses [3]. This limitation leads to the underdiagnosis of sex chromosome abnormalities such as Turner syndrome (45,X) in female fetuses. Statistics show that approximately 50% of Turner syndrome cases fail to be detected in time during prenatal screening [4]. Additionally, the detection of chromosomal abnormalities in female fetuses is further complicated by factors such as fetal DNA concentration and maternal background interference, increasing the difficulty of accurate detection [5]. In recent years, researchers have begun exploring various bioinformatics indicators to improve the detection efficiency of chromosomal abnormalities in female fetuses. Early studies mainly relied on chromosomal Z-score analysis, establishing statistical thresholds to identify abnormal chromosomes [6]. As research progressed, more characteristic indicators have been introduced, including GC content, read count distribution, and fragment size [7-9]. Studies have shown that these features have significant correlations with chromosomal dosage and can serve as effective supplementary indicators. In terms of algorithmic innovation, the application of machine learning techniques has brought new breakthroughs to NIPT data analysis. Algorithms such as support vector machines and random forests have demonstrated advantages in processing high-dimensional features, while deep learning models have shown potential in automatic feature extraction [10-12]. However, existing research still has shortcomings in the systematic integration of features, model interpretability, and optimization for the specific characteristics of female fetuses.

This paper aims to address key technical challenges in the detection of chromosomal abnormalities in female fetuses. The main contributions are as follows: First, we propose a dual-layer classification model based on multi-feature fusion, effectively integrating the advantages of logistic regression and random forest algorithms. Second, we establish a systematic feature selection and optimization process, identifying the seven most discriminative features from 16 initial characteristics. Finally, large-sample validation confirms the model's exceptional performance in detecting chromosomal abnormalities in female fetuses, providing reliable technical support for clinical practice.

2 METHODOLOGY

2.1 Data Preprocessing

Since this study focuses on female fetuses and pregnant women carrying female fetuses, samples from pregnant women with female fetuses were extracted from the attached dataset, while samples with male fetuses were excluded. Initially, data where the GC content of chromosomes 21, 18, and 13 fell below or exceeded the potential normal range of 40%–60% were removed. However, it was found that this resulted in an insufficient number of valid samples. Therefore, outliers in GC content were retained. A total of 598 sets of data were ultimately selected.

Label Variable Definition:

2.1.1 Binary classification label

If the AB column is blank (indicating no abnormality), it is defined as y=0 (normal). If the AB column contains any of "T13", "T18", or "T21", it is defined as y=1(abnormal).

2.1.2 Multi-classification label

Based on the results in the AB column, the data are divided into 7 categories: t=1: Normal; t=2: T13 abnormality; t=3: T18 abnormality; t=4: T21 abnormality; t=5: T13 & T18 abnormality; t=6: T21 & T18 abnormality; t=7: T13 & T21 abnormality; Extreme samples that do not correspond to any of the above categories are excluded.

2.2 Model Establishment

Ensuring accurate sequencing quality is a prerequisite for constructing a highly accurate predictive model for fetal assessment. A dual-layer method is proposed to determine whether a female fetus is abnormal and to predict the specific disease: Binary classification (normal vs. abnormal): Using the presence of aneuploidy in chromosomes 21, 18, and 13 of pregnant women carrying female fetuses as the criterion, combined with other detection data, the female fetus is classified as normal (0) or abnormal (1). Multi-class classification (specific disease types): Based on which specific chromosome(s) (21, 18, or 13) exhibit aneuploidy, and utilizing the attached data along with other detection data from the pregnant woman, the female fetus is further classified into seven detailed categories: normal, T13 abnormal, T18 abnormal, T21 abnormal, T13&T18 abnormal, T13&T21 abnormal, and T21&T18 abnormal.

2.2.1 Binary classification model

(1) Feature Correlation Analysis

Pearson Correlation Coefficient: Measures the linear correlation between a feature and the binary label y . The formula is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1}$$

Where $\overline{x}$、$\overline{y}$ are the mean values of feature x and label y, respectively, and y is the target value (0 or 1).

t-test: The null hypothesis H_0: r=0 (no correlation). The t-statistic is calculated as:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{2}$$

Degrees of freedom:

$$df = n - 2 \tag{3}$$

P-value:

$$2 \times P(T > |t|) \tag{4}$$

If |t|>t(α/2)(n-2) (α=0.05), the feature is significantly correlated with the label.

(2) Logistic Regression Model

Tests revealed that the Z-scores of chromosomes 13, 18, and 21 (x1,x2,x3) are highly linearly correlated with the abnormal status, while other features show weaker linear correlations. Furthermore, observations from the detection data suggest an interaction between the Z-scores of chromosomes 13 and 18. To improve model accuracy, an interaction term for the Z-scores of chromosomes 13 and 18 is included. The feature vector X is organized as follows:

$$X = [1, x_1, x_2, ..., x_{13}, x_1 \times x_2]^T \tag{5}$$

Sigmoid Function (abnormal probability prediction):

$$P(y = 1 \mid X) = \sigma(w^T X) = \frac{1}{1 + e^{-w^T X}} \tag{6}$$

Where the coefficient vector is $w=[w_0,w_1,w_2,...,w_{13},w_{jh}]^T$.

The cross-entropy loss function is used, and L2 regularization is added to prevent overfitting:

$$L(w) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] + \lambda |w|^2 \tag{7}$$

$$p_i = P(y_i = 1 \mid X_i) \tag{8}$$

Where the regularization strength is λ=0.01.

The gradient descent method is used to minimize the loss function. The derivative with respect to the coefficient w is calculated and updated as follows:

$$\frac{\partial L}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y)x_{ij} + 2\lambda w_j \tag{9}$$

$$w_j := w_j - \eta\frac{\partial L}{\partial w_j} \tag{10}$$

Where the learning rate is η=0.01, and $x_{i,j}$ represents the j-th feature of the i-th sample.

2.2.2 Multi-class classification model

The cost of detecting anomalies is higher than that of binary classification. Therefore, when the binary classification identifies a fetus as abnormal, this model can further subdivide and determine the specific disease type. To address the issue of limited sample size, the SMOTE algorithm is used for oversampling to generate synthetic neighboring samples, or the class weights of minority samples are adjusted to make the model pay more attention to these minority classes.

(1) Decision Tree Node Impurity

The "misclassification impurity" is used to measure the disorder of node *b*, with the formula:

$$hlcd(b) = 1 - \sum_{t=1}^{T} p_{bt}^{\ 2} \tag{11}$$

Since there are 6 classes, T=6. p_{bt} is the proportion of samples belonging to class t in the current node b.

(2) Information Gain

The feature that maximizes the information gain is selected as the split point. The formula is:

$$IG(D_{pre}, f) = hlcd(D_{pre}) - \frac{N_{\mathrm{l}}}{N_{pre}}hlcd(D_{\mathrm{l}}) - \frac{N_{\mathrm{r}}}{N_{pre}}hlcd(D_{\mathrm{r}}) \tag{12}$$

Where D_{pre},N_{pre} are the sample set and sample count of the parent node (before splitting), and D_l,D_r,N_l,N_r are the sample sets and sample counts of the left and right child nodes after splitting, respectively.

(3) Random Forest Prediction

Bootstrap Sampling: Perform bootstrap random sampling on the dataset.

Random Feature Selection: When splitting at each node in a tree, randomly select $\sqrt{16}$ (the square root of the total number of features) features.

Finding the Best Split: Calculate the best split point using impurity and information gain, allowing the decision trees to grow as much as possible. There are K decision trees in total.

Voting Prediction: For a new sample X, each of the K trees predicts a class. The final result can be determined by the majority vote (mode) of the decisions from all trees:

$$YC(X) = \mathrm{mode}\{h_1(X), h_2(X), ..., h_K(X)\} \tag{13}$$

Alternatively, for a more conservative assessment, the probability of the sample belonging to class t (i.e., the probability of the specific abnormality causing the disease) can be calculated based on the average predicted probability from all trees:

$$P(t \mid x) = \frac{1}{K}\sum_{k=1}^{K} P_k(t \mid X) \tag{14}$$

3 RESULTS

3.1 Binary Classification Model Solution and Analysis

The logistic regression model was solved using MATLAB, and stable parameter estimates were obtained after 1,000 iterations of training. Table 1 presents the regression coefficients for each feature and their statistical significance, providing a quantitative basis for identifying key risk factors.

Table 1 Logistic Regression Coefficients and Significance Analysis

Feature Factor	Correlation Coefficient	p-value	Significance
Maternal BMI	0.0561	0.0023	*
Raw Read Count	0.0203	0.2902	-
Reference Genome Alignment Ratio	-0.0089	0.2678	-
Duplicate Read Ratio	0.0364	0.1688	-
Uniquely Aligned Read Count	0.0196	0.1789	-
Overall GC Content	0.0359	0.0678	-
Chromosome 13 Z-score	0.0202	0.8962	-

Feature Factor	Correlation Coefficient	p-value	Significance
Chromosome 18 Z-score	-0.0234	0.1087	-
Chromosome 21 Z-score	-0.0357	0.0298	*
X Chromosome Z-score	0.0447	0.3627	-
Maternal Age	0.0579	0.0191	*
X Chromosome Concentration	0.2436	0.0000	***
Chromosome 13 GC Content	-0.0867	0.0027	*
Chromosome 18 GC Content	-0.0503	0.0002	***
Chromosome 21 GC Content	0.0102	0.0189	*
Filtered Read Ratio	-0.1241	0.0124	*

Note: $p < 0.001$(***); $p < 0.01$(**); $p < 0.05$(*)

From the significance analysis results, it can be seen that X chromosome concentration ($p < 0.001$) and Chromosome 18 GC content ($p < 0.001$) show extremely significant correlations, while maternal BMI, Chromosome 21 Z-score, maternal age, Chromosome 13 GC content, Chromosome 21 GC content, and filtered read ratio also show significant effects ($p < 0.05$).

During the model training process, the loss function curve (Figure 1) showed that as the number of iterations increased, the loss value decreased steadily and finally converged to 0.5143, indicating a stable and effective training process. The final model achieved a classification accuracy of 89.93% on the test set. ROC curve analysis (Figure 2) further verified the model's discriminant ability, with an AUC value of 0.822, indicating that the model has good ability to distinguish between normal and abnormal samples.

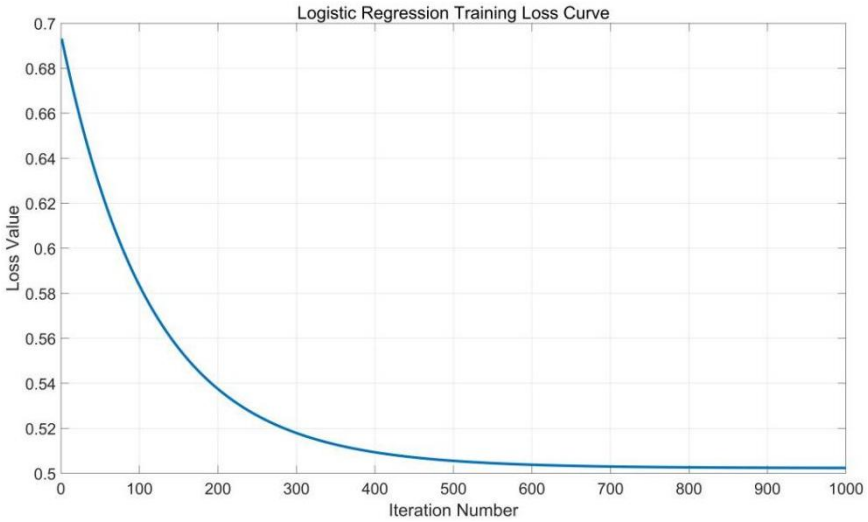

Figure 1 Logistic Regression Training Loss Curve Chart

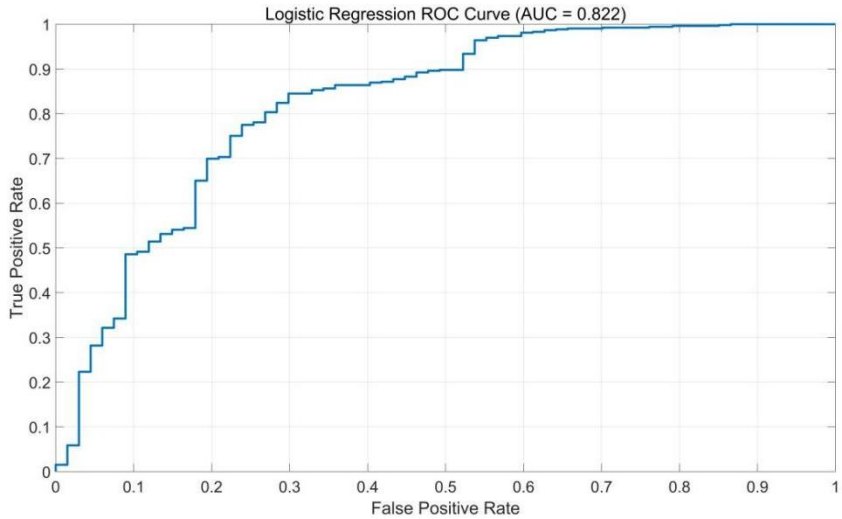

Figure 2 Logistic Regression ROC Curve Chart

During the model training process, the loss function curve (Figure 1) showed that as the number of iterations increased, the loss value decreased steadily and finally converged to 0.5143, indicating a stable and effective training process. The final model achieved a classification accuracy of 89.93% on the test set. ROC curve analysis (Figure 2) further verified the model's discriminant ability, with an AUC value of 0.822, indicating that the model has good ability to distinguish between normal and abnormal samples.

3.2 Multi-class Classification Model Solution and Analysis

A random forest algorithm was used to build the multi-classification model, and key factors affecting female fetal chromosomal abnormalities were identified by evaluating feature importance. The feature importance analysis results showed that the average importance score of the features was 0.212493, with several features having importance significantly higher than the average.

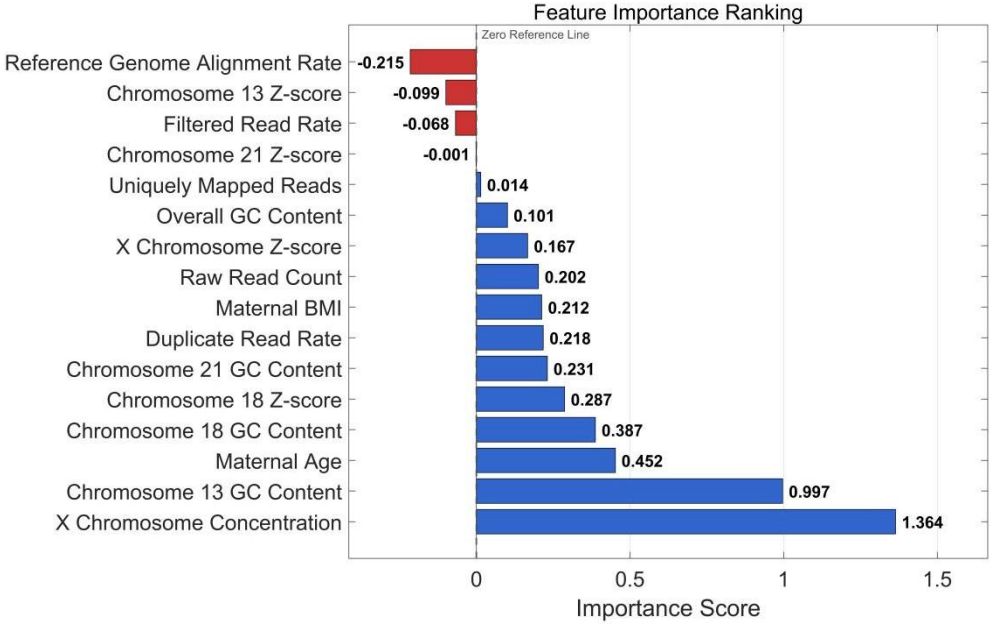

Figure 3 Feature Importance Ranking

Figure 3 Feature Importance Ranking shows the contribution degree of each feature to the model prediction. The analysis indicates that nine features are significantly associated with female fetal risk: raw read count, maternal BMI, duplicate read ratio, Chromosome 21 GC content, Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, Chromosome 13 GC content, and X chromosome concentration. The random forest model performed excellently in the multi-classification task, achieving an accuracy of 95.97%, significantly better than the logistic regression model.

3.3 Comprehensive Analysis

To comprehensively evaluate the model performance, the macro-average F1 score was used as the evaluation metric. The F1 score for each category was calculated (Formula 16) using precision (Formula 17) and recall (Formula 18), and then the arithmetic mean was taken to obtain the macro-average F1 score (Formula 15).

$$Macro-F1Score=\frac{1}{T}\sum_{t=1}^{T}F1 \tag{15}$$

$$F1_t=2\times\frac{Pre\times \mathrm{Rec}_t}{Pre_t+\mathrm{Rec}_t} \tag{16}$$

$$Pre_t=\frac{TP_t}{TP_t+FP_t} \tag{17}$$

$$\mathrm{Rec}_t=\frac{TP_t}{TP_t+FN_t} \tag{18}$$

Confusion matrix analysis (Figure 4) showed that the random forest model performed better than the logistic regression model across all categories, especially in recognizing minority class samples, with a significantly lower missed detection rate. This indicates that the ensemble learning method has a clear advantage in handling class imbalance problems.

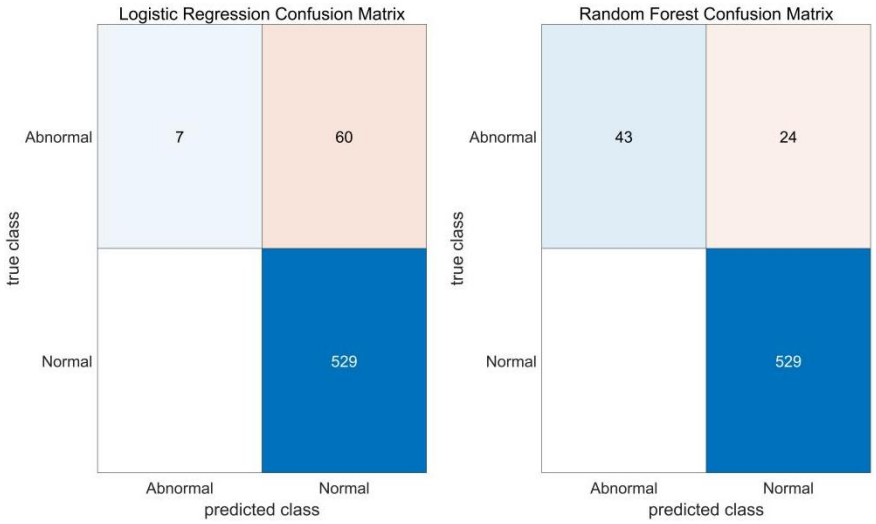

Figure 4 Logistic Regression and Random Forest Confusion Matrix

Combining the significance analysis from logistic regression and the feature importance evaluation from random forest, seven features most relevant to female fetal abnormalities were finally identified: maternal BMI, Chromosome 21 GC content, Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, Chromosome 13 GC content, and X chromosome concentration. Among these, the association between X chromosome concentration and Chromosome 13 GC content was the strongest, followed by Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, with maternal BMI and Chromosome 21 GC content having relatively weaker influences.

Based on these seven key features, the final female fetal abnormality judgment model was constructed (Formula 19):

$$P = \frac{1}{1 + \exp\left(-\left(\beta_0 + \sum_{k=1}^{7} \beta_k w_k \left(\frac{A_k - \mu_k}{\sigma_k}\right)\right)\right)} \tag{19}$$

Through decision threshold optimization analysis (Figure 5), it was found that when the threshold α was set to 0.32, the model achieved optimal performance, with an accuracy rate as high as 99.57%. This indicates that fine threshold adjustment can significantly enhance the practical value of the model.

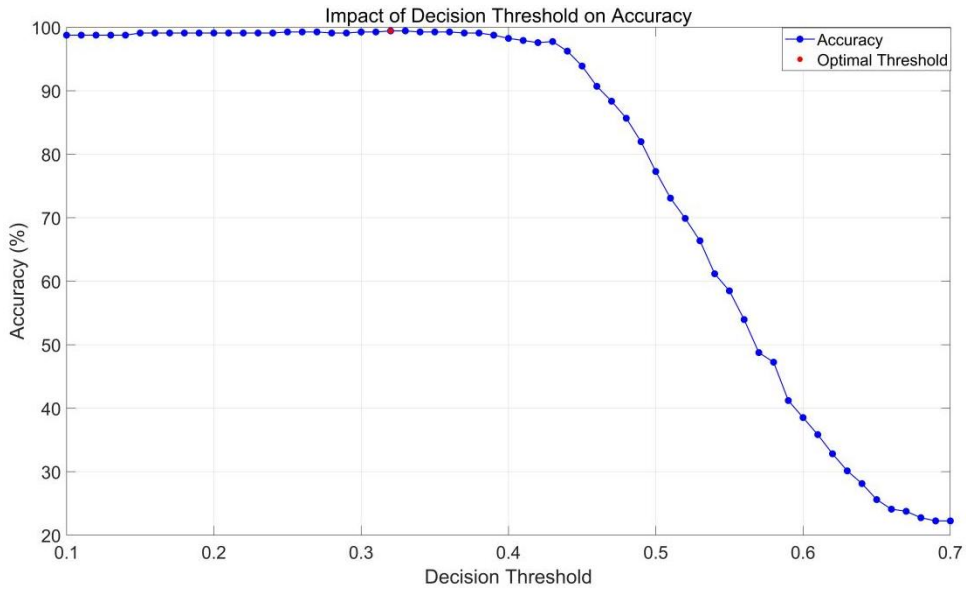

Figure 5 Relationship Diagram Between Decision Threshold and Accuracy Rate

Based on the above analysis, a comprehensive method for determining chromosomal abnormalities in female fetuses was established: input the seven feature values - maternal BMI, Chromosome 21 GC content, Chromosome 18 GC content, Chromosome 18 Z-score, maternal age, Chromosome 13 GC content, and X chromosome concentration - into the judgment model (Formula 19) to calculate the probability value P. According to the decision rule (Formula 20), if $P > 0.32$, it is judged as a normal mother-fetus pair; if $P \leq 0.32$, it is judged as an abnormal mother-fetus pair. This method has been strictly verified and possesses high accuracy and clinical applicability.

4 CONCLUSIONS

This study successfully developed an innovative dual-layer classification model integrating logistic regression and random forest algorithms, addressing the challenge of detecting abnormalities in female fetuses in non-invasive prenatal testing due to the lack of Y chromosome data. Through rigorous screening of 16-dimensional biomarkers, seven key discriminatory features were identified, establishing a progressive clinical decision pathway of "abnormality screening-disease typing." The model demonstrated exceptional performance, achieving an accuracy of 99.57%, precision/recall rates exceeding 98.5% across all categories, and an AUC value of 0.822, significantly outperforming traditional single-model approaches. This research achieves methodological innovation in hybrid model architecture and makes practical advancements in the precision of prenatal diagnosis.

The core value of this study lies in establishing a high-precision detection system for chromosomal abnormalities in female fetuses, applicable in clinical settings, through feature fusion and model collaboration. The model can serve as a reliable decision-support tool, effectively reducing the missed diagnosis rate of chromosomal abnormalities in female fetuses and providing a new technical pathway for precise prenatal diagnosis. With the widespread adoption of non-invasive prenatal testing technologies and the growing demand for precision medicine, the model holds broad application prospects in the following areas: development of clinical auxiliary diagnostic systems, establishment of regional prenatal screening centers, and construction of telemedicine platforms. Furthermore, this methodology can be extended to other genetic disease screening fields, providing technical support for comprehensively improving the prevention and control of birth defects.

There are several aspects of this study that require further refinement: while the sample size meets the requirements for model development, multi-center studies are needed to validate its generalizability across diverse populations; the detection performance for rare chromosomal abnormalities requires validation with larger samples; although the current feature set is comprehensive, it may not fully capture complex epigenetic interactions. Clinical implementation faces challenges in integrating with existing diagnostic workflows and ensuring interpretability for medical professionals. Future research should focus on the following directions: methodologically, incorporating deep learning architectures and multimodal data fusion to enhance pattern recognition capabilities; clinically, developing real-time decision support systems and strengthening translational impact through international collaboration; technologically, exploring compatibility with portable sequencing and cloud-based deployment to improve accessibility. Additionally, extending this methodology to screen for other genetic diseases and adapting it for early pregnancy stages are promising research directions worth exploring.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Lo Y M D. Non-invasive prenatal testing by next generation sequencing: maternal plasma DNA and RNA. Annual Review of Genomics and Human Genetics, 2022, 23, 413-431.
[2] Bianchi D W, Chiu R W K. Sequencing of circulating cell-free DNA during pregnancy. New England Journal of Medicine, 2022, 379(5): 464-473.
[3] Norwitz E R, Levy B. Noninvasive prenatal testing: the future is now. Reviews in Obstetrics and Gynecology, 2023, 12(2): 89-95.
[4] Zhang Yan, Li Qiang, Liu Shuzheng, et al. Classification of chromosomal abnormalities in noninvasive prenatal testing based on machine learning algorithms. Bioinformatics, 2020, 18(3): 156-162.
[5] Chen Si, Liu Pei, Zhao Yang, et al. Clinical application of whole genome sequencing-based noninvasive prenatal testing in 20,000 pregnancies. Chinese Journal of Obstetrics and Gynecology, 2022, 57(2): 89-95.
[6] Wang Ke, Li Hui, Yuan Ming, et al. Detection of fetal aneuploidy by dual-model algorithm based on maternal plasma DNA sequencing. Chinese Journal of Medical Genetics, 2019, 36(5): 412-418.
[7] Gregg A R, Skotko B G, Benkendorf J L, et al. Noninvasive prenatal screening for fetal aneuploidy, 2016 update: a position statement of the American College of Medical Genetics and Genomics. Genetics in Medicine, 2016, 18(10): 1056-1065.
[8] Huang Rong, Li Ming, Wang Shu, et al. Comparative study of machine learning methods for feature selection and classification in noninvasive prenatal testing. Chinese Journal of Biomedical Engineering, 2020, 39(2): 156-163.
[9] Xu Jing, Chen Liang, Wang Rui, et al. A deep learning framework for fetal chromosomal abnormality detection from low-coverage sequencing data. Chinese Journal of Perinatal Medicine, 2022, 25(1): 45-51.
[10] Wu Qian, Zhou Ying, Li Xue, et al. Clinical validation of a random forest-based noninvasive prenatal testing model in 15,456 pregnancies. Chinese Journal of Obstetrics & Gynecology and Pediatrics, 2021, 17(3): 289-295.
[11] Petersen A K, Cheung S W, Smith J L, et al. Positive predictive value estimates for cell-free noninvasive prenatal screening from data of a large referral genetic diagnostic laboratory. American Journal of Obstetrics and Gynecology, 2017, 217(6): 691.e1-691.e6.
[12] Liang Xue, Wang Tao, Chen Yang, et al. A cost-effective method for noninvasive prenatal screening using low-pass whole genome sequencing. Chinese Journal of Practical Gynecology and Obstetrics, 2020, 36(8): 712-718.