A Model for NIPT Timing Selection and Foetal Abnormality Detection Based on Multi-Objective Optimisation and Logistic Regression

Haohui Wang[1], Xinrui Wang[2], Zhen Xing[2]

[1]Reading Academy, Nanjing University of Information Science and Technology, Nanjing, China

[2]College of Ecology and Applied Meteorology, Nanjing University of Information Science and Technology, Nanjing, China

Keywords: Non-Invasive Prenatal Testing; Multi-Objective Optimization; Multiple Linear Regression; Body Mass Index, Logistic Regression

Abstract: In today's era of rapid advancements in medical technology, non-invasive prenatal testing (NIPT) enables the detection of foetal abnormalities. The timing of such testing is crucial in mitigating risks associated with the narrowing treatment window. This study first investigates the relationship between Y chromosome concentration and both gestational age and maternal body mass index (BMI). Employing both multiple linear regression and polynomial regression models, it further establishes the functional relationship between Y chromosome concentration and these maternal parameters, calculating mean squared errors of 0.0978 and 0.0987 respectively for the two models. Building upon this, an optimised model was established incorporating five additional indicators: height, weight, age, detection error, and the proportion of Y chromosome concentrations meeting the standard. Subsequently, a Monte Carlo method was employed to introduce random perturbations of 0.5%–2% to the Y chromosome concentration. Results demonstrated that the three groups defined in this study achieved accuracy rates exceeding 90% under various perturbation levels. Finally, a logistic regression model calculated the regression coefficients and p-values for each feature, yielding a ranked importance order. This enabled the extraction of coefficients for the five features to determine the formula for the classification method.

1. Introduction

Since China established its family planning policy as a fundamental national strategy in 1982, the concept of eugenics and optimal childbirth has gradually taken root in public consciousness. However, due to the impact of natural and social environmental changes, the rate of foetal defects has continued to rise. Statistics indicate approximately 900,000 new cases of birth defects occur annually in China [1]. Therefore, determining foetal health and identifying compromised pregnancies at the earliest stage is crucial, as delay risks shortening the therapeutic window. Early and standardised prenatal diagnosis plays a vital role in effectively preventing the occurrence of congenital anomalies. Among these, the primary foetal abnormalities include Down syndrome, Edwards syndrome, and Patau syndrome. These three conditions are determined by whether the proportion of foetal chromosome 21, 18, and 13 ‘free DNA fragments’ (referred to as ‘chromosome concentration’) is abnormal. The accuracy of this method is determined by the concentration of foetal sex chromosomes, which in turn correlates with other physiological indicators. Studying these indicators to understand variations in sex chromosome concentration enables the determination of optimal timing for NIPT and more precise assessment of foetal abnormalities.

2. Model Assumptions

Assumption 1: The error between the predicted values and actual values of the multiple linear regression model follows a normal distribution;

Assumption 2: The BMI in the appendix has been accurately calculated;

Assumption 3: Gestational age is calculated based on the date of last menstrual period and the time of examination, and the examination data for gestational age in the known dataset is accurate and

  DOI: 10.25236/iwmecs.2025.012

error-free.

3. Model Formulation and Solution

3.1 Data Preprocessing and Analysis

First, scatter plots were plotted separately for the relationship between Y chromosome concentration and both gestational age and maternal body mass index (BMI) to preliminarily assess their correlation. The results are shown in Figures 1 and 2:

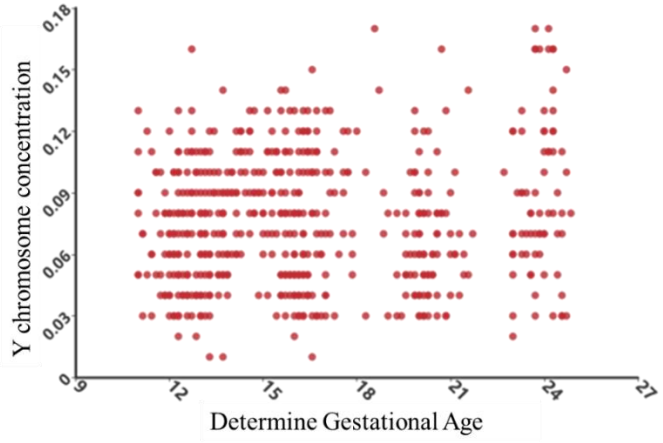

Fig. 1 Scatter Plot of Fetal Y Chromosome Concentration Versus Gestational Age at Testing

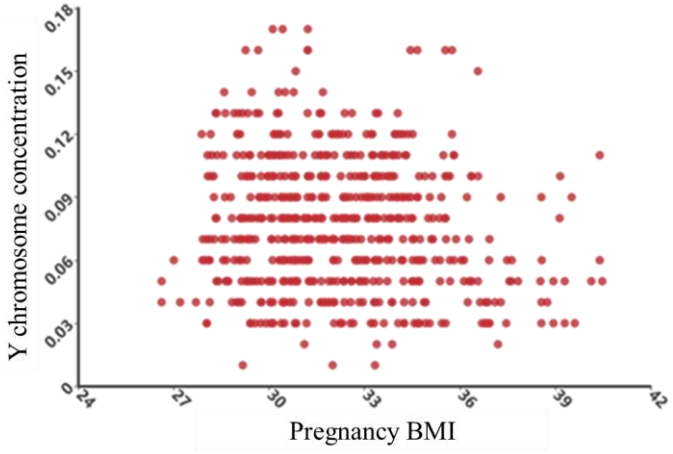

Fig. 2 Scatter plot of foetal Y chromosome concentration versus maternal BMI

As observed in Figure 1, Y chromosome concentration exhibits a weak trend of increasing with gestational age. However, sampling issues result in pronounced segmentation within the data, indicating significant challenges in constructing a model correlating gestational period with Y chromosome concentration. In contrast, a more pronounced relationship exists between maternal BMI and Y chromosome concentration. Figure 2 reveals that Y chromosome concentration exhibits a tendency to decrease with increasing BMI. When maternal BMI is at lower levels, the distribution of Y chromosome concentrations, though dispersed, maintains a higher average concentration. As BMI increases, a downward trend emerges. However, sampling limitations result in considerable dispersion in Y-chromosome concentration, particularly with fewer samples in the higher BMI range, indicating a complex relationship between BMI and Y-chromosome concentration. Consequently, a scientific approach is required to analyse pairwise correlations.

Subsequently, Spearman's correlation analysis was performed on the raw data. First, the two datasets under analysis underwent rank transformation. Taking ‘maternal BMI’ and ‘foetal Y

chromosome concentration' as examples, the raw data for each variable were sorted in ascending order, with each data point assigned a rank. Where identical values occurred, the average rank was taken. Subsequently, the sum of squared rank differences was computed. Let R_i denote the rank of 'maternal BMI' and S_i denote the rank of Y chromosome concentration. For each sample, rank interpolation was performed as follows:

$$d_i = R_i - S_i \tag{1}$$

Substitute into the formula to calculate the Spearman correlation coefficient, the formula for which is given in [2]:

$$r = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{2}$$

Here, r denotes the sample Spearman correlation coefficient, and n represents the total number of samples. This formula enables the determination of the strength of correlation between maternal BMI and Y chromosome concentration.

When multiple identical values occur within the variables—for instance, where several samples share the same maternal BMI index—the formula requires adjustment to eliminate the influence of identical ranks. The corrected formula is:

$$r = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{[\sum_{i=1}^{n}(R_i - \overline{R})^2 - \frac{\sum t_x^3 - \sum t_x}{12}][\sum_{i=1}^{n}(S_i - \overline{S})^2 - \frac{\sum t_y^3 - \sum t_y}{12}]}} \tag{3}$$

Significance testing determines whether r are statistically significant. When sample sizes are large, the sample Spearman correlation coefficient r may be approximated as following a normal distribution, allowing calculation of its test statistic:

$$Z = r\sqrt{n-1} \tag{4}$$

The value can be calculated based on the test statistic:

$$p = 2 \times (1 - \phi(|Z|)) \tag{5}$$

At a significance level of 0.05, compare the magnitude of with 0.05 to determine whether a significant correlation exists. The final results are presented in Table 1.

Table 1. Spearman's correlation coefficient and p-value

Correlation coefficient	p-value
Concentration and gestational age: 0.0681	0.0927
Concentration and BMI: -0.1309	0.0012

3.2 Multi-objective Optimisation Model

Prior to establishing a multi-objective optimisation model, pregnant women's BMI values were categorised. Given the significant weak negative correlation between Y chromosome concentration and maternal BMI, this study first employed the widely used K-means clustering method [3] for analysis.

K-means clustering is an unsupervised learning technique that divides unlabelled data into groups sharing common characteristics. Its fundamental approach involves initially guessing the number of categories, then randomly selecting several cluster centres. Subsequently, all data points are assigned to their nearest centre, forming preliminary groupings. The centres within each group are recalculated,

and this process is iterated until the groupings stabilise. Ultimately, similar data points are grouped together, aiding in the discovery of inherent structures and patterns within the data. The K-means clustering results are presented in Figure 3:

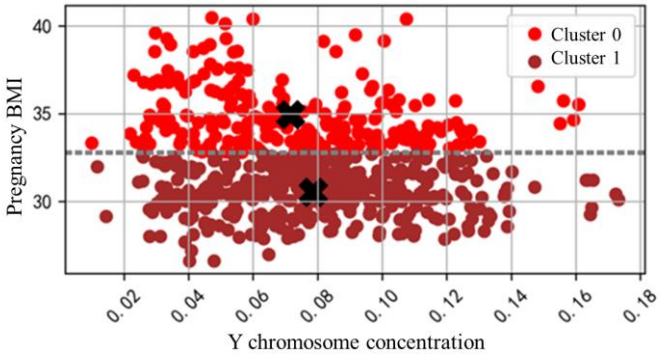

Fig. 3 K-means clustering results

Additionally, the classification performance of the model was analysed using the contour coefficient, CH index, and DBI index, yielding the results for each metric as shown in Table 2:

Table 2. K-means clustering evaluation metric values

Coefficient of Contour	CH	DBI
0.5869	1207.33	0.5697

The K-means clustering method was employed to classify pregnant women's BMI values, initially dividing them into two categories. However, considering that a mere two-category division struggles to adequately reflect the heterogeneity within the pregnant population, the classification results proved rather crude. This inadequacy hinders the subsequent determination of the optimal timing for NIPT testing. Such oversimplified grouping may compromise the scientific rigour and individualisation of testing timing, failing to meet the refined management requirements of actual clinical practice. Consequently, it is imperative to refine the classification strategy to enhance its practical significance.

To ensure sufficient data volume within each BMI group, interval-based average sample sizes were employed for grouping, yielding five data sets: [26,30), [30,31.18), [31.18,32.57), [32.57,34.32), and >34.32.

We now proceed to establish a multi-objective optimisation model [4], whose objective functions are: (1) maximising the probability of Y-chromosome concentration exceeding 0.04; (2) minimising the gestational age at testing. This ensures NIPT accuracy while minimising risks associated with shortened therapeutic windows.

To maximise the probability of Y chromosome concentration exceeding 0.04, the objective function is designed as the sum of achievement probabilities across each BMI group:

$$\max F_1 = \frac{1}{G}\sum_{g=1}^{G} P_g(t_g) \tag{6}$$

Here, G denotes the total number of BMI groups, with the subscript g corresponding to each BMI group; t represents the testing time point; $P_g(t_g)$ denotes the probability that the Y chromosome concentration in the g group of pregnant women is greater than or equal to 4% at time point t_g;

To minimise the gestational age required for examination and detection, the objective function is designed as follows:

$$\min F_2 = \frac{1}{G}\sum_{g=1}^{G} \frac{t_g - 10}{15} \tag{7}$$

This formula normalises the detection time point, converting it to a value between 0 and 1 to

eliminate dimensional effects.

As the aforementioned process requires balancing two objective functions, this paper opts to assign corresponding weights to them, thereby converting multi-objective optimisation into single-objective optimisation. The objective function is established as follows:

$$\max F = \omega_1 \cdot \frac{1}{G}\sum_{g=1}^{G} P_g(t_g) - \omega_2 \cdot \frac{1}{G}\sum_{g=1}^{G} \frac{t_g - 10}{15} \tag{8}$$

The relative importance of the two objectives can be set by adjusting the weights. Based on clinical experience, ensuring test accuracy is more critical during earlier gestational weeks, at which point the F1 weight is set to a higher value. During later gestational weeks, testing should be conducted as early as possible to mitigate risks, at which point the F2 weight is set to a higher value. To meet these requirements, dynamically varying weights are set according to different testing time points. The model testing time point must fall within 10-25 weeks: $10 \le t \le 25$.

The probability that Y chromosome concentration exceeds 0.04 in each group of pregnant women undergoing testing must not be excessively low:

$$0.8 \le P_g(t_g) \tag{9}$$

At this point, the optimised model expression can be obtained:

$$\begin{cases} \max F = \omega_1 \cdot \dfrac{1}{G}\displaystyle\sum_{g=1}^{G} P_g(t_g) - \omega_2 \cdot \dfrac{1}{G}\sum_{g=1}^{G} \dfrac{t_g - 10}{15} \\ s.t. \begin{cases} 10 \le t \le 25 \\ 0.8 \le P_g(t_g) \end{cases} \end{cases} \tag{10}$$

3.3 Multiple Linear Regression Model

To comprehensively address the issue of determining foetal abnormalities, this paper extends the aforementioned multi-objective optimisation model by incorporating additional influencing factors: height, weight, and age. It simultaneously accounts for detection errors and the proportion of foetuses meeting the Y chromosome concentration threshold. No processing is required for height, weight, and age data. The proportion of foetuses meeting the Y chromosome concentration threshold y_i is defined as:

$$a_{ij} = \begin{cases} 0, x_{ij} < 0.04 \\ 1, x_{ij} \ge 0.04 \end{cases}, i \in [10, 25],\ j \in [1, n] \tag{11}$$

$$y_i = \frac{\sum_{j=1}^{n} a_{ij}}{n} \tag{12}$$

Where x_{ij} represents the Y chromosome concentration in the foetus of the jth pregnant woman during the i-th pregnancy period, and y_i denotes the ratio of the number of pregnant women with foetal Y chromosome concentrations not below 0.04 during the i-th pregnancy week to the total number of pregnant women during that pregnancy period.

Regarding detection errors, given that GC content serves as a crucial indicator for sequencing quality assessment, its abnormal levels may impact foetal Y chromosome concentration, ultimately leading to detection errors. In the first question, a significant proportion of data (accounting for 41.68% of the total) exhibited GC content values at the margins of the normal range. When constructing the multiple linear regression model, data falling outside this range were directly excluded. This approach

is unreasonable. This question further examines the impact of detection errors arising from this factor. To quantify this effect, a random error is introduced $\varepsilon' \sim N(0,\sigma^2)$. To reflect the influence of GC content on this error, an error weighting factor w_i is defined:

$$a_i = \begin{cases} 0, x_i \geq 0.4 \\ x_i, x_i < 0.4 \end{cases}, i \in [1,n] \tag{13}$$

$$w_i = 100 \times |a_i - 0.4| \tag{14}$$

where x_i represents the GC content for each pregnant woman, and denotes the sample size. Finally, construct a new column to detect data errors:

$$y_i = w_i \varepsilon', i \in [1,n] \tag{15}$$

The general form of a multiple linear regression model is:

$$\hat{y} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \alpha_7 x_7 + \varepsilon \tag{16}$$

Where $\hat{y}$ is the Y chromosome concentration, x_1 denotes the gestational age at testing, x_2 represents the mother's BMI, x_3 indicates age, x_4 denotes height, x_5 denotes weight, x_6 denotes the proportion of foetuses meeting the Y chromosome concentration threshold, x_7 denotes the detection error, α_0 is the constant term, $\partial_1 \sim \partial_7$ denotes the coefficient for each variable, and $\varepsilon \sim N(0,\sigma^2)$ is the error term.

Perform parameter estimation using the method of least squares:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{17}$$

where y_i denotes the actual observed value and $\hat{y}_i$ denotes the predicted value.

$$\begin{aligned} y = & -0.4208 + 0.0009x_1 + 0.0061x_2 - 0.0008x_3 \\ & + 0.0028x_4 - 0.0030x_5 + 0.1379x_6 - 0.0155x_7 \end{aligned} \tag{18}$$

K-means clustering was performed on height, weight, age, examination error, and the proportion meeting Y-chromosome concentration standards. The optimal partitioning scheme was selected by plotting an elbow plot and calculating the sum of squares within clusters (SSE), contour coefficient, CH index, and DBI index. The elbow plot is depicted in Figure 4:

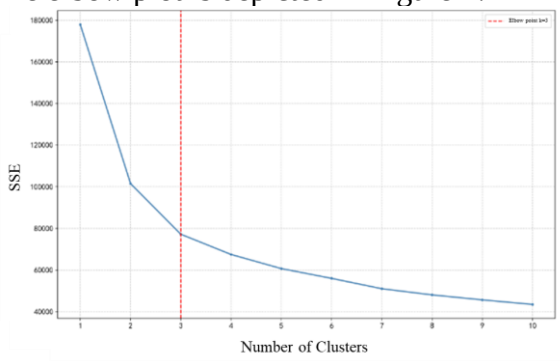

Fig. 4 k-means clustering elbow diagram

As shown in Figure 4, the intra-cluster sum of squares decreases sharply between cluster numbers 1 and 5, while the rate of decline slows for cluster numbers exceeding 5. An excessively low number

of clusters fails to meet the precision requirements of medical research, whereas an excessively high number significantly increases the complexity of clinical implementation. Consequently, relevant metrics were calculated for 3 to 6 clusters to facilitate further selection. The computational results are presented in Table 3:

Table 3. Cluster-specific correlation metrics

Cluster number K	Coefficient of Contour	CH Index	DBI Index
3	0.2787	703.9638	1.1885
4	0.2518	588.1545	1.2164
5	0.2427	520.0769	1.3413
6	0.2028	468.7164	1.3751

The data in the summary table indicates that classification performance at k=3 is marginally superior to other cluster numbers. Employing the optimal classification cluster number of 3, the probability function is calculated using a multiple linear regression model incorporating additional factors. Subsequently, the K-means classification results are substituted into the objective function to determine the optimal NIPT timing for each group. The computational results are presented in Table 4:

Table 4. Three Scheme comparing

Croup Num	Optimal timing (weeks)
1	16.0
2	14.7
3	15.2

As shown in Table 4, the optimal detection time point occurs earlier in Group 2 (low height and low weight with low BMI), whereas it occurs later in Group 1 (tall height and large weight with high BMI).

3.4 Logistic Regression Model

Considering the chromosomal differences between female and male foetuses, this paper independently explores the influence of multiple factors on foetal abnormalities in females, providing a method for determining such abnormalities. This constitutes a binary classification problem. Given that logistic regression is a widely applied statistical learning method for classification tasks, particularly suited to binary classification, this approach is employed to establish the analytical model.

The core principle of logistic regression involves mapping the output of linear regression onto the interval (0,1) via a sigmoid function, thereby representing the probability of a sample belonging to a particular class. When this probability is greater than or equal to a predetermined threshold (typically 0.5), the model predicts the sample belongs to the positive class; otherwise, it belongs to the negative class. The specific form of the sigmoid function is:

$$\begin{cases} \sigma(z) = \dfrac{1}{1+e^{-z}} \\ z = w^T x_i + b \end{cases} \tag{19}$$

Among these, z is the result of a linear combination; w represents the weight vector, indicating the importance of each feature; x_i denotes the input feature vector; b is the bias term.

Logistic regression solves model parameters through maximum likelihood estimation, thereby deriving the likelihood function. This function serves as the starting point for the objective function when deducing optimal parameters. The likelihood function identifies a set of parameters that maximises the probability of the observed outcomes. The form of the likelihood function [5] is:

$$L(\mathbf{w},b) = \prod_{i=1}^{N} P\left(y_i = 1 \mid \mathbf{x}_i\right)^{y_i} \cdot \left(1 - P\left(y_i = 1 \mid \mathbf{x}_i\right)\right)^{1-y_i} \tag{20}$$

Here, L denotes the probability of observing the current data given the model parameters w and b; N denotes the total sample size;

As logical regression lacks an analytical solution, optimal parameters cannot be directly determined. Consequently, iterative optimisation algorithms are required. The gradient descent method is thus employed, which calculates the gradient of the loss function with respect to the model parameters to determine the direction of parameter updates—namely, adjusting parameters in the opposite direction of the gradient to progressively reduce the loss. This process enables the model's output probability distribution to approximate the true distribution. At each iteration, gradient descent adjusts the weights based on the prediction error under the current parameters, ultimately converging upon the parameter values that minimise the loss. This achieves an effective solution for the model parameters.

By solving the logistic regression model, the regression coefficients and corresponding p-values for the 20 feature variables were obtained. Selected results are presented in Table 5:

Table 5. Feature variable regression coefficients and p-values

Feature Name	Coefficient of regression	p-value
Age	-0.2224	0.0007
Height	0.0192	0.9753
Weight	0.0069	0.9909
Maternal BMI	-0.0729	0.9636
Z-score for chromosome 21	0.4096	0.0483
GC content for chromosome 18	629.1031	0.1457
Percentage of filtered read segments	33.6380	0.2563

The regression coefficient reflects the extent to which a feature influences classification outcomes; a larger absolute value indicates a stronger impact on predictive results. Furthermore, if the p-value falls below the significance threshold (e.g., 0.05), the feature is deemed to exert a statistically significant effect on classification results. The results indicate the six most influential features for predicting foetal abnormalities are: GC content on chromosome 21, GC content on chromosome 18, GC content on chromosome 13, GC content, proportion of alignment on the reference genome, and X chromosome concentration. Their corresponding regression coefficients are -1006.266, 629.103, 576.524, -236.889, -73.561, and -61.348, with corresponding p-values of 0.00017, 0.14569, 0.06277, 0.05683, 0.36793, and 0.00007. According to medical knowledge, GC content on different chromosomes reflects their specific genetic composition and structural characteristics. Alterations in GC content may affect processes such as DNA unwinding and transcription, leading to abnormal gene expression and triggering foetal developmental abnormalities. Therefore, the model results appear reasonable.

Observation of the above results reveals instances where regression coefficients exhibit large absolute values alongside large p-values. This is a common and reasonable occurrence. It typically indicates that the feature may exhibit a strong tendency to influence outcomes, but due to factors such as small sample size, multicollinearity, insufficient variance, or noise interference, the current data fails to provide sufficient statistical evidence to confirm its significance.

Following the results, this paper conducted an effectiveness analysis of the logistic regression model, calculating the likelihood ratio chi-squared value, p-value, and classification accuracy, and plotting the ROC curve as shown in Figure 5:

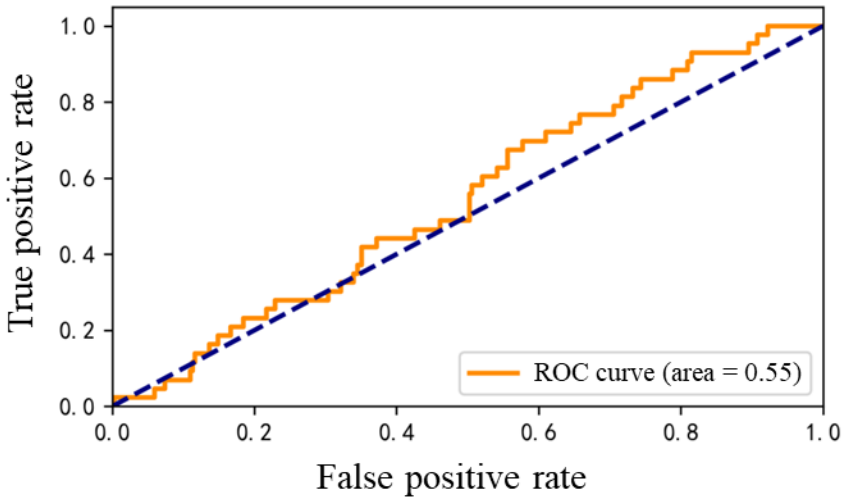

Fig. 5 ROC curve diagram

The likelihood ratio chi-squared value obtained is 107.035, with a p-value of 6.799×10^{-14}. This indicates that the logistic regression model is significantly effective overall, demonstrating a robust ability to explain the relationship between the independent variables and the dependent variable "chromosomal aneuploidy". The calculated classification accuracy reached 0.887, indicating strong predictive capability for sample classification. The model accurately identifies chromosomal aneuploidy in most cases. However, the ROC curve area under the curve (AUC) of 0.55, as shown in the figure, is only marginally above 0.5. This suggests room for improvement in distinguishing positive from negative samples, with potential suboptimal prediction for certain cases. In summary, the logistic regression model is generally effective with decent classification accuracy, though room for optimisation remains in precisely distinguishing sample categories.

Five of the top six most influential features identified by both the logistic regression and random forest models overlap: GC content of chromosome 13, GC content of chromosome 18, GC content of chromosome 21, GC content, and X chromosome concentration. Consequently, a method for determining female foetal abnormalities was established based on the metric values of these five features. By re-training the logistic regression model to learn the relationship between features and the dependent variable, the coefficients and intercept for these five features were extracted. This ultimately yielded the following classification formula:

$$\begin{aligned} \log it(p) = &-20.5021 - 25.4383x_1 - 73.2661x_2 \\ &+33.0427x_3 + 23.5054x_4 + 14.7007x_5 \end{aligned} \tag{21}$$

If $\log it(p) \geq 0$, it is determined to be an abnormal female foetus; if $\log it(p) < 0$, it is determined to be a normal female foetus.

4. Summary

This study presents both qualitative and quantitative analyses of the relationship between foetal Y chromosome concentration and maternal BMI alongside gestational age at detection, establishing a multiple linear regression model. By converting multi-objective optimisation into single-objective optimisation through assigning weights according to target importance, the complexity of the model was reduced. Whilst exploring multiple characteristic indicators affecting female foetal abnormalities and determining diagnostic methods, the logistic regression model—a widely applied statistical learning method for classification problems, particularly suited to binary classification tasks—demonstrated high compatibility with Problem 4, exhibiting favourable model adaptability. However, certain limitations persist, such as suboptimal model fit due to insufficient consideration of influencing factors. This may stem from the limited number of adopted factors, potentially obscuring

numerous latent relationships. Further research will address this issue in subsequent stages.

References

[1] Huang Xiu-jie, Li Jing. Analysis of the Current Status and Influencing Factors of Birth Defects in Newborns [J]. Medical Frontiers, 2025, 15(19): 34-37.

[2] Ding Le, Han Feng, Wang Jia-qi, et al. Improvement of High-G Acceleration Rapid Load Reduction Analysis and Control Strategy Based on Spearman's Correlation Coefficient [J]. Energy Engineering, 2025, 45(03): 16-20.

[3] He Pailing, Cheng Meiyue, Li Dekun, et al. Optimisation of Logistics Distribution Routes Using a Two-Stage Fusion Model Based on K-means Algorithm and Genetic Algorithm [J]. Modern Information Technology, 2025, 9(15): 168-173+178.

[4] Feng Daguang, Feng Sizhe, Li Zhaoxing, et al. Comparison of Solar Greenhouse Temperature Prediction Methods Based on Multiple Linear Regression Models [J]. Journal of Shenyang Normal University (Natural Science Edition), 2025, 43(01): 75-81.

[5] Jiang Jun, Feng Shuo, Sun Yinggui, et al. Hypothermia Risk Prediction Model Following Transurethral Holmium Laser Enucleation of the Prostate: Based on Logistic Regression, Decision Trees and Support Vector Machines [J/OL]. Journal of Southern Medical University, 1-7 [2025-09-07].