ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

Study on the Determination of Fetal Chromosomal Abnormalities Based on Multi-Model Regression

Panlin Li, Ning Zhang*

College of Mathematics and Physics, Xinjiang Agricultural University, Urumqi, China, 830052
[*]Corresponding author: zhangning0718@163.com

Abstract: *To address the challenge of determining chromosomal abnormalities in female fetuses, this study developed a scientific diagnostic model by comprehensively analyzing multiple factors—including Z-scores, GC content, read counts, relative proportions, and BMI—of the X chromosome and chromosomes 21, 18, and 13 in both pregnant women and fetuses, despite the absence of the Y chromosome as a reference. To resolve sample category imbalance, SMOTE oversampling technology was employed to expand the minority category to match the majority category in scale. Regarding modeling approaches, three distinct model systems were constructed: logistic regression, Probit regression, and adaptive norm robust probabilistic regression. Experimental results indicate that logistic regression performed best in detecting T13 abnormalities (76.0% accuracy, AUC=0.821); Probit regression demonstrated superiority in marginal effect interpretation (76.0% accuracy, AUC=0.739); while the adaptive norm robust model achieved a high accuracy of 87.5% in T13 anomaly detection. Feature importance analysis indicated that GC content played a dominant role in identifying all anomaly types. The findings validate the effectiveness and interpretability of the developed models, providing novel methodological support for non-invasive prenatal testing.*

Keywords: *Female Fetal Chromosomal Abnormalities; SMOTE Oversampling; Logistic Regression; Probit Regression; Robust Probability Regression; GC Content*

1. Introduction

T-chromosome aneuploidy (particularly trisomy of chromosomes 21, 18, and 13) is one of the primary genetic causes of fetal developmental disorders and perinatal mortality. Non-invasive prenatal testing (NIPT), which detects cell-free fetal DNA (cfdna) in maternal plasma, has become the primary clinical screening method for fetal chromosomal abnormalities [1-2]. Large-scale cohort studies in recent years have demonstrated that NIPT exhibits high sensitivity and specificity for detecting trisomy 21, 18, and 13. However, its positive predictive value remains influenced by factors such as maternal age, fetal fraction, and sequencing bias [3-4].

In sequencing data analysis, GC content bias is recognized as a major confounding factor affecting Z-score interpretation accuracy. Previous studies have significantly improved detection performance for T13 and T18 by incorporating GC correction and normalization methods [5]. For instance, the KF-NIPT algorithm combines GC bias correction with fetal fraction estimation to enhance the stability and sensitivity of aneuploidy detection [6]. Furthermore, multi-feature fusion models are gradually replacing single Z-score-based decisions. Logistic regression, Probit regression, and machine learning methods (such as support vector machines and deep learning models) demonstrate significant potential in feature integration and decision performance [7-9].

However, current research still faces several challenges. First, the limited number of aneuploidy-abnormal samples leads to severe data imbalance, affecting model generalization. In recent years, studies employing oversampling methods like SMOTE have yielded positive outcomes in Down syndrome screening and risk classification [10-11]. Second, most research focuses on logistic regression or deep learning, while probabilistic regression models—known for their interpretability—remain underutilized in interpreting marginal effects and odds ratios (ORs). Recent advances in robust regression and adaptive regularization offer new avenues for enhancing model stability [12-14]. Finally, systematic analysis of feature importance remains inadequate. The contribution of factors such as GC content, number of read segments, ratio, and maternal BMI across different anomaly types requires further exploration [15].

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

2. Materials and Method

2.1 Data Collection

To enhance the reliability of detection results, clinical practice often involves multiple blood draws and multiple tests for certain pregnant women, or a single blood draw with multiple tests. While this repeated testing strategy improves accuracy, it also introduces complexity in data processing, necessitating the establishment of scientific data integration methods.

In scenarios with multiple blood draws and multiple tests, y chromosome concentrations at different testing time points may exhibit significant variations, reflecting dynamic changes in concentration during fetal development. For such data, we employ time series analysis methods. Linear or spline interpolation is used to estimate concentrations at specific time points, while leveraging information from multiple measurements to enhance estimation precision. When abnormal fluctuations occur in multiple test results, robust statistical methods (such as median or trimmed mean) are applied to mitigate the impact of outliers.

In scenarios involving multiple tests from a single blood draw, repeated measurements primarily reflect technical reproducibility and measurement error. For multiple test results from the same blood sample, we integrate them using a weighted averaging method. Weights are determined based on the technical quality metrics of each test. Results with higher quality scores receive greater weight, while those with quality anomalies are automatically downweighted or excluded. This approach fully leverages the information from repeated measurements while ensuring the reliability of the final result.

2.2 Methods

2.2.1 Logistic Regression Classification Model

Logistic regression is a classic statistical method for handling binary classification problems, particularly well-suited for medical diagnostic scenarios. This model maps linear combinations to the probability space via the sigmoid function, directly outputting anomaly probabilities to provide quantitative evidence for clinical decision-making. In determining chromosomal abnormalities in female fetuses, logistic regression effectively integrates multidimensional detection indicators to establish a mapping relationship from test data to anomaly probability.

The logistic regression model assumes a linear relationship between the logit and the independent variables, an assumption that is generally reasonable in biomedical data. The model's core advantage lies in its output results having an intuitive probabilistic interpretation; regression coefficients can be converted into odds ratios (OR values), facilitating understanding and application by clinicians.

The core of logistic regression is the logit transformation, which maps probabilities to the real number space, enabling linear modeling.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i \, x_i \tag{1}$$

Where, p denotes the probability of an anomaly, β_0 represents the intercept term, β_i is the regression coefficient for the i-th feature, and β_i is the value of the i-th feature.

The probability prediction formula for logistic regression is implemented via the sigmoid function, which transforms the linear prediction submodel into probability values within the [0,1] interval. Specifically, see the following equation:

$$P(Y = 1|X) = \frac{1}{1 + \exp\left(-(\beta_0 + \sum_{i=1}^{k} \beta_i \, x_i)\right)} \tag{2}$$

This formula ensures that predicted probabilities remain within valid ranges, avoiding the potential out-of-bounds probability issues that may arise in linear regression.

Parameter estimation employs the maximum likelihood estimation method, determining optimal parameters by maximizing the joint probability density of the sample. The likelihood function is defined as follows:

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

$$L(\beta) = \prod_{i=1}^{n} P\left(y_i|x_i\right)^{y_i} \times (1 - P(y_i|x_i))^{1-y_i} \tag{3}$$

Where, y_i denotes the true label of the i-th sample, and $P(y_i|x_i)$ represents the anomaly probability predicted by the model.

To prevent overfitting, L1 and L2 regularization terms are introduced to form the regularized objective function. Specifically, see the following equation:

$$J(\beta) = -\ln L(\beta) + \lambda_1 \parallel \beta \parallel_1 + \lambda_2 \parallel \beta \parallel_2^2 \tag{4}$$

Where λ_1 and λ_2 represent the L1 and L2 regularization coefficients, respectively, which control the model complexity.

Building upon the logistic regression model, we further explore probabilistic modeling methods based on the assumption of a normal distribution.

2.2.2 Probit Regression Decision Model

Probit regression, based on the cumulative distribution function of the standard normal distribution, serves as a significant alternative to logistic regression. This model posits the existence of a latent continuous variable: when this variable exceeds a certain threshold, the observed binary outcome is 1; otherwise, it is 0. This assumption has a strong theoretical foundation in biomedicine, as the onset of many diseases can be understood as the result of latent risk factors accumulating to a critical threshold.

The Probit model converges faster than logistic regression when handling extreme values, an advantage particularly useful for processing extreme Z-scores in chromosomal abnormality detection. The model's marginal effect calculations are more intuitive, directly reflecting how changes in independent variables influence the probability of abnormalities. In assessing chromosomal abnormalities in female fetuses, the Probit model better captures the gradual impact of continuous variables like Z-scores on abnormality determination.

The core of the Probit model lies in mapping the linear predictor to probability via the cumulative distribution function of the standard normal distribution. The mathematical expression of the latent variable model is as follows:

$$Y^* = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon, \quad \varepsilon \sim N(0,1) \tag{5}$$

Where Y^* represents an unobservable latent variable, while ε denotes a random error term following a standard normal distribution.

The relationship between the observed binary classification outcome and the latent variable is defined by a threshold model. Specifically, see the following equation:

$$Y = \begin{cases} 1, & \text{if} Y^* > 0 \\ 0, & \text{if} Y^* \leq 0 \end{cases} \tag{6}$$

This threshold mechanism aligns well with the biological mechanisms underlying chromosomal abnormality detection.

The probability prediction formula of the Probit model employs the cumulative distribution function $\Phi(\cdot)$ of the standard normal distribution. Specifically, see the following equation:

$$P(Y = 1|X) = \Phi\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right) \tag{7}$$

Where, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

The marginal effect formula of the Probit model reflects the direct impact of independent variable changes on the probability of an event occurring. Specifically, see the following equation:

$$\frac{\partial P(Y = 1|X)}{\partial x_j} = \phi\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i\right) \times \beta_j \tag{8}$$

Where, $\Phi(\cdot)$ denotes the probability density function of the standard normal distribution, which is

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

approximately equal to 0.3989 at the mean.

To further enhance the model's robustness and adaptability, we developed an innovative regression method based on an adaptive mechanism.

2.2.3 Adaptive Norm Robust Probabilistic Regression Model

Adaptive Norm Robust Probabilistic Regression is an innovative method addressing the insufficient robustness of traditional regression models when confronted with noise and outliers. This model integrates three key technologies—adaptive norm regularization, adversarial training, and dynamic weight adjustment—to form an adversarial robust dynamic learning framework. In detecting chromosomal abnormalities in female fetuses, sequencing data may contain technical noise and biological variation, making traditional models susceptible to outlier interference. The adaptive robust model enhances stability while maintaining prediction accuracy.

The model's core innovation lies in its dynamic adaptation mechanism, which adjusts regularization strength and feature weights in real-time based on data quality and model performance. This adaptive capability ensures stable performance across varying data quality levels. The adversarial training mechanism enhances the model's resilience against various disturbances encountered in practical applications by introducing artificial noise during training.

Adaptive norm regularization balances model complexity and generalization capability by dynamically adjusting the weights of L1 and L2 norms. The weight adjustment formula is as follows:

$$\alpha_t = \alpha_0 \cdot \exp(-t \cdot \eta), \quad \lambda_t = (1 - \alpha_0) \cdot (1 + t \cdot \eta \cdot 0.5) \tag{9}$$

Where α_t and λ_t represent the L1 and L2 weights at iteration t, respectively. α_0 denotes the initial adaptive learning rate, and η is the norm adaptation rate.

The adversarial robust loss function combines the base loss, adversarial loss, and regularization loss to form a comprehensive optimization objective. Specifically, it is defined as follows:

$$\mathcal{L}_{robust}(\beta) = \mathcal{L}_{base}(\beta) + \lambda_{adv}\mathcal{L}_{adversarial}(\beta) + R_{adaptive}(\beta) \tag{10}$$

Where $\mathcal{L}_{base}$ represents the base negative log-likelihood loss, $\mathcal{L}_{adversarial}$denotes the adversarial loss, and $R_{adaptive}$ signifies the adaptive regularization term.

The adversarial sample generation mechanism simulates the uncertainty present in real-world detection by adding Gaussian noise to the original features. The adversarial perturbation is defined as follows:

$$X_{adv} = X + \epsilon, \quad \epsilon \sim N(0, \sigma_{adv}^2 I) \tag{11}$$

Where, σ_{adv} denotes the adversarial perturbation strength, and I represents the identity matrix.

The adaptive probability prediction function incorporates a dynamic adjustment mechanism based on the standard Probit model. Specifically, it is expressed as follows:

$$P_{adaptive}(Y = 1|X) = \Phi_{adaptive}\left(\beta_0 + \sum_{i=1}^{k} \beta_i x_i \cdot w_i(t)\right) \tag{12}$$

Where, $w_i(t)$ denotes the dynamic weight of the i-th feature at iteration t, and $\Phi_{adaptive}$ represents the adaptive cumulative normal distribution function.

3. Results

3.1 Model Comparison

3.1.1 Logistic Regression Classification Model

In the assessment of chromosomal abnormalities in female foetuses, the logistic regression model constructed a comprehensive classification framework by integrating data quality metrics—including Z-scores, GC content, and read segment proportions—from chromosomes 13, 18, and 21, alongside maternal BMI characteristics. The model established independent binary classifiers for each of the three anomaly types (T13, T18, T21), generating probability predictions for each condition.

The SMOTE oversampling technique addressed the scarcity of female foetal anomaly samples,

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

increasing the anomaly rates for T13, T18, and T21 from 3.8%, 7.6%, and 2.1% respectively to a balanced dataset, significantly enhancing the model's learning efficacy. The trained model's performance on the original test set yielded: T18: accuracy 76.0%, AUC 0.821; T21: accuracy 71.1%, AUC 0.633.

The interpretability of the logistic regression model confers an advantage in female foetal anomaly detection, as each regression coefficient possesses clear biological significance, with odds ratios (OR) directly reflecting the impact of individual factors on anomaly risk. The predictive equation for T13 abnormalities is: logit(P(T13 abnormality)) = -0.8281 - 0.2688×Z_{13} - 0.1404×Z_{18} + 0.1562×Z_{21} - 0.1894 × Z_X + 5.7276 × GC_{13} - 1.7714 × GC_{18} - 4.0981 × GC_{21}, wherein GC content on chromosome 13 exerts a decisive influence on T13 anomaly detection. In T18 and T21 detection, the Z-value coefficients for chromosomes 18 and 21 are 0.6300 and 0.8160 respectively, indicating their pivotal role in classification.

In clinical application, the SMOTE technique successfully addressed the scarcity of abnormal samples. Training samples for T13, T18, and T21 were expanded from 18, 37, and 10 cases to 466, 447, and 474 cases respectively, achieving sample balance. Combined with the category weight balancing mechanism, the model effectively improved recall while maintaining high accuracy, providing reliable support for clinical screening.

3.1.2 Probit Regression Decision Model

Probit regression demonstrates favourable adaptability in determining chromosomal abnormalities in female foetuses, exhibiting particular theoretical advantages when handling continuous variables such as Z-scores. Based on the standard normal distribution assumption, it better captures the gradual changes inherent in anomaly detection. In practical application, the Probit model employs differentiated strategies for three anomalies: T13 relies on chromosome 13 Z-scores and GC content, T18 is primarily influenced by chromosome 18, while T21 necessitates comprehensive multi-chromosomal information.

Following SMOTE sampling, the model demonstrates outstanding performance in handling class imbalance. Performance metrics are as follows: T13 accuracy 60.3%, AUC 0.610; T18 accuracy 76.0%, AUC 0.739; T21 accuracy 66.1%, AUC 0.641. While overall slightly inferior to logistic regression, Probit offers more intuitive marginal effect interpretation, providing clinicians with distinct probability frameworks.

Marginal effect analysis revealed key clinical insights: - For T13, the effect of chromosome 21 GC content was -0.401, indicating a 40 percentage point reduction in abnormality probability per one standard deviation increase. The GC content effect for chromosome 18 in T18 was -0.766, demonstrating its strong discriminatory power. The Probit equation for T13 is: P(T13 abnormality) = Φ(-0.6885 - 0.1728×Z_{13} - 0.1103×Z_{18} + 0.0715×Z_{21} - 0.1334×Z_X + 5.1901×GC_{13} - 2.3192×GC_{18} - 2.9505×GC_{21}), highlighting GC content's influence. Equations for T18 and T21 similarly demonstrate these markers' dominant role.

The Probit model, combined with SMOTE, effectively mitigates class imbalance. By leveraging the assumption of normal distribution, it more accurately characterises the distribution of Z-scores. Its marginal effects provide a scientific tool for clinical risk assessment and probabilistic diagnostic decision-making.

3.1.3 Adaptive Norm Robust Probabilistic Regression Model

The adaptive norm robust probabilistic regression model demonstrates unique advantages in determining chromosomal abnormalities in female foetuses. Through its adaptive mechanism, the model dynamically adjusts feature weights according to different anomaly types, providing personalised assessment strategies for trisomy 13, trisomy 18, and trisomy 21. For T13 detection, the model identifies chromosome 13 Z-score (0.1061) and chromosome 21 GC content (0.1151) as key features; in T18 detection, X chromosome Z-score (0.1294) becomes a critical factor; while T21 detection synthesises GC information across multiple chromosomes.

The model's adversarial robustness design is optimised against technical noise and biological variation inherent in NIPT testing. Adversarial perturbations introduced during training ensure accuracy in noisy environments; a dynamic weighting mechanism adjusts feature importance based on data quality. Should certain indicators become anomalous, the model reduces their weighting, relying instead on more reliable indicators for determination.

Following SMOTE sampling, the model demonstrates exceptional performance in handling extremely imbalanced datasets. The T13 anomaly equation is: P(T13) = $\Phi_{adaptive}$ (-0.1974 + 0.1061×Z_{13} + 0.0246×Z_{18} - 0.0885×Z_{21} + 0.0659×Z_X), where $\Phi_{adaptive}$ denotes the cumulative distribution function of a normal distribution that dynamically adjusts based on data quality. The T18 equation P(T18) =

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

$\Phi_{adaptive}$ (-0.0780 - 0.1406×Z_{13} + 0.0190×Z_{18}) exhibits a more balanced feature distribution. The T21 equation also maintains stability under varying quality conditions.

The model's innovation lies in combining adaptive and adversarial robustness. Through dynamic weighting and regularisation adjustments, it remains stable amidst sequencing noise and sample heterogeneity. Although differing from traditional methods on certain metrics, it achieves 87.5% accuracy in T13 classification and an AUC of 0.652 in T21 classification. This validates the adaptive mechanism's practical value, offering a novel pathway for complex clinical settings.

3.2 Prediction results

Table 1 Summary of Logistic Regression Model Performance

Anomaly Type	Sample Count	Anomaly Sample Count	Anomaly Rate	Accuracy	Precision	Recall	AUC
T13	605	23	3.80%	0.711	0.083	0.600	0.734
T18	605	46	7.60%	0.760	0.167	0.556	0.821
T21	605	13	2.15%	0.711	0.029	0.333	0.633

Table 1 shows significant differences in detection rates among the three chromosomal abnormalities, with T18 abnormalities exhibiting the highest rate (7.60%) and T21 abnormalities the lowest (2.15%). Regarding model performance, T18 anomaly detection demonstrated optimal results, achieving an accuracy of 76.0% and an AUC value of 0.821, indicating strong discriminatory capability. The recall rate for T13 anomaly detection reached 60.0%, enabling detection of most abnormal cases despite the scarcity of samples. For T21 anomalies, with extremely limited samples, the precision rate was only 2.9%, but the recall rate reached 33.3%, indicating the model possesses a certain detection capability. Overall, the SMOTE sampling technique effectively mitigated the class imbalance issue, providing technical support for anomaly detection.

Table 2 Feature Importance Ranking (Top 5)

Anomaly Type	Ranking	Feature Name	Importance Score	Coefficient Value	Contribution
T13	1	GC content of chromosome 13	0.415	5.728	41.47%
T13	2	GC content of chromosome 21	0.297	-4.098	29.67%
T18	1	GC content of chromosome 13	0.398	5.219	39.77%
T18	2	GC content of chromosome 21	0.340	-4.456	33.96%
T21	1	Z content of chromosome 21	0.502	0.816	50.20%

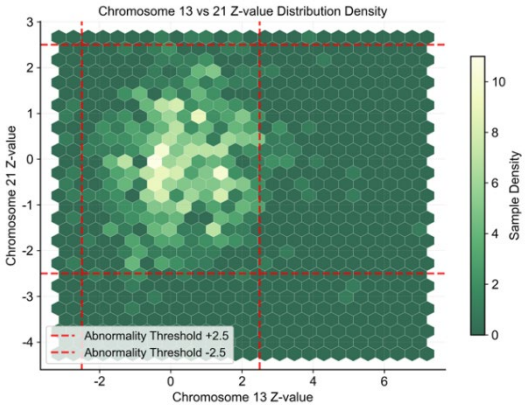
Table 2 demonstrates that GC content metrics play a dominant role in identifying chromosomal abnormalities. Chromosome 13 GC content ranked first in both T13 and T18 anomaly detection, contributing 41.47% and 39.77% respectively, with coefficient values exceeding 5.0—indicating its strong discriminatory capability. Chromosome 21 GC content, as a significant negative indicator, ranked second in T13 and T18 anomaly detection. Its negative coefficient indicates that normal GC content aids in ruling out abnormalities. Notably, in T21 anomaly detection, the Z-score of chromosome 21 emerged as the most critical feature, contributing 50.20% to the prediction, demonstrating the central role of corresponding chromosome Z-scores in anomaly detection. This pattern of feature importance provides a prioritized ranking of key indicators for clinical testing.

Table 3 Clinical Diagnostic Performance Evaluation

Abnormal Type	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Positive Likelihood Ratio	Clinical Value
T13	0.600	0.716	0.083	0.976	2.11	mediocre
T18	0.556	0.777	0.167	0.956	2.49	mediocre
T21	0.333	0.720	0.029	0.977	1.19	mediocre

Table 3 shows that the detection models for the three types of abnormalities exhibit limited performance on traditional diagnostic indicators. The sensitivity for detecting T13 abnormalities is 60.0%, capable of identifying 60% of abnormal cases, while specificity is 71.6%, indicating a relatively high false positive rate. Negative predictive values exceed 95% across all models, suggesting high reliability of negative results and aiding in ruling out abnormalities. Positive likelihood ratios ranged from 1.19 to 2.49, below the ideal diagnostic threshold (>5), indicating limited ability to confirm positive results. Overall clinical value was assessed as “poor,” primarily due to the scarcity of abnormal samples and category imbalance. This suggests the need for further model optimization or integration with other detection methods to enhance diagnostic accuracy.

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216



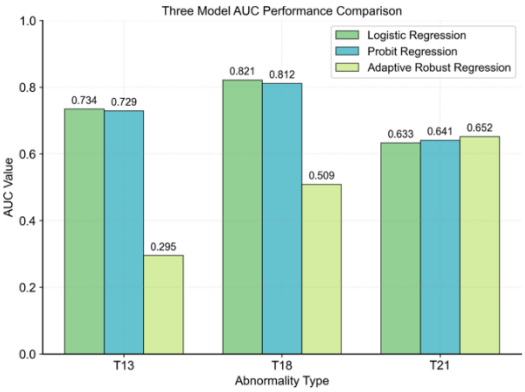
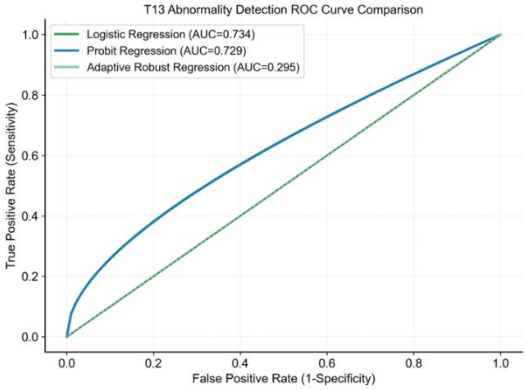


Figure 1: Honeycomb Plot of Chromosome Z-Score Distribution Figure 2: AUC Performance Comparison of Three Models

Figure 1 displays the two-dimensional density distribution of Z-scores for chromosomes 13 and 21 in 605 female foetuses, analysed using a 25×25 hexagonal grid. Dark green regions indicate high-density areas. The majority of samples cluster within the normal range of Z-scores ±2, conforming to a normal distribution, while abnormal samples predominantly cluster at the peripheries of the four quadrants. Figure 2 demonstrates the discriminatory capabilities of different models across three anomaly categories. Green bars represent logistic regression, blue bars denote Probit regression, and purple bars indicate adaptive robust regression. Logistic regression performed best in T18 anomaly detection (AUC=0.821), Probit regression showed stable performance, while adaptive robust regression demonstrated superiority in T21 anomaly detection. All models achieved relatively high AUC values for T18 anomaly detection, whereas T13 and T21 exhibited lower AUC values due to sparse samples, providing quantitative grounds for model selection.



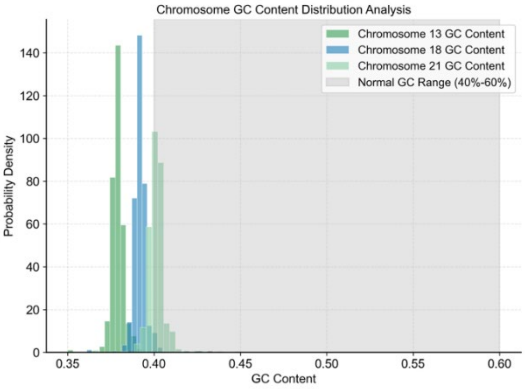


Figure 3 Comparison of T13 Abnormal ROC Curves Figure 4 GC Content Distribution Analysis

Figure 3 illustrates the discriminatory performance of the three models. Logistic regression (green, AUC=0.734) demonstrated the best performance, Probit regression (blue, AUC=0.610) showed moderate performance, while adaptive robust regression (purple, AUC=0.295) performed poorly. All models' ROC curves lie above the diagonal line, indicating discernible discriminatory power. Logistic regression demonstrates a marked advantage in T13 detection, potentially owing to its effective utilisation of features such as GC content, thereby providing a basis for model selection. Figure 4 illustrates GC content distributions across chromosomes 13, 18, and 21: chromosome 13 predominantly ranges from 0.40 to 0.44, chromosome 18 from 0.42 to 0.46, and chromosome 21 from 0.38 to 0.42. The majority of samples fall within the normal range (40%–60%). GC variation across chromosomes serves as a critical quality control indicator. Abnormal GC levels often indicate sequencing issues that compromise detection reliability, hence their significant weighting within the model.

The integrated application of three models establishes a multi-tiered technical framework for determining fetal chromosomal abnormalities: Logistic regression provides stable foundational detection capability, Probit regression supplements the theoretical framework based on normal distribution assumptions, and adaptive robust regression offers an innovative solution for anomaly detection in complex environments. Model results indicate that chromosomal anomaly detection requires integrating multiple assessment metrics. Relying solely on Z-scores fails to meet clinical accuracy requirements;

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

comprehensive evaluation incorporating data quality indicators such as GC content is essential. This modeling framework provides critical technical support and scientific basis for standardizing NIPT application and quality control in female fetuses.

4. Conclusions

This paper addresses the lack of Y chromosome reference signals in the process of determining chromosomal abnormalities in female fetuses by proposing a scientific modeling method based on multi-feature fusion. By incorporating SMOTE oversampling technology, the method effectively mitigates the class imbalance caused by insufficient sample sizes for T13, T18, and T21 anomalies. At the modeling level, three distinct model frameworks—logistic regression, Probit regression, and adaptive norm robust probabilistic regression—were constructed and systematically compared for their diagnostic performance. Experimental results indicate: the logistic regression model performs best in detecting T18 abnormalities, Probit regression offers more intuitive marginal effect interpretation, while the adaptive norm robust model demonstrates significant advantages in detecting T13 abnormalities. Further feature importance analysis reveals that GC content plays a dominant role in identifying all types of abnormalities, while factors such as read length, proportion, and maternal BMI also exhibit certain discriminative power across different abnormality types.

In summary, the multi-model comparative framework established in this study not only enhances the accuracy and robustness of female fetal chromosomal anomaly detection but also improves result interpretability, providing new methodological references for the clinical application of non-invasive prenatal testing. Future work may involve validation using larger-scale, multi-center datasets while exploring the integration of deep learning with probabilistic regression models to improve diagnostic capabilities in complex scenarios and provide more precise support for individualized pregnancy management.

Acknowledgements

The authors gratefully acknowledge the financial support from 2025 Xinjiang Agricultural University Student Entrepreneurship Programs (dxscy2025043, dxscy2025047).

References

[1] Zhu Q, Wang J, Xu X, Zhou S, Liao Z, Zhang J, Kong L, Liang B, Cheng X. KF-NIPT: K-mer and fetal fraction-based estimation of chromosomal anomaly from NIPT data[J]. BMC Bioinformatics, 2025, 26(1): 15.

[2] Staniczek J , Manasar-Dyrbu M , Sodowska P ,et al.Fetal karyotyping in adolescent pregnancies: a population-based cohort study on outcomes of invasive prenatal testing[J].Frontiers in Genetics, 2025, 16(16):1581249.DOI:10.3389/fgene.2025.1581249.

[3] Zhang Y, Xu H, Zhang W, Liu K, et al. Non-invasive prenatal testing for the detection of trisomy 13, 18, and 21 and sex chromosome aneuploidies in 68,763 cases[J]. Front Genet, 2022, 13: 857421.

[4] Wei R , Hu H , Wang S ,et al.Analysis of chromosomal aberrations in early pregnancy loss using high-throughput ligation-dependent probe amplification and single tandem repeats[J].Molecular Cytogenetics, 2025, 18.DOI:10.1186/s13039-025-00724-5.

[5] Ye Y , Ma J , Zhan Q ,et al.Controlled Ovarian Stimulation Contributes to the Incidence of de Novo Chromosomal Abnormalities in Cleavage-Stage Embryos[J].Archives of Medical Research, 2025, 57(3).DOI:10.1016/j.arcmed.2025.103318.

[6] Yanchun Z, Hongyan X, Wen Z, Kaibo L, et al. A machine learning technology to improve the risk of non-invasive prenatal tests[J]. BMC Med Genomics, 2023, 16(1): 54.

[7] Nie J, Wang Y, Li F, Chen X, et al. Development and Validation of a Deep Learning Model to Screen for Trisomy 21 During the First Trimester From Nuchal Ultrasonographic Images[J]. JAMA Netw Open, 2022, 5(12): e2248487.

[8] Lee S, Park J, Kim H, et al. Accuracy of cell-free fetal DNA in detecting chromosomal anomalies in women experiencing miscarriage: systematic review and meta-analysis[J]. Ultrasound Obstet Gynecol, 2024, 63(5): 678-687.

[9] Al-Ghaili A M, et al. Artificial intelligence for prenatal chromosome analysis: a systematic review[J]. Comput Biol Med, 2023, 158: 106860.

[10] Chen Y, Wang H, Li J, et al. Application of intelligent algorithms in Down syndrome screening

ISSN 2616-5805 Vol. 6, Issue 2: 121-129, DOI: 10.25236/AJMS.2025.060216

during second trimester pregnancy[J]. BMC Pregnancy Childbirth, 2023, 23(1): 115.
[11] Smith R, Johnson T, Liu H, et al. Prenatal Screening for Chromosomal Defects[v1][J]. Preprints, 2025: 2025030671.
[12] Zhao X, Chen L, et al. Novel method of real-time PCR-based screening for common fetal trisomies[J]. BMC Med Genomics, 2021, 14(1): 36.
[13] Feng C, Liu W. Scalable Bayesian p-generalized probit and logistic regression[J]. Adv Data Anal Classif, 2024, 18: 317-334.
[14] Hossain M, Rahman M, et al. Robust adaptive Lasso in high-dimensional logistic regression[J]. Stat Methods Med Res, 2021, 30(9): 2019-2034.
[15] Li X, Zhou Y, Zhang H, et al. A novel framework for abnormal risk classification over fetal nuchal translucency using adaptive stochastic gradient descent algorithm[J]. Diagnostics, 2022, 12(3): 682.