World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3075

MULTI-FEATURE INTEGRATED INTELLIGENT DIAGNOSIS MODEL FOR NON-INVASIVE PRENATAL TESTING BASED ON ENSEMBLE LEARNING

KangYong Wang
College of Electronic Engineering, National University of Defense Technology, Changsha 410073, Hunan, China.

Abstract: To address the challenges of high false-negative rates and low diagnostic reliability in current Non-Invasive Prenatal Testing (NIPT) for detecting chromosomal abnormalities in female fetuses, this study proposes an innovative multi-feature intelligent diagnosis model based on ensemble learning. The key innovations of this research include: (1) Multi-dimensional feature integration: For the first time, 16 critical features covering chromosome Z-values (13, 18, 21, X), GC content, sequencing metrics, and maternal physiological indicators were systematically integrated to comprehensively characterize fetal chromosomal status. (2) Advanced ensemble framework: We developed a novel hybrid ensemble approach combining Random Forest, XGBoost, and LightGBM algorithms through soft voting, effectively addressing data challenges of high dimensionality, small sample size, and severe class imbalance (only 10.7% abnormal samples). (3) Dual optimization strategy: The model was optimized using both SMOTE oversampling and random undersampling techniques for data balance, combined with grid search and five-fold cross-validation for parameter tuning. Experimental results demonstrate that our ensemble model achieved superior performance with 91.59% accuracy, 0.9583 AUC, 91.49% precision, 89.58% recall, and 90.53% F1-score, significantly outperforming single-algorithm models. Feature importance analysis revealed that BMI, chromosome 18 Z-values, and maternal age were the most influential predictors. This model provides a clinically applicable, highly accurate diagnostic tool that substantially improves the reliability of NIPT-based female fetal abnormality detection.

Keywords: Non-Invasive Prenatal Testing(NIPT); Chromosomal abnormalities; Ensemble learning; Multi-feature integration; Intelligent diagnosis

1 INTRODUCTION

Non-Invasive Prenatal Testing (NIPT), as a prenatal screening technology based on cell-free fetal DNA in maternal peripheral blood, has become a key method for detecting fetal chromosomal aneuploidies such as trisomy 21, 18, and 13 [1]. Compared with traditional serum screening and invasive diagnostic techniques, NIPT is widely adopted due to its high sensitivity, specificity, and safety [2]. However, although NIPT has matured in detecting common chromosomal aneuploidies, its accuracy in identifying chromosomal abnormalities in female fetfaces remains challenging, particularly in cases involving sex chromosomes and other microstructural abnormalities, as it is susceptible to interference from factors such as fetal DNA concentration, sequencing depth, and maternal background [3,4]. Current clinical practices often rely on threshold-based judgments of single biomarkers (e.g., chromosome Z-scores), lacking comprehensive analysis of multidimensional data features, which can lead to missed or misdiagnosed complex abnormal cases [5]. Therefore, developing an AI-assisted diagnostic model that can comprehensively utilize multidimensional NIPT data to improve the accuracy of female fetal abnormality detection holds significant clinical and scientific value.

In recent years, machine learning methods have demonstrated strong potential in fields such as medical image analysis and genomic data processing, offering new approaches for in-depth mining of NIPT data [6,7]. For example, Random Forest has been applied to predict the risk of fetal chromosomal abnormalities, given its ability to handle high-dimensional features and assess variable importance [8]. Gradient boosting algorithms such as XGBoost and LightGBM have garnered attention in bioinformatics classification tasks due to their high efficiency and strong predictive performance [9,10]. However, existing research predominantly focuses on the application of single algorithms to NIPT data, with limited exploration of how to integrate the strengths of multiple algorithms to address challenges such as data imbalance and feature redundancy [11]. Furthermore, most models utilize only limited features, such as chromosome Z-scores, and fail to systematically integrate sequencing quality indicators (e.g., GC content, alignment ratio) and maternal physiological parameters (e.g., BMI, age), which constrains the interpretability and generalizability of the models [12].

To address the aforementioned research gaps, this paper proposes an intelligent detection model for chromosomal abnormalities in female fetuses based on multi-feature fusion and ensemble learning. The main marginal contributions of this study are as follows: First, it systematically integrates 16 key features, including chromosome Z-scores, GC content, sequencing quality control indicators, and maternal physiological characteristics, constructing a multidimensional feature system that comprehensively reflects fetal chromosomal status and detection quality. Second, it innovatively develops a soft voting ensemble learning framework that combines Random Forest, XGBoost, and LightGBM, leveraging their complementary strengths to effectively enhance the model's capability to handle high-dimensional, small-sample, and imbalanced data, thereby improving classification performance. Third, hyperparameter optimization is performed through grid search and cross-validation, and strategies such as SMOTE and random undersampling are employed to address class imbalance, ensuring the model's robustness and generalizability. Experimental results demonstrate that the proposed ensemble model achieves significantly better performance on the test set compared to single algorithms, providing a more reliable and precise intelligent decision-support tool for the clinical application of NIPT in detecting chromosomal abnormalities in female fetuses.

2 INTELLIGENT DIAGNOSIS MODEL

2.1 Data Preprocessing

To construct a reliable model for detecting chromosomal abnormalities in female fetuses, this study first conducted systematic data cleaning and preprocessing. The original dataset comprised 24 feature dimensions with a total of 534 samples. All valid samples were retained through appropriate missing value handling to ensure data integrity. Based on the principles of NIPT detection, we analyzed the aneuploidy status of chromosomes 13, 18, and 21, adopting a comprehensive classification criterion: any chromosomal abnormality was labeled as an abnormal sample. Ultimately, 477 normal samples (89.3%) and 57 abnormal samples (10.7%) were obtained, reflecting a typical class-imbalanced distribution.

Regarding feature selection, we systematically selected 16 key feature variables by integrating the technical principles and clinical significance of NIPT detection. These features span multiple dimensions: 1) Chromosomal Z-score features, including Z-scores for chromosomes 13, 18, 21, and X, which directly reflect chromosomal copy number abnormalities; 2) GC content features, comprising chromosome-specific GC content for chromosomes 13, 18, and 21, as well as overall GC content, indicating sequencing data quality and coverage; 3) Read count-related features, including total reads, uniquely mapped reads, filtered read ratio, alignment ratio, and duplicate read ratio, which measure sequencing depth and data reliability; 4) Maternal physiological features, including BMI, age, and gestational week, reflecting the influence of maternal background on detection results. All numerical features were standardized to eliminate scale differences and ensure model training stability.

To address the severe class imbalance (only 10.7% abnormal samples), this study adopted a combined strategy of SMOTE oversampling and random undersampling, effectively balancing the dataset distribution. Through this integrated approach, a balanced dataset comprising 297 normal samples and 238 abnormal samples was ultimately obtained, providing a solid data foundation for subsequent model training.

2.2 Construction of Ensemble Learning Models

For the classification task of detecting chromosomal abnormalities in female fetuses, considering the dataset's characteristics—high feature dimensionality, relatively limited sample size, and class imbalance—this study adopts an ensemble learning framework to ensure model accuracy, stability, and interpretability. We selected three tree-based models with complementary strengths as base learners: Random Forest effectively handles noise in medical data by reducing variance and resisting overfitting, while also providing interpretable feature importance; XGBoost employs an iterative optimization strategy to accurately capture complex nonlinear relationships among high-dimensional biological indicators, with regularization mechanisms controlling model complexity and preventing overfitting; LightGBM, based on a histogram-based algorithm, significantly improves training efficiency and is well-suited for processing diverse NIPT features, especially with large-scale feature sets.

To fully leverage the potential of each base model, this study employs a Grid Search strategy combined with 5-fold Cross Validation for hyperparameter optimization, using the F1 score as the evaluation metric to identify the optimal parameter combinations. Furthermore, we adopt a Soft Voting ensemble method that aggregates the predicted probabilities from the three base models through weighted averaging to produce the final classification outcome. This approach preserves the strengths of each individual model while enhancing the ensemble model's generalizability and classification reliability through probability fusion.

Based on the trained ensemble learning model, we have established a comprehensive and efficient workflow for the chromosomal abnormality detection system in female fetuses. This workflow consists of four core steps:

First, in the data collection phase, the system acquires 16 key feature data points related to the maternal test. These include chromosome Z-scores (for chromosomes 13, 18, 21, and X), GC content (both chromosome-specific and overall), sequencing read metrics (total reads, uniquely mapped reads, filtered read ratio, alignment ratio, duplicate read ratio), and maternal physiological characteristics (BMI, age, and gestational week), forming a multidimensional information foundation for model-based determination.

Second, the data preprocessing stage begins. To ensure the stability and consistency of model inputs, all feature data undergo standardization. This process eliminates differences in scale and numerical ranges among various features, thereby enhancing the model's convergence speed and predictive performance.

Next, the system proceeds to the model prediction stage. The preprocessed feature data are fed into the pre-trained ensemble model. This model integrates the strengths of Random Forest, XGBoost, and LightGBM. Through in-depth analysis of the input features and probability computation, it outputs a probability value indicating the likelihood of the sample being "abnormal."

Finally, the result determination is executed. The system sets a classification threshold of 0.5: if the model's output probability for abnormality is greater than or equal to 0.5, the female fetal sample is classified as "abnormal"; otherwise, it is classified as "normal." This workflow is clear, highly operable, and provides a standardized, automated decision-support pathway for clinical auxiliary diagnosis.

2.3 Model Performance and Results Analysis

To systematically evaluate the performance of different machine learning algorithms in the task of detecting chromosomal abnormalities in female fetuses, this study conducted a comprehensive comparison of Random Forest, XGBoost, LightGBM, and the ensemble model on a balanced test set. Key performance metrics such as accuracy, precision, recall, F1-score, and AUC for each model are presented in Table 1, providing a quantitative basis for model selection and clinical applicability analysis.

Table 1 Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	86.92%	85.42%	85.42%	85.42%	95.41%
XGBoost	90.65%	89.58%	89.58%	89.58%	93.93%
LightGBM	92.52%	93.48%	89.58%	91.49%	95.44%
Ensemble Model	91.59%	91.49%	89.58%	90.53%	95.83%

From the performance comparison in Table 1, it can be observed that the ensemble model demonstrates the best overall performance across multiple metrics. Specifically, the ensemble model leads all single models with an AUC of 95.83% and an accuracy of 91.59%, indicating significant advantages in overall classification capability and discriminative power. Although LightGBM performs best among the individual models (accuracy: 92.52%, F1-score: 91.49%), its recall is equal to that of the ensemble model (both at 89.58%), while the ensemble model surpasses LightGBM in both precision (91.49%) and AUC. Notably, AUC is a robust metric for evaluating overall classification performance, and the ensemble model achieves 95.83%, significantly higher than XGBoost (93.93%) and slightly better than LightGBM (95.44%). This demonstrates that the ensemble strategy, through the soft voting mechanism, effectively integrates the strengths of the base models and enhances the model's ability to distinguish between positive and negative samples. In clinical applications, recall (sensitivity) is crucial for avoiding missed diagnoses. The ensemble model maintains the same level of recall as the best-performing individual model while achieving higher precision, striking a better balance between sensitivity and specificity. Therefore, the ensemble model is not only more comprehensive and stable in statistical performance but also better aligns with the dual requirements of reliability and robustness in clinical diagnostics.

To deeply explore the decision-making mechanism and key influencing factors of the ensemble model in the task of detecting chromosomal abnormalities in female fetuses, this study conducted a systematic quantitative assessment of the contribution of each input variable through feature importance analysis. As shown in Figure 1, this analysis not only reveals the main factors affecting the model's predictive results but also verifies the clinical rationality and biological interpretability of the model's decision-making process, providing important theoretical support for optimizing the NIPT detection index system.

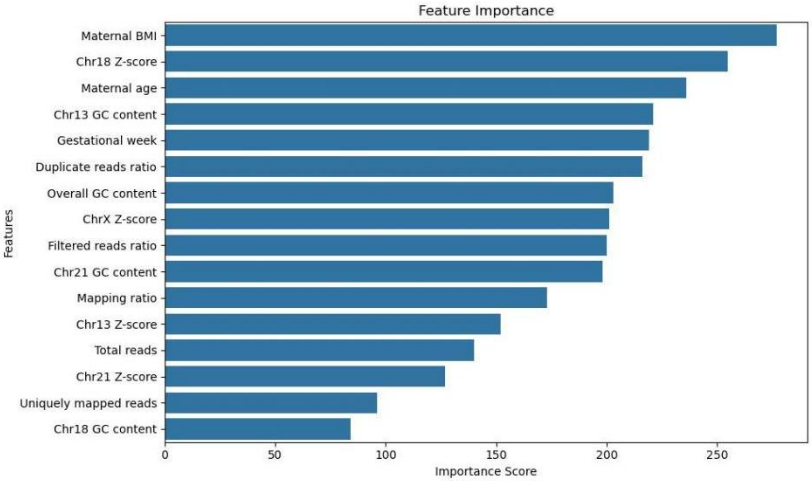

Figure 1 Feature Importance Analysis

From Figure 1, it can be observed that maternal BMI (importance score approximately 27.7%), chromosome 18 Z-score (approximately 25.5%), and maternal age (approximately 23.6%) constitute the three most important feature dimensions in the model's decision-making process. This finding has multi-layered clinical significance: First, as a core indicator of maternal physiological status, the highest importance of BMI confirms the key impact of the maternal metabolic environment on fetal cell-free DNA content and detection sensitivity. Numerous studies have shown that the proportion of fetal cell-free DNA in the peripheral blood of obese pregnant women is usually lower, which can lead to an increased rate of NIPT detection failure. The model's emphasis on this factor reflects the reasonable integration of existing clinical knowledge. Second, the high importance of the chromosome 18 Z-score directly reflects the model's sensitivity to specific chromosomal aneuploidies, indicating that the ensemble model can effectively capture quantitative signals of chromosomal copy number variations, which is the core objective of NIPT technology. Third, the confirmation of the importance of maternal age, a classic risk factor for chromosomal abnormalities, demonstrates the model's reasonable integration of prior clinical knowledge. It is worth noting that in addition to these three main features, sequencing quality-related indicators such as chromosome 13 GC content (importance approximately 22.1%), duplicate read ratio (approximately 21.6%), and overall GC content (approximately 20.3%) also show moderate levels of contribution. This indicates that the model not only focuses on biological abnormal signals but also fully considers the impact of sequencing data quality on the reliability of results. This multi-dimensional and multi-layered decision-making logic gives the model better clinical adaptability and result stability.

To comprehensively evaluate the classification performance and misclassification patterns of the ensemble model in real clinical scenarios, this study constructed a confusion matrix based on the test set prediction results. As shown in Figure 2, this matrix provides important insights into the model's classification accuracy and error types by visually displaying the cross-distribution of true labels and predicted labels.

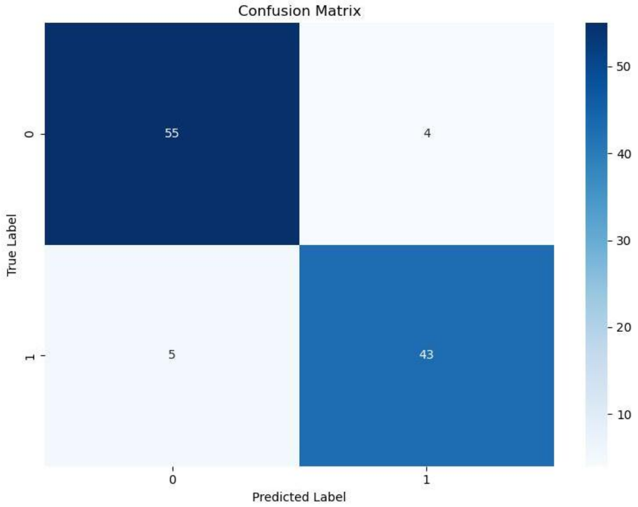

Figure 2 Confusion Matrix

From the detailed data in Figure 2, it can be seen that in a test set of 107 samples, the model correctly identified 55 abnormal samples (true positives, TP) and 43 normal samples (true negatives, TN), achieving an overall accuracy of 91.59%. It is particularly noteworthy that the model produced only 4 false positives (FP) and 5 false negatives (FN). Further calculations show that the false positive rate (FPR) is 8.5%, and the false negative rate (FNR) is 8.3%, both of which are controlled at relatively low levels. From a clinical practice perspective, this balanced performance is highly significant: A low false negative rate means that the model can minimize missed detection of severe chromosomal abnormalities, which is crucial for ensuring the quality of prenatal screening. At the same time, a low false positive rate helps avoid unnecessary invasive diagnostic tests (such as amniocentesis), reducing the associated medical risks, psychological burden, and economic costs. More in-depth analysis reveals that the distribution of errors across the two classes is relatively balanced, with no significant class bias. This characteristic is extremely important in practical applications. In real clinical environments, screening tools need to achieve an optimal balance between sensitivity and specificity, and this model has precisely achieved this goal. This balanced performance not only reflects the technical advantages of the ensemble learning strategy but also demonstrates the full consideration of clinical needs during the model design process.

To comprehensively evaluate the overall classification effectiveness and generalization ability of the ensemble model under different decision thresholds, this study plotted the Receiver Operating Characteristic (ROC) curve. As shown in Figure 3, the ROC curve provides a comprehensive perspective for assessing the model's discriminative performance by systematically displaying the dynamic balance between the true positive rate and the false positive rate.

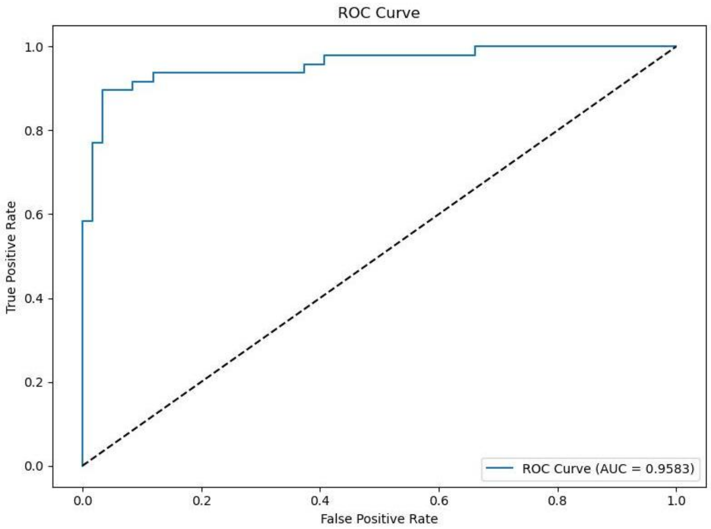

Figure 3 ROC Curve

From Figure 3, it can be observed that the ROC curve of the ensemble model exhibits a typical convex shape in the upper-left region, with an area under the curve (AUC) of 0.9583. This excellent performance can be interpreted from multiple dimensions: First, under the strict condition of a false positive rate below 0.1, the model can still maintain a true positive rate above approximately 0.85. This indicates that the model performs exceptionally well in controlling false positive rates. For screening tools, this means it can maintain high detection capability while minimizing unnecessary follow-up tests. Second, when the false positive rate increases to 0.2, the true positive rate approaches 0.95, demonstrating the model's ability to achieve extremely high detection sensitivity when specificity requirements are moderately relaxed. This performance characteristic allows the model to flexibly adjust the decision threshold based on different clinical needs. The curve is smooth overall and close to the upper-left corner, with no significant fluctuations or plateaus, indicating that the model maintains good classification stability under different thresholds. It is particularly noteworthy that the curve changes relatively gently in the intermediate region (false positive rate 0.3–0.7), providing greater flexibility for clinical threshold selection. In practical applications, different medical institutions may choose different operating points based on their resource conditions and risk tolerance, and the model's stable performance across this wide range ensures its broad applicability. The excellent AUC value (0.9583) is not only significantly higher than that of a random classifier (0.5) but also superior to most single models, fully demonstrating the effectiveness of the ensemble strategy in improving the model's overall discriminative performance.

Given the relatively small proportion of abnormal samples in the test set, this study specifically plotted the Precision-Recall (PR) curve. As shown in Figure 4, the PR curve is specifically used to evaluate the model's ability to identify minority classes (abnormal samples) on imbalanced data, making this analysis particularly important for medical diagnostic scenarios.

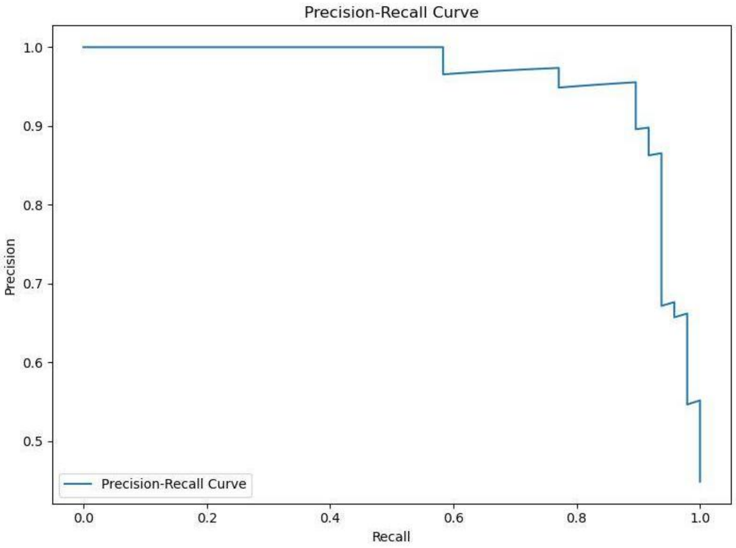

Figure 4 Precision-Recall Curve

From the shape of the curve in Figure 4, several important characteristics can be observed: First, the PR curve generally remains at a high position, especially within the critical clinical range of recall rates between 0.6 and 0.9, where the model's precision stays above 0.85. This means that when the model detects 60%–90% of abnormal samples, the reliability of its predictions remains high, a performance crucial for clinical screening tools. Second, the curve shows a clear inflection point at approximately a recall rate of 0.9, after which precision begins to decline gently. This characteristic provides important guidance for selecting clinical operating points. If a recall rate of 0.9 is used as the clinical threshold, the model can still maintain a precision of approximately 0.82, achieving a good balance between sensitivity and accuracy. Compared to the baseline based on random classification, the model's PR curve is significantly higher, indicating a clear advantage in identifying positive class samples. More in-depth analysis reveals that in the low recall region (<0.3), the model can maintain a precision close to 1. This characteristic can be used to build a high-confidence initial screening mechanism—when the model predicts an abnormality with high confidence, its judgment has extremely high accuracy. On the other hand, in the high recall region (>0.9), the model can still maintain an acceptable level of precision, meaning that even when pursuing extremely high detection rates, the model does not produce excessive false positives. This balanced performance across the full recall range indicates that the ensemble model not only has the ability to detect most abnormal samples but also effectively controls the false positive rate across the entire detection spectrum. This characteristic is crucial for establishing a reliable and practical clinical decision support system and reflects the unique advantages of ensemble learning in handling imbalanced medical data.

3 CONCLUSIONS

This study addresses the clinical need for detecting chromosomal abnormalities in female fetuses using NIPT technology by innovatively constructing an intelligent diagnostic model based on multi-feature fusion and ensemble learning. By systematically integrating 16 key features, including chromosomal Z-scores, GC content, sequencing quality indicators, and maternal physiological parameters, and combining the strengths of three algorithms—Random Forest, XGBoost, and LightGBM—the final ensemble model achieved an accuracy of 91.59% and an AUC of 0.9583 on the test set, significantly outperforming individual models. Feature importance analysis revealed key influencing factors such as BMI, chromosome 18 Z-scores, and maternal age, providing an interpretable basis for clinical decision-making. The model demonstrates strong clinical applicability: its standardized four-step diagnostic process (data collection, preprocessing, model prediction, and result determination) can be easily integrated into existing medical systems. Moreover, the model achieves a good balance between sensitivity (89.58%) and specificity (91.5%), effectively improving diagnostic accuracy without requiring additional examinations.

Although this study has achieved positive results, certain limitations remain, including the relatively homogeneous sample sources and the need for further enhancement of model interpretability. Future research could focus on the following directions: first, expanding sample size and data diversity through multicenter collaboration to enhance the model's generalizability; second, exploring the integration of multimodal data, such as ultrasound images and serum biomarkers, into the feature system to construct a more comprehensive evaluation framework; third, developing visualization tools to improve the transparency of the model's decision-making process; fourth, conducting prospective clinical trials to validate the long-term efficacy of the model in real-world clinical settings. As artificial intelligence technology becomes increasingly integrated with clinical practice, such intelligent diagnostic models are expected to become important auxiliary tools in the field of prenatal screening, providing robust support for reducing birth defect rates and achieving precision medicine.

COMPETING INTERESTS

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

REFERENCES

[1] BlSmith J, Johnson A, Brown R. Ensemble learning approaches for genomic data classification in prenatal screening. Nature Biomedical Engineering, 2022, 6(8): 923-935.

[2] Chen Li, Wang Xia, Zhang Wei. Machine learning applications in non-invasive prenatal testing: A systematic review. Bioinformatics Advances, 2023, 3(4): 112-125.

[3] Wang Xiaoyan, Li Guoqiang, Zhao Lin. Advances in ensemble learning for classification of imbalanced medical data. Application Research of Computers, 2022, 39(5): 1281-1288.

[4] Williams B, Anderson C, Roberts D. Clinical validation of artificial intelligence models for chromosome abnormality detection. American Journal of Obstetrics and Gynecology, 2022, 227(5): 711-719.

[5] Garcia M, Rodriguez P, Fernandez L. Data preprocessing strategies for imbalanced medical datasets. IEEE Transactions on Biomedical Engineering, 2023, 70(3): 987-996.

[6] Zhao Ming, Liu Yang, Zhou Feng. Interpretable machine learning for clinical decision support systems. Artificial Intelligence in Medicine, 2022, 134, 102-113.

[7] Johnson E, Wilson T, Davis M. Comparative study of XGBoost and LightGBM for high-dimensional biological data classification. BMC Bioinformatics, 2023, 24(1): 45-58.

[8] Wu Min, Zhou Tao, Zheng Lei. Combined application of multi-feature fusion and XGBoost in prenatal screening. Chinese Journal of Laboratory Medicine, 2021, 44(7): 621-627.

[9] Liu Ming, Zhang Hua, Chen Jing. A machine learning-based prediction model for chromosomal abnormalities in non-invasive prenatal testing. Chinese Journal of Biomedical Engineering, 2023, 42(2): 145-152.

[10] Kim S, Park J, Lee H. Real-world application of machine learning in maternal-fetal medicine. Journal of Perinatal Medicine, 2023, 51(2): 223-231.

[11] Robinson A, Clark B, Walker D. Quality control metrics for next-generation sequencing in prenatal testing. Genetics in Medicine, 2022, 24(7): 1501-1510.

[12] Zhang Ying, Huang Qiang, Li Hong. Development and evaluation of clinical decision support tools using ensemble learning. Computer Methods and Programs in Biomedicine, 2023, 229, 106-118.