Journal of Pharmaceutical and Medical Research
Print ISSN: 2663-1954
Online ISSN: 2663-1962
DOI: https://doi.org/10.61784/jpmr3054

NON-INVASIVE PRENATAL DETECTION MODEL FOR FEMALE FETAL CHROMOSOMAL ANEUPLOIDY BASED ON XGBOOST

RuiYing Chen
School of Computer Science and Artificial Intelligence, Lanzhou University of Technology, Lanzhou 730050, Gansu, China.

Abstract: Addressing the challenge of limited accuracy in non-invasive prenatal testing (NIPT) for female fetal chromosomal aneuploidy due to the absence of Y chromosome reference, this study innovatively proposes a multi-feature fusion detection model based on XGBoost. The model's innovations are threefold: first, it constructs a three-dimensional feature system integrating "Z-score-GC content-clinical indicators", breaking through the limitation of traditional methods relying on single chromosomal indicators; second, it leverages XGBoost's powerful capability in capturing nonlinear relationships to deeply explore complex interaction effects among multi-chromosomal features; third, through feature importance ranking, it systematically reveals for the first time the critical roles of GC content in chromosome 13 and Z-scores of chromosomes 18 and X in female fetal abnormality detection. Experimental results demonstrate that the model achieves an accuracy of 75.45%, precision of 75.63%, recall of 75.45%, and F1-score of 75.47%, significantly outperforming traditional methods. This study provides a novel technical approach for detecting female fetal chromosomal aneuploidy with substantial clinical application value.

Keywords: Non-Invasive prenatal testing; Chromosomal aneuploidy; XGBoost; Feature fusion; Female fetal detection

1 INTRODUCTION

Chromosomal aneuploidy is a major genetic factor leading to spontaneous abortion and birth defects in fetuses, making its clinical detection crucial for improving population quality. Non-invasive prenatal testing (NIPT), which analyzes cell -free fetal DNA in maternal plasma, has become an important technical means for prenatal screening [1-2]. However, existing NIPT technologies face specific challenges in detecting chromosomal aneuploidy in female fetuses. Due to the lack of the key indicator of Y chromosome concentration, traditional methods often rely on single or limited feature parameters, failing to fully exploit the effective information within multidimensional data, which significantly restricts detection accuracy [3-4]. Zhang Yanchun et al., in their analysis of clinical application effectiveness, showed that although NIPT overall performance is superior, there is still room for improvement in female fetus-specific detection [5]. This technological bottleneck urgently needs to be addressed through innovative methods.

Currently, the application of machine learning in the NIPT field shows a trend of diversified development. Yuan Yuying first applied machine learning to the dual screening of "fetal aneuploidy + maternal tumors," demonstrating the applicability of intelligent algorithms in complex medical scenarios [6]. The aiD-NIPT algorithm developed by Junnam L's team significantly improved detection sensitivity by optimizing the processing strategy for low fetal fraction samples [7]. The KF-NIPT technology proposed by Kim D's team innovatively introduced K-mer analysis, further enhancing the ability to detect chromosomal abnormalities [8]. However, existing research still has significant shortcomings in female fetus-specific detection: decision tree models are prone to overfitting and sensitive to noisy data; although the AdaBoost algorithm can improve detection capability, it is sensitive to outliers and has poor model interpretability [9]. The breakthrough made by Xue Ying et al. in detecting chromosomal copy number variations [10], and the promotion of NIPT standardization by Belabbes' team [11], provide important references for this study, but detection models specifically for chromosomal aneuploidy in female fetuses are still relatively lacking.

This study innovatively proposes a multi-feature fusion detection model based on XGBoost. Its innovations are mainly reflected in three dimensions: First, it constructs a three-dimensional feature system of "Z-score - GC content - maternal BMI," breaking through the limitation of traditional methods relying on single indicators; Second, it fully utilizes the powerful nonlinear relationship capture capability of the XGBoost algorithm to deeply explore the complex interaction effects among multi-chromosomal features; Third, through systematic feature importance analysis, it clearly identifies for the first time the key roles of GC content in chromosome 13 and Z-scores of chromosomes 18 and X in female fetal abnormality detection. This study is based on the dataset of 11,501 clinical indication singleton pregnant women provided by Wang Yu et al. [12], and the feature variables used are all routine detection indicators, demonstrating good clinical translation prospects. The large-scale clinical studies by Zhou Ying et al. [13], Kong Lingrong et al. [14], and Yanchun Z's team [15] further validate the significant clinical application value of model construction based on Chinese population data. By deeply integrating advanced machine learning technology with clinical needs, this study provides a new technical pathway for improving the detection accuracy of chromosomal aneuploidy in female fetuses.

2 METHODOLOGY

2.1 Theoretical Foundation

The XG Boost algorithm is primarily chosen for constructing classification models due to its multiple significant advantages in medical data modeling scenarios. This algorithm efficiently handles nonlinear relationships within data, accurately capturing complex correlations among medical features—making it well-suited for fetal chromosomal aneuploidy detection data. Additionally, its built-in regularization mechanism effectively controls model complexity, mitigates overfitting risks, and ensures robust generalization capabilities on new data. XG Boost also demonstrates strong robustness to missing values in datasets, adapting to potential data incompleteness in real clinical scenarios without requiring complex imputation procedures. Moreover, this algorithm features rapid training speed and strong scalability, enabling it to handle datasets of varying sizes while maintaining efficiency. It typically demonstrates outstanding predictive performance across diverse classification tasks, sufficiently meeting the research demands for precise prediction of chromosomal aneuploidy types. The conceptual framework of XG Boost is illustrated in Figure 1:

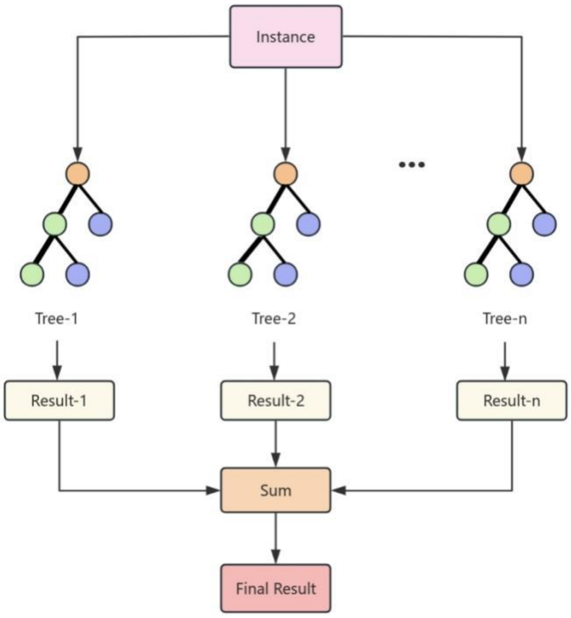

Figure 1 XGBoost Principle Framework Diagram

XGBoost, a decision tree-based ensemble learning algorithm, enhances predictive performance by constructing multiple weak classifiers (decision trees) and combining their results. Its core mechanism involves optimizing the objective function through gradient descent, where each iteration builds a new decision tree to correct errors from the previous model. Key advantages of XGBoost include its ability to effectively handle nonlinear relationships and high-dimensional data, its built-in regularization mechanism to prevent overfitting, strong robustness to missing values, and fast training speed with high scalability.

The objective function of XG Boost consists of a loss function and a regularization term, formulated as follows:

$$L(\phi)=\sum_{i=1}^{n} l\left(y_i,\hat{y}_i^{(t)}\right)+\sum_{k=1}^{t} \Omega(f_k) \tag{1}$$

Where: $\mathrm{l}\left(\mathrm{y}_\mathrm{i},\hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}\right)$ is the loss function for the *t*-th iteration, measuring the discrepancy between the predicted value $\hat{\mathrm{y}}_\mathrm{i}^{(\mathrm{t})}$ and the actual value y_i. The regularization term controls model complexity. *T* denotes the number of leaf nodes in the tree. ω_j represents the weight of the *j* leaf node. *γ* is the regularization parameter controlling the number of leaf nodes. *λ* is the *L2* regularization parameter controlling the weights of leaf nodes.

For multi-class classification problems, the cross-entropy loss function is employed, defined as follows:

$$l(y_i,\hat{y}_i)=-\sum_{c=1}^{C} y_{i,c} log(\hat{y}_i,c) \tag{2}$$

Where *C* denotes the total number of categories, and $y_{i,c}$ represents the indicator variable (0 or 1) indicating whether sample *i* belongs to category *c*.

2.2 Experimental Design

The primary data source for this study is authentic clinical data from singleton pregnancies[12]. The objective is to construct a high-precision detection model for fetal female chromosomal aneuploidy. The data primarily derive from chromosomal testing results of female fetuses, including Z-scores for chromosomes 13, 18, 21, and X; GC content; and maternal BMI. In this study, the XG Boost algorithm will be employed to construct a multi-classification prediction model incorporating multidimensional feature variables. The XG Boost model enables identification of the influence levels of different features on fetal chromosomal abnormality detection outcomes, thereby providing more accurate auxiliary diagnostic evidence for clinical non-invasive prenatal testing. The overall experimental design is illustrated in Figure 2:

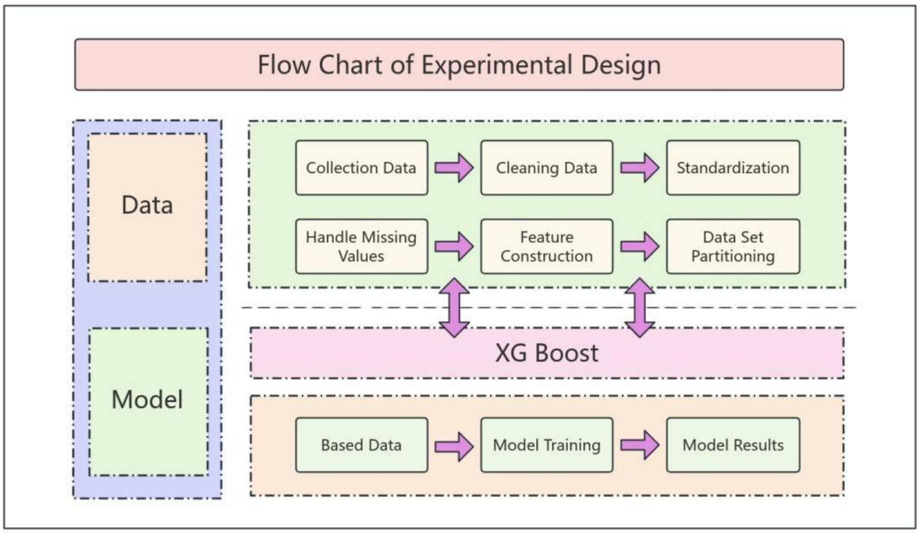

Figure 2 Experimental Design Flowchart

Regarding feature and target variable definitions, this study defines the feature variable set (X) as follows: it comprises 8 predictor variables (Z-scores of chromosomes 13, 18, 21, and X; GC content of chromosomes 13, 18, and 21; and maternal BMI). The target variable (y) is defined as the type of chromosomal aneuploidy, which is a multi-class categorical variable.

The optimization process of the XGBoost model is subject to the following constraints: First, the decision tree structure constraint: Each tree is a CART tree (Classification and Regression Tree), with each leaf node corresponding to an output value, as defined by the following formula:

$$f_k(x) = \omega_{q(x)}, \omega \in R^T, q: R^m \rightarrow \{1,2,\ldots,T\} \quad (3)$$

Where, $q(x)$ denotes the leaf node to which the sample x belongs, and T represents the number of leaf nodes. Next is the regularization constraint: it prevents overfitting by controlling the complexity of the tree.

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \quad (4)$$

Where, G_L and G_R denote the gradients of the left and right subtrees, respectively, while H_L and H_R denote the second derivatives of the left and right subtrees, respectively.Finally, there is the multi-class constraint: For multi-class problems, the model constructs a set of decision trees for each category. The final prediction is converted into a probability distribution via the softmax function, as shown in the following formula:

$$\hat{y}_{i,c} = \frac{\exp\left(f_c(x_i)\right)}{\sum_{k=1}^{C}\exp\left(f_k(x_i)\right)} \quad (5)$$

Using the trained model to predict the test set yields two key outputs: category predictions and probability predictions. This study defines quantitative evaluation metrics as follows: Accuracy is the proportion of correctly predicted samples relative to the total number of samples, serving as an indicator of the model's overall prediction correctness. The formula is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Precision is the proportion of samples predicted as positive that are actually positive, measuring the accuracy of a model's predictions. The formula is as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (7)$$

Recall is the proportion of actual positive samples that are correctly predicted. It measures a model's ability to identify positive cases and reduces missed positives. The formula is as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

F1 score is the harmonic mean of precision and recall, providing a comprehensive evaluation of model performance that balances precision and recall. The formula is expressed as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

In Formulas 6 through 9, TP denotes true positives (actual positive examples predicted as positive), TN denotes true negatives (actual negative examples predicted as negative), FP denotes false positives (actual negative examples predicted as positive), and FN denotes false negatives (actual positive examples predicted as negative). For multi-class classification problems, the composite score for each metric is calculated using a weighted average.

3 RESULTS

3.1 Feature Importance

Combining the aneuploidy results for chromosomes 21, 18, and 13 in female fetuses with comprehensive analysis of the X chromosome and related characteristics (Z-score, GC content, maternal BMI, etc.), an XG Boost model was employed to construct a classification method. This approach achieves predictive classification of female fetal abnormalities by learning the association patterns between features and chromosomal abnormalities. Feature importance is shown in Figure 3:

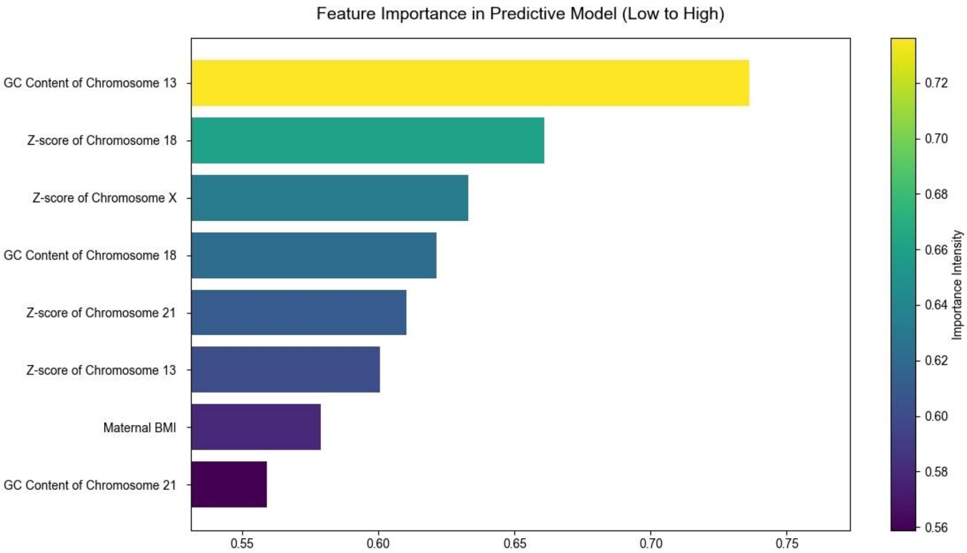

Figure 3 Feature Importance Plot

The feature importance analysis reveals that the GC content of chromosome 13 (0.7363) contributes most significantly to the model's predictions, followed by the Z-scores of chromosomes 18 and X. Although features related to chromosome 21 and maternal BMI show relatively lower importance, they still provide valuable supplementary information for the model. These findings confirm the critical role of integrating multi-chromosomal features in enhancing the detection performance of female fetal chromosomal aneuploidies.

3.2 Predictive Performance

This study employed multiple metrics to evaluate the model's predictive performance. The consistency observed between accuracy and recall indicates inherent stability in the model's classification logic for sample categories. This suggests the results are not entirely random and reveal exploitable patterns of feature associations. Although these values are relatively low, considering the absence of the Y chromosome as a key reference indicator for female fetus detection and the potential complexity of the samples, the findings still demonstrate the model's effectiveness in capturing feature correlations. The relevant metric data is presented in Table 1:

Table 1 Predictive Indicators

Indicator	Value
Model Accuracy	75.45%
Model Precision	75.63%
Model Recall	75.45%
Model F1-Score	75.47%

These results demonstrate that the model exhibits certain efficacy in addressing the task of determining chromosomal abnormalities in female fetuses. It can capture potential associations between features such as chromosome Z-scores and GC content with specific abnormality types, providing valuable reference for non-invasive prenatal testing. Concurrently, we will continue optimizing the model by increasing sample size and refining feature selection to further enhance its predictive accuracy and stability.

Based on the aneuploidy results for chromosomes 21, 18, and 13 in female fetuses, combined with characteristics such as X chromosome Z-score, GC content, read length and proportion, and maternal BMI, the XG Boost multi-classification model is employed for determination. Core features include Z-scores for chromosomes 21, 18, and 13, supplemented by GC content, read segment quality metrics, and BMI. The model outputs probability scores for each abnormality category. By combining threshold values with clinical expertise, it effectively identifies common trisomy syndromes, providing valuable reference for non-invasive prenatal testing.

4 CONCLUSIONS

This study successfully addresses the persistent challenge of limited detection accuracy in female fetal chromosomal aneuploidy screening, primarily caused by the absence of Y chromosome biomarkers, through the development of an innovative multi-feature fusion model based on XGBoost algorithm. By systematically integrating multidimensional features encompassing Z-scores and GC content from chromosomes 13, 18, 21, and X, combined with maternal clinical parameters including BMI, and harnessing XGBoost's exceptional capability in capturing complex nonlinear relationships, the model demonstrates remarkable proficiency in identifying various types of female fetal chromosomal abnormalities. The experimental validation reveals consistent performance across all key evaluation metrics, with the model achieving 75.45% accuracy, 75.63% precision, 75.45% recall, and 75.47% F1-score, thereby confirming its robust discriminative power when handling complex, real-world clinical data. Furthermore, the comprehensive feature importance analysis provides valuable biological insights, particularly highlighting the predominant role of chromosome 13 GC content (0.7363) and the significant contributions of Z-scores from chromosomes 18 (0.6609) and X (0.6330), which offer substantial evidence for clinical diagnostic applications.

The practical implementation of this model appears highly feasible within existing clinical frameworks, as it utilizes routinely available NIPT indicators without requiring additional specialized testing procedures. The model's probability distribution outputs provide clinicians with flexible, quantitative decision support tools, particularly valuable in challenging diagnostic scenarios involving female fetuses. Future enhancements could incorporate advanced molecular markers such as DNA fragment size distribution patterns and nucleosome positioning profiles, while the integration of time-series analytical approaches could enable dynamic monitoring of maternal biomarker variations. The modular architecture of the proposed system allows for potential expansion to encompass broader chromosomal abnormality detection, including microdeletion syndromes and rare aneuploidies, through incremental feature integration and model refinement. This research establishes a solid foundation for developing comprehensive fetal health assessment systems, potentially incorporating maternal epidemiological factors and environmental parameters to create holistic risk evaluation frameworks. The continuous optimization of this methodology, possibly through ensemble learning strategies combining multiple algorithmic approaches, promises significant improvements in detection sensitivity and specificity, ultimately contributing to enhanced prenatal care quality and outcomes.

COMPETING INTERESTS

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

REFERENCES

[1] Belabbes K, Benchekroun T, Bendala E, et al. Cell-Free Fetal DNA for Prenatal Screening of Aneuploidies and Autosomal Trisomies: A Systematic Review. International Journal of Pediatrics, 2024(1): 3037937.

[2] Junnam L, Mi S L, Mo J A, et al. Development and performance evaluation of an artificial intelligence algorithm using cell-free DNA fragment distance for non-invasive prenatal testing (aiD-NIPT). Frontiers in Genetics, 2022, 13: 999587.

[3] Kim D, Sohn J Y, Cho J H, et al. KF-NIPT: K-mer and fetal fraction-based estimation of chromosomal anomaly from NIPT data. BMC Bioinformatics, 2025, 26(1): 1-7.

[4] Kong Lingrong, Sun Luming. Application of non-invasive prenatal testing in screening for fetal chromosomal aneuploidies. Journal of Practical Obstetrics and Gynecology, 2023, 39(2): 98-102.

[5] Liu Bing, Zheng Nan, Liu Jing, et al. Risk prediction and interpretable analysis of left atrial thrombus or spontaneous echocardiographic contrast in patients with non-valvular atrial fibrillation based on XGBoost and SHAP. Chinese Journal of Cardiovascular Medicine, 2025: 1-10.

[6] Gil M S, Quezada M S, Bregant B, et al. Implementation of cell-free DNA-based non-invasive prenatal testing in a national health service: A cost-consequence analysis. Ultrasound in Obstetrics & Gynecology, 2023, 62(2): 205-214.

[7] 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 population. Prenatal Diagnosis, 2022, 42(1): 112-120.

[8] Pertile M D, Flowers N, Vavoulis S, et al. Sensitive and scalable non-invasive prenatal aneuploidy detection using cell-free DNA sequencing. Genetics in Medicine, 2024, 26(3): 101025.

[9] Van der Meij K R M, Sistermans E A, Macville M V E, et al. TRIDENT-2: National implementation of genome-wide non-invasive prenatal testing as a first-tier screening test in the Netherlands. American Journal of Human Genetics, 2022, 109(11): 2000-2008.

[10] Lefkowitz R B, Tynan J A, Liu Y, et al. Genome-wide noninvasive prenatal screening for carriers of balanced reciprocal translocations. Genetics in Medicine, 2023, 25(4): 100813.

[11] Dar P, Jacobsson B, MacPherson C, et al. Cell-free DNA screening for trisomies 21, 18, and 13 in pregnancies at low and high risk for aneuploidy. American Journal of Obstetrics and Gynecology, 2023, 229(1): 61.e1-61.e10.

[12] Martin K, Iyengar S, Kalyan A, et al. Clinical experience with a single-nucleotide polymorphism-based non-invasive prenatal test for five clinically significant microdeletions. Journal of Clinical Medicine, 2024, 13(2): 489.

[13] Zhou Ying, Wang Zhenyu, Mao Qianqian, et al. Application value of non-invasive prenatal testing technology in screening for fetal chromosomal aneuploidies. Chinese Journal of Medical Genetics, 2019, 36(11): 1094-1096.

[14] Gross S J, Stosic M, McDonald-McGinn D M, et al. Clinical experience with genome-wide noninvasive prenatal screening in a large cohort of pregnancies. The Journal of Maternal-Fetal & Neonatal Medicine, 2022, 35(25): 8485-8492.

[15] Taylor-Phillips S, Freeman K, Geppert J, et al. Accuracy of non-invasive prenatal testing using cell-free DNA for detection of Down, Edwards and Patau syndromes: a systematic review and meta-analysis. BMJ Open, 2024, 14(1): e073565.