A Machine Learning Approach for Fetal Chromosome Abnormality Identification Based on Multi-Feature Fusion

Yi Wang, Rouyi Fan, Yaning Yin, Tianyuan Liu*
School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China
**Corresponding Author*

Abstract: One of the prenatal screening techniques employed by NIPT (Non-Invasive Prenatal Testing) is high-throughput sequencing of foetal cell-free DNA isolated from maternal peripheral blood. Due to its high sensitivity, potential for early diagnosis, and non-invasive nature, it has emerged as a crucial method for identifying chromosomal abnormalities in fetuses. The Y chromosome's concentration is a vital reference point for quality assessment and anomaly analysis, necessitating its presence for the identification of male foetal abnormalities. However, because the Y chromosome lacks a definitive marker, the only techniques available for identifying chromosomal defects in female embryos are multidimensional feature fusion analysis and the X chromosome. For the diagnosis of female foetal anomalies, this methodology creates serious gaps and difficulties in feature utilisation, model stability, and interpretability. This study utilizes regional NIPT data to develop a multi-feature fusion and machine learning-based approach for identifying chromosomal abnormalities in female fetuses. SMOTE was employed to address the class imbalance brought on by the lack of aberrant samples in the training dataset. A feature set comprising Z-scores, GC content, and read duration metrics was methodically created. The LightGBM model was used to identify foetal chromosome abnormalities in females. Experimental results demonstrate that LightGBM outperformed Random Forest, XGBoost, CatBoost, and logistic regression algorithms, achieving 78.99% accuracy, 82.29% precision, 78.99% recall, and an F1 score of 80.52% on the test set. The most important diagnostic characteristics are the chromosome 21 Z-score, the percentage of duplicated reads, and the GC content, according to SHAP analysis, which was used to improve clinical interpretability. This study closes a gap in the present NIPT technology systems for detecting female foetal anomalies and offers an efficient method for accurate, interpretable screening of female foetal chromosomal abnormalities.

Keywords: Non-Invasive Prenatal Testing; Female Foetal Anomaly Detection; Feature Engineering; LightGBM; SHAP Explainability

1. Introduction

The main causes of perinatal mortality are death around birth, stillbirth, spontaneous abortion, birth malformations, and illnesses are foetal chromosomal abnormalities [1]. In unscreened pregnant populations, the inherent likelihood of foetal chromosomal abnormalities ranges from 0.36% to 6% [2]. Down syndrome, Edward's syndrome, and Patau syndrome are the most prevalent chromosomal disorders in fetuses. Abnormal amounts of free DNA fragments from foetal chromosomes 21, 18, and 13 are indicative of these disorders [3]. Finding anomalies in these chromosomes is essential for identifying problems in the fetus.

Based on sample gathering techniques, prenatal screening technologies for expectant mothers can be divided into invasive and non-invasive testing. Amniocentesis is the main invasive technique, which was initially used in the middle of the 1960s [3, 4]. This method, which offers excellent technical maturity and accuracy, entails putting a needle into the uterus to extract amniotic fluid for diagnosis. However, because it is an intrusive surgery, there is a slight but genuine chance of miscarriage, and the likelihood is directly related to the mother's unique situation and the doctor's expertise. This concern prompted the creation of non-invasive prenatal screening devices.

Mid-pregnancy serum screening, a non-invasive prenatal screening method, made its debut in the 1980s. This technique assesses the risk of foetal abnormalities by taking a mother's blood in the middle of her pregnancy. It doesn't harm the fetus and only needs a blood sample, unlike invasive procedures. Its drawbacks include a delayed testing schedule, high false-positive rates, and poor detection rates. The screening window was moved to the first trimester in the 1990s. While this early assessment allowed for earlier risk assessments, it still encountered issues with false positives and insufficient accuracy [5,6].

Non-invasive prenatal testing [5-7] became widely used in 2011. This method allows for early, non-invasive screening for foetal chromosomal aneuploidies by obtaining foetal cell-free DNA from maternal peripheral blood for high-throughput sequencing. This technology has emerged as a key technique for prenatal screening of trisomy 21, 18, 13, and sex chromosomal anomalies because of its high sensitivity, non-invasive nature, and detection rate. It has significantly improved the prevention of birth abnormalities globally by taking the place of invasive prenatal diagnostic techniques.

Theoretically, NIPT can now screen for a wide range of genetic variants, including trisomy, microdeletion syndromes, sex chromosomal abnormalities, and even monogenic illnesses. Its clinical use is still restricted, nevertheless. For instance, due to limited positive predictive values for rare autosomal and structural chromosomal abnormalities, it is typically recommended to avoid reporting such findings clinically. The main reason for this is that the foetal DNA examined in NIPT really comes from the placenta, hence the results frequently show chromosomal abnormalities in the placenta rather than the fetus [8].

Even with this technical framework and cautious application guidelines, NIPT still has several limitations when it comes to identifying chromosomal abnormalities in female fetuses. According to research on NIPT accuracy by Yunyun Z et al. [9,10], the existence of the Y chromosome is linked to the predictive value of NIPT for identifying foetal chromosomal abnormalities. The detection accuracy is significantly higher for fetuses with the Y chromosome than for those without. This suggests that Y chromosome concentration is an essential quality control and diagnostic reference for male fetuses. However, this marker is entirely absent in female fetuses, hence abnormality diagnosis requires integrated and indirect examination of multidimensional X chromosome and autosome properties. This discrepancy makes it difficult to identify abnormalities in female fetuses from a technical standpoint.

Current techniques for detecting female fetuses encounter a number of significant obstacles: First, feature utilisation is still low, and an over-reliance on Z-scores prevents multi-source data like maternal BMI, read length distribution, and GC content from being successfully integrated. Second, there is a greater chance of clinical missed diagnoses when there is a significant class imbalance, as uncommon anomalous data result in low model recognition rates. Third, clinical trust and adoption are hampered by "black-box" decision procedures and inadequate model interpretability. Fourth, individual variability is disregarded, as the generalisability of the model is hampered by physiological variations among maternal populations.

Machine learning algorithms, with strong feature learning and classification abilities, have recently shown tremendous promise in complex biological data mining. While previous research attempted to apply these strategies to NIPT to improve detection, most studies did not optimise algorithms or validate systems specifically for the unique female foetal population [11,12]. To address this gap, our study introduces a multi-feature fusion and machine learning approach for identifying chromosomal anomalies in female fetuses. We systematically created a high-dimensional feature set—comprising read length, GC content, Z-scores, and other indicators—and examined several popular ensemble learning techniques to develop an accurate, automated screening model tailored for female fetuses. Furthermore, our study employs the SHAP explainability methodology to tackle the inherent complexity of machine learning models.

2. Data Preprocessing

2.1 Data Collection

An NIPT testing dataset supplied by a regional institution served as the source of the data for this investigation. 605 female foetal samples were included in the 1,687 records that made up the original dataset. Table 1 describes the precise

data content and structure of the 31 feature variables that were present in each record.

Table 1. Description of Dataset Variables

Category	Variable Name	Description
Sample Label	Sample ID, Maternal ID	Unique identifiers for samples and pregnant women
Pregnant Woman's Basic Information	Age, Height, Weight, BMI	Physiological and physical index of pregnant women
Clinical Testing Information	Last Menstrual Period Date, Test Date, Gestational Age, Number of Blood Draws	Time-related and procedural records pertaining to pregnancy cycle and sampling
Sequencing Quality Metrics	Total reads, alignment rate, duplicate read ratio, GC content	Key parameters evaluating overall sequencing data quality
Chromosome-Specific Metrics	Quantitative and quality control measurements for target chromosomes	Quantitative and quality control measurements for target chromosomes
Clinical Diagnosis Results	Chromosomal aneuploidy determination, foetal health status	Clinical diagnosis and foetal health status based on test results

2.2 Data Processing

2.2.1 Missing value treatment

To guarantee data completeness and logical coherence, a statistical analysis of missing values for important variables was carried out, and focused imputation techniques were created. Table 2 provides specific handling options.

Table 2. Missing Value Handling Strategy

Variable Name	Missing Proportion	Handling Method	Processing Basis
Last Menstrual Period	0% (female fetus)	No action required	The female data column is complete
Gravida BMI	0.17%	Median imputation	BMI is a continuous variable; the median is less sensitive to outliers and better represents central tendency.
Chromosomal aneuploidy (AB column)	88.93%	Uniformly labeled as "Normal"	Per data notes, blank entries indicate no abnormal detection results; thus, missing values are considered clinically normal.

2.2.2 Sample screening

According to clinical consensus, the effective testing window for NIPT is between 10 and 25 weeks of gestation. This study strictly adhered to this range for sample selection to exclude potential testing noise introduced by overly early or late gestational ages.

2.2.3 Outliers and threshold adjustment

For the GC content metric, the proposed normal range is 40%-60%. However, preliminary analysis indicates that a large number of samples exhibit GC content distributed within the 38%-39% range. Strictly capping the threshold at 40% would exclude nearly one-third of valid samples, introducing significant selection bias. Referencing relevant literature and accounting for biological and technical variability in actual testing, this study adjusted the lower bound for valid GC content to 37%. This maximizes sample retention while ensuring data reliability. For other continuous variables such as maternal age and body mass index (BMI), this study applied screening based on medically recognized reasonable ranges: BMI was restricted to 20–40, and age was controlled within 20–45 years. Values within these established reasonable ranges were retained to reflect genuine individual variation.

2.2.4 Data standardization and encoding

To achieve uniform data formatting and meet the input requirements of machine learning models, the following conversions were performed.

For gestational age data conversion, information in the "weeks + days" format (e.g., 12w+3) was uniformly converted to a floating-point number representing weeks. The conversion formula is:

$$Gestational\ age = \mathrm{n+m/7} \quad (1)$$

Where n represents the number of full weeks and m represents the number of days. This facilitates model reception and processing.

For continuous numerical features such as GC content and read length, Z-score normalization is applied to set their mean to 0 and standard deviation to 1. This aims to eliminate dimensional differences between features and prevent model training bias caused by uneven numerical ranges. The formula is:

$$z = (x - \mu) / \sigma \quad (2)$$

This processing eliminates dimensional differences between features, preventing certain features with large numerical ranges from dominating the model training process. For the categorical variable "Number of Pregnancies," one-hot encoding is applied. Since no inherent ordinal relationship exists among its categories (1 time, 2 times, ≥3 times), one-hot encoding converts it into a binary feature, making it more suitable for machine learning models.

2.2.5 Extract key features

After completing preliminary feature engineering, a strategy combining model-based feature importance ranking with correlation analysis was employed for feature selection. To identify core indicators most strongly associated with female foetal abnormalities, this study first assessed feature importance using the random forest algorithm. Subsequently, to eliminate the impact of multicollinearity, a correlation

coefficient threshold was set to remove highly correlated features. The specific screening process is as follows: Iterate through all features in descending order of importance. For feature i, if feature j exists in feature set S such that $|R_{ij}|>0.8$, then exclude feature i; otherwise, add feature i to set S. To determine the optimal number of features, performance curves were constructed for subsets containing 1 to 10 features. Using 5-fold cross-validation, the average F1 score of the LightGBM model is evaluated across different feature subsets. It is observed that model performance stabilizes and reaches a plateau when the number of features reaches 6. Ultimately, while retaining low-correlation features, the top 6 most important features are selected as core predictive indicators. Parameter fitting and analysis are then performed using a Logit model. Let the probability of female foetal abnormality be P(y=1|X). Using the Logit model, the following 6 core features were incorporated:

$$\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)=\beta_0+\beta_1 S'+\beta_2 R' +\beta_3 Q'+\beta_4 Q_{seq}+\beta_5 BMI'+\beta_6 T' \quad (3)$$

Where β epresents the feature weight (determined via maximum likelihood estimation; positive values indicate that the feature increases the probability of an anomaly, while negative values decrease it); Results from the log-likelihood model fitted to the training data show the estimated weights for each feature parameter as follows: $\beta_0 = -2.5$, $\beta_1 = 3.2$, $\beta_2 = 2.7$, $\beta_3 = 2.3$, $\beta_4 = -1.8$, $\beta_5 = 0.9$, $\beta_6 = 0.6$. Analysis indicates that the Z-score of chromosome 21 exhibits the most significant positive correlation with anomaly probability (β_1= 3.2), followed by the Z-scores of chromosomes 18 and 13. The test quality score feature (β_4= -1.8) shows a significant negative correlation, indicating that higher sequencing quality predicts a lower risk of foetal abnormalities. Additionally, both maternal BMI and X chromosome Z-score showed positive correlations, though their influence was relatively modest.

2.3 Category Distribution Balancing

The classification target of this study is foetal chromosomal abnormality types, including: 0 (normal), 1 (Trisomy 13), 2 (Trisomy 18), and 3 (Trisomy 21). The original data exhibits severe class imbalance, with the vast majority of samples belonging to the normal category (89.5%), while the three abnormal categories collectively account for only 10.5% of the samples. Training a model directly on this dataset would result in severe overfitting to the majority class. To address this issue, this study employs the Synthetic Minority Over-sampling Technique (SMOTE) on the training set during the model training phase. The fundamental principle of SMOTE is to synthesize new samples for the minority class by identifying the k-nearest neighbors of minority samples in the feature space and generating new samples through linear interpolation. Specifically, for each sample in the minority class, one sample is randomly selected from its k nearest neighbors, and a new sample is generated. The mathematical representation is as follows:

$$\begin{aligned} x_{new} &= x_i + \lambda(x_j - x_i)x_{new} \\ &= x_i + \lambda(x_j - x_i) \end{aligned} \quad (4)$$

Among them are minority class samples, each randomly selected as a neighbor with λ being a random number between [0,1]. Through SMOTE, we reconstruct the class distribution of the training set to achieve balance, thereby significantly enhancing the model's ability to identify anomalous samples.

3. Methodology

3.1 Model Selection and Training Framework

We developed a research framework for fetal chromosomal abnormality detection by comprehensively comparing multiple machine learning models and incorporating SHAP for interpretability analysis, as detailed in Figure 1.

Experiments employed stratified sampling to divide the dataset into training and test sets at an 8:2 ratio, maintaining consistent class distribution. All models underwent hyperparameter tuning via 5-fold cross-validation, with F1 score as the primary evaluation metric to ensure comparability of model performance under imbalanced data scenarios.

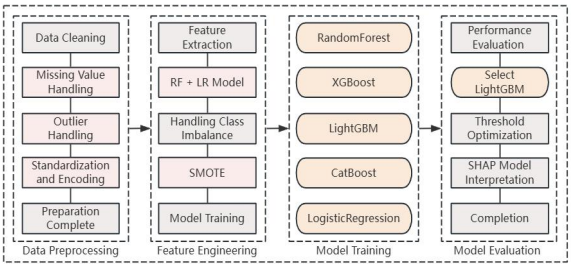

Figure 1. Research Framework

3.2 LightGBM

After completing category imbalance handling, rigorous cross-validation, and comprehensive comparison across multiple evaluation metrics revealed that the LightGBM model demonstrated the best overall performance on the test set. Its accuracy, F1 score, and AUC value were significantly superior to those of other comparison models. The model's principles are illustrated in Figure 2:

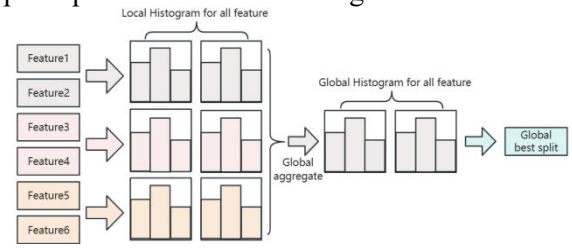

Figure 2. LightGBM Model Schematic Diagram

LightGBM demonstrated outstanding performance in this study, with its balance of high efficiency and high accuracy stemming from several core design principles: First, the histogram-based decision tree algorithm discretizes continuous features into finite bins, significantly reducing computational complexity and memory consumption; Second, the Leaf-wise growth strategy selects nodes with the greatest gain for splitting, achieving lower loss function values for the same number of splits. This enhances model accuracy while maintaining efficiency. Additionally, the model natively supports categorical feature handling without requiring pre-coding and integrates L1 and L2 regularization terms to constrain model complexity, effectively suppressing overfitting.

For the multi-classification task addressed in this study, LightGBM employs a one-vs-rest strategy. It trains independent binary classifiers for each category and ultimately determines the prediction result by comparing the output probabilities across categories.

3.3 Threshold Optimization

The model outputs predicted probabilities for each sample's chromosome status (normal, T13, T18, T21). To align with clinical decision-making needs, this study first transformed the problem framework into an anomaly risk identification task, defining the sample's anomaly probability as:

$$P(abnormal) = P(\mathrm{T13})+P(\mathrm{T18})+P(\mathrm{T21}) = 1\text{-}P(normal) \quad (5)$$

Given that the risk of false negatives far exceeds that of false positives in clinical settings, directly adopting the default probability threshold of 0.5 would result in insufficient sensitivity for identifying abnormal samples. Therefore, this study employs grid search to systematically optimize the decision threshold, targeting maximization of the F2 score. The F2 score is a weighted variant of the F1 score. By assigning twice the weight to recall compared to precision, it better aligns with the core requirement for high sensitivity in clinical practice. The specific calculation formula is as follows:

$$F2 = (5 \times P \times R) / (4 \times P + R) \quad (6)$$

Here, P denotes, R denotes, and the optimization process is conducted on the validation set. The principle is as follows: a series of candidate thresholds is generated with fine increments within the probability range [0, 1], and the corresponding F2 score is evaluated for each candidate threshold when used as the classification criterion. Ultimately, the threshold that maximizes the F2 score is selected as the optimal decision boundary.

3.4 Explainability Analysis

This study employs the SHAP (Shapley Additive exPlanations) method for interpretability analysis to gain deeper insight into the decision-making mechanism of LightGBM models and enhance their predictive transparency. The SHAP framework, based on Shapley value theory from cooperative game theory, fairly quantifies each feature's contribution to the model's prediction outcomes.

The SHAP method decomposes the predicted value of a single sample into the base value plus the sum of contributions from each feature. Its additive explanatory model can be expressed as:

$$f(x) = \phi_0 + \sum \phi_i (i = 1 \text{ to M}) \quad (7)$$

Here, f(x) denotes the model's predicted output for the sample, ϕ_0 is the base value (the average prediction across all training samples), and ϕ_i represents the SHAP value of the i-th feature. A positive value indicates that the feature increases the prediction probability, while a negative value suggests it decreases it.

Based on this framework, this study systematically analyzes model behavior through three dimensions. First, by integrating SHAP values across all samples, we calculate the average absolute value of each feature's contribution to identify globally significant

features influencing model decisions. Second, using SHAP dependency graphs and feature value distribution scatter plots, we reveal directional relationships between specific features and prediction outcomes, clarifying whether their value changes have positive or negative impacts on predictions. Finally, for high-risk samples, we employ SHAP waterfall diagrams to trace decision-making paths, visually demonstrating key evidence features and their contribution directions and magnitudes, thereby providing clinicians with case-level decision-making references.

Finally, for specific high-risk samples, SHAP force-waterfall plots trace the decision path leading to their predictions, visually displaying key evidence features along with their contribution direction and magnitude, thereby providing case-level decision support for clinicians.

This analysis method not only verifies the consistency between the model decision logic and clinical prior knowledge, but also constructs a complete interpretation system from global feature importance to case decision interpretation, which significantly enhances the credibility and acceptability of the model in clinical application.

4. Results and Analysis

4.1 Feature Selection and Parameter Tuning

Through a feature selection framework combining random forest and logistic regression (with feature importance rankings shown in Figure 3), six core features with high significance and strong independence were identified for subsequent modeling. These features include: chromosome 21 Z-value, chromosome 18 Z-value, chromosome 13 Z-value, repeat sequence ratio, GC content, and maternal BMI.

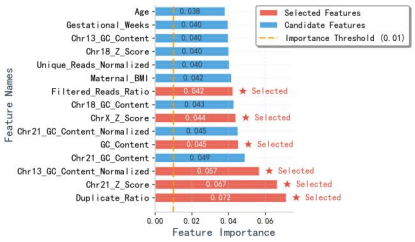

Figure 3. Feature Importance Screening

This study employed SMOTE oversampling technique to equalize the number of three types of chromosomal abnormality samples with normal categories, thereby constructing a fully balanced dataset (specific category distribution comparison before and after sampling is shown in Table 3) to ensure the model can learn features of all categories equally.

Table 3. Comparison of Class Distributions before and After SMOTE Sampling

Category	Pre-sampling Sample Size	Pre-sampling Proportion (%)	Post-sampling Sample Size	Post-sampling Proportion (%)	Growth Multiple
0	530	89.5%	530	25.0%	1.0x
1	25	4.2%	530	25.0%	21.2x
2	30	5.1%	530	25.0%	17.7x
3	7	1.2%	530	25.0%	75.7x

To optimize model performance and control complexity, this study systematically tunes model hyperparameters using a grid search method. Cross-validation was employed to evaluate the generalization capabilities of different parameter combinations, ultimately determining the optimal parameter configuration of the model as shown in Table 4.

4.2 Model Performance Comparison

Figure 4 presents a comparison of five models' performance on independent test sets and their overall trends, with detailed metrics listed in Table 5. LightGBM outperforms all models in accuracy, precision, recall, and F1 score, achieving 80.52% F1 score and 78.99%.

It proved its superiority in processing such complex and high-dimensional medical data. The worst performance of the logistic regression model indicates that there are complex nonlinear relationships in the data, which are difficult to capture effectively by the linear model.

Table 4. Optimal Hyperparameters

Model Name	Number of Decision Trees	Maximum Depth	Learning Rate	Random Seed
Random Forest	100	10	-	42
XGBoost	100	6	0.1	42
LightGBM	100	8	0.1	42
CatBoost	100	6	0.05	42
Logistic Regression	-	-	-	42

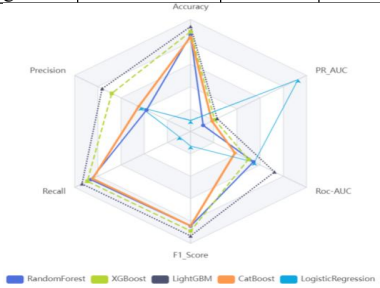

Figure 4. Machine Learning Model Performance Comparison

Table 5. Performance Comparison of Different Machine Learning Models

model	Accuracy	Precision	Recall	F1_Score	Roc-AUC	PR_AUC
RandomForest	0.7731	0.7880	0.7731	78.05%	0.4543	0.0984
XGBoost	0.7899	0.8180	0.7899	80.01%	0.4499	0.1153
LightGBM	0.8151	0.8263	0.8151	81.99%	0.4724	0.1190
CatBoost	0.7647	0.7947	0.7647	77.94%	0.4383	0.1113
LogisticRegression	0.3529	0.7929	0.3529	45.89%	0.4550	0.2377

4.3 Feature Importance and SHAP Analysis

Based on the SHAP (Shapley Additive exPlanations) framework, this study conducted an in-depth analysis of the predictive mechanism of the LightGBM model, revealing the decision-making basis for chromosome abnormality classification from two dimensions: feature importance and contribution.

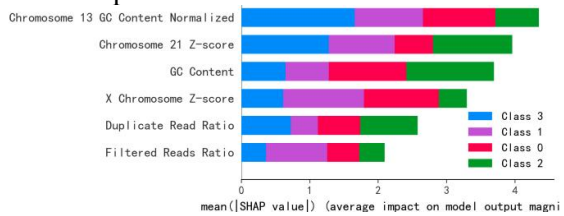

Figure 5. LightGBM Feature Importance Ranking

Figure 5 illustrates the global feature importance ranking based on SHAP values. The feature importance analysis indicates that the model successfully identified key biomarkers closely associated with chromosomal abnormalities. Among these, the normalized GC content of chromosome 13, the Z-score of chromosome 21, and the genome-wide GC content ranked as the top three most important features. This ranking aligns closely with the clinical diagnostic focus for Down syndrome (T21) and Patau syndrome (T13). Data quality metrics such as the Z-score of the X chromosome and the proportion of duplicate reads also demonstrated significant contributions, highlighting the critical impact of sequencing quality on diagnostic accuracy.

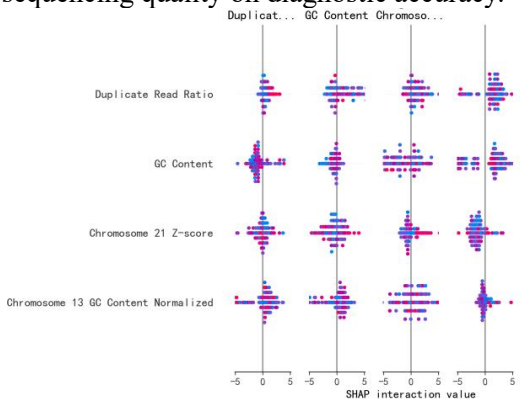

Figure 6. Feature Importance Summary Based on SHAP Values

The SHAP interaction analysis in Figure 6 further reveals the complex relationship between features and prediction outcomes. Chromosome 13's GC content_normalized exhibits a distinct bidirectional influence pattern, where its numerical changes directly correlate with increases or decreases in abnormal risk. The Z-score on chromosome 21 exhibits a typical threshold effect: its contribution to anomaly prediction sharply increases when exceeding the statistically significant threshold, aligning with clinical practice standards based on Z-score interpretation. The overall importance of the GC content feature group reflects the complementary value of global and local GC content metrics, collectively forming a comprehensive quality control system.

From a clinical interpretability perspective, this analysis validates the biological plausibility of model decisions. The prominent importance of chromosome-specific indicators aligns with the pathological mechanisms of aneuploidy detection, while the threshold response characteristics of Z-score indicators resonate with established statistical judgment criteria. This SHAP-based white-box analysis not only enhances the model's credibility in medical applications but also provides valuable feature importance references for prenatal diagnosis, advancing the transparent application of machine learning in clinical decision-making.

4.4 Threshold Optimization Effect

The performance comparison before and after threshold optimization is shown in Table 6. After optimization, the model's recall rate increased from 72.5% to 85.2%, indicating that the model can detect more true abnormal cases and significantly reduces the risk of clinical missed detections. Although the precision rate decreased, the F2 score improved significantly, better meeting the risk control requirements of this application scenario.

Table 6. Performance Comparison before and After Threshold Optimization

Scenario	Threshold	Precision	Recall	F1 Score	F2 Score
Before Optimization	0.5	85.1%	72.5%	78.3%	75.2%
After Optimization	0.796	80.3%	85.2%	82.7%	84.3%

5. Discussion and Conclusions

5.1 Discussion

This study successfully developed an efficient and interpretable machine learning model for detecting chromosomal abnormalities in female fetuses. The model demonstrates four key advantages: First, it establishes a systematic workflow from data preprocessing to model interpretation, ensuring reproducibility and reliability. Second, through multi-model comparison and fine-tuning, LightGBM outperforms other models in processing high-dimensional NIPT data. Third, threshold optimization prioritizes abnormal sample detection (high recall rate), aligning with prenatal screening's clinical principle of prioritizing sensitivity for high-risk cases. Finally, SHAP analysis enhances decision transparency, transforming the "black box" model into a clinically understandable "gray box" that helps clinicians interpret diagnostic criteria and facilitates human-machine collaborative decision-making.

However, this study has several limitations: First, the training data originates from a single region with predominantly high maternal BMI, limiting sample representativeness. The model's generalizability requires further validation on more diverse datasets. Second, current features are extracted solely from routine NIPT test reports. Integrating deeper sequencing data (e.g., coverage depth in specific genomic regions) may further enhance model performance in future studies.

5.2 Conclusions

To address the unique challenges in identifying female foetal chromosomal abnormalities using NIPT technology, this study developed a machine learning-based composite decision-making process. This approach first identifies key features through medical a priori knowledge and feature importance evaluation, then employs a two-stage classification strategy to determine chromosomal abnormalities. In the first stage, a LightGBM binary classification model is established. The total probability of a sample being predicted as any of the T13, T18, or T21 anomaly types serves as the basis for determining the presence of chromosomal abnormalities. The second stage further employs a “one-versus-many” strategy for samples initially screened as high-risk. Independent binary classifiers are trained, and the precise identification of the abnormal type is achieved by comparing the predicted probabilities of each category, where the final type is determined as argmax (P (T13), P (T18), P (T21)).

Through decision curve optimization, the optimal probability threshold for distinguishing abnormal from normal samples was determined as 0.796. Based on this threshold, the judgment rule is established: if P (abnormal)> 0.796, the sample is classified as high-risk and requires confirmation; otherwise, it is deemed normal. This approach balances screening sensitivity with diagnostic specificity, providing a reliable and interpretable solution for identifying chromosomal abnormalities in female fetuses.

References

[1] Oyovwi S O M, Ohwin P E, Rotu A R, et al. Internet-Based Abnormal Chromosomal Diagnosis during Pregnancy Using a Noninvasive Innovative Approach to Detecting Chromosomal Abnormalities in the Fetus: Scoping Review. JMIR bioinformatics and biotechnology, 2024, 5e58439.

[2] Zaiaeva, E. E.; Andreeva, E. N.; Demikova, N. S. Prevalence of Rare Chromosomal Anomalies Based on Epidemiological Monitoring of Congenital Malformations in the Moscow Oblast. Medical Genetics. 2021, 20 (7), 59–66.

[3] Frisova V. Prenatal Screening for Chromosomal Defects. Reproductive Medicine, 2025, 6(2):15-15.

[4] Jindal A, Sharma M, Karena ZV, Chaudhary C. Amniocentesis. 2023 Aug 14. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan–. PMID: 32644673.

[5] Zheng L, Y in N, Wang M, et al. Comparative performance and health economic analysis of prenatal screening for down syndrome in Fujian province, China. Scientificreports, 2025, 15(1): 23940. DOI: 10.1038/S41598-025-08592-0.

[6] Bulbul A G, Kirtis E, Kandemir H, et al. Is intermediaterisk really intermediate? Comparison of karyotype and non-invasive prenatal testing results of pregnancies at intermediate risk of trisomy 21 on maternal serum screening. Journal of genetic counseling, 2024, 34(2):e1973-e1973.

[7] George K, Achilleas A, Michalis N, et al. Targeted capture enrichment followed by NGS: development and validation of a single comprehensive NIPT for chromosomal aneuploidies, microdeletion syndromes and monogenic diseases. Molecular cytogenetics, 2019, 12(1): 48.

[8] Meij D V R K, Sistermans A E, Macville V M, et al. TRIDENT-2: National Implementation of Genome-wide Non-invasive Prenatal Testing as a First-Tier Screening Test in the Netherlands. The American Journal of Human Genetics, 2019, 105(6):1091-1101.

[9] Yunyun Z, Shanning W, Yinghui D, et al. Clinical experience regarding the accuracy of NIPT in the detection of sex chromosome abnormality. The journal of gene medicine, 2020, 22(8):e3199.

[10] Luo W, He B, Han D, et al. The clinical performance of foetal sex chromosome abnormalities in serum biochemical screening in the second trimester. Scientific Reports, 2024, 14(1):29011-29011.

[11] Huang Q, Zhu J, Lu J, et al. A noninvasive prenatal test pipeline with a well-generalized machine-learning approach for accurate foetal trisomy detection using low-depth short sequence data. Expert Systems with Applications, 2024, 249(PC):123759-.

[12] Xianfeng X, Liping W, Xiaohong C, et al. Machine learning-based evaluation of application value of the USM combined with NIPT in the diagnosis of foetal chromosomal abnormalities. Mathematical biosciences and engineering: MBE, 2022, 19(4): 4260-4276.

