A Foetal Abnormality Detection Model Based on Cox Proportional Hazards and Multi-Layer Perceptron Neural Networks

Yu'ang Zhou[1], Xiaolin Liu[1], Jiayi He[1]

[1]Silesian College of Intelligent Science and Engineering, Yanshan University, Qinhuangdao, China

Keywords: Cox proportional hazards model; k-means clustering; MLP; Neural networks

Abstract: With the widespread adoption of NIPT prenatal testing technology for assessing foetal health, unhealthy foetuses can be detected more rapidly, thereby extending the therapeutic window. However, NIPT can only guarantee its validity when performed under specific conditions. This paper therefore commences by applying k-means clustering to relevant data using the Cox proportional hazards model, thereby determining the optimal testing timeframe based on existing data. An error analysis of various influencing factors revealed that height exhibits the lowest sensitivity. Considering the differences between male and female chromosomes, this study constructed features from foetal data and developed a neural network model based on a multi-layer perceptron (MLP) for training and learning. This model achieved a sensitivity of 91%, specificity of 82%, and accuracy of 86%.

1. Introduction

Non-invasive prenatal testing (NIPT) [1] is a prenatal screening technique that analyses foetal chromosomal abnormalities by collecting and examining foetal cell-free DNA fragments present in the maternal bloodstream during pregnancy. This technology is primarily employed for the early identification of common chromosomal aneuploidy disorders such as Down syndrome (Trisomy 21), Edwards syndrome (Trisomy 18), and Patau syndrome (Trisomy 13). The accuracy of NIPT is highly dependent on the concentration of foetal sex chromosomes (XY for male foetuses, XX for female foetuses). Generally, testing can be conducted between 10 and 25 weeks of gestation. Results are considered substantially reliable when the concentration of the male Y chromosome reaches or exceeds 4% and the female X chromosome concentration is within normal limits.

In practice, the concentration of the male Y chromosome correlates closely with the gestational age and the mother's Body Mass Index (BMI). To optimise detection outcomes, pregnant women are typically grouped according to BMI, with optimal testing windows determined for each group. Given individual variations in maternal age, BMI, and pregnancy status, a uniform testing strategy or single empirical grouping may cause some women to miss the optimal treatment window, thereby increasing potential health risks. Consequently, scientifically sound grouping and timing selection are crucial for enhancing the accuracy and safety of NIPT [2][3].

2. Model Assumptions

Assuming that the concentration of foetal-derived cell-free DNA generally increases progressively with gestational age, with inter-individual variation in the rate of increase but no decline during the mid-to-late pregnancy period. Therefore, gestational age may be regarded as the primary temporal driver of concentration variation.

Assuming that maternal physical indicators such as BMI, weight, and height significantly influence the timing of concentration attainment. Higher BMI tends to delay concentration attainment, with consistent overall trends despite individual variations in the magnitude of this effect.

Assuming that fluctuations in experimental conditions such as sequencing depth, alignment coverage, and GC content may be regarded as systematic variations. These do not introduce directional bias to the overall concentration metric. Consequently, modelling may incorporate quality factors to correct for such variations.

Assuming that random errors during detection are independent of core variables such as gestational

DOI: 10.25236/iwmecs.2025.017

age and BMI, and collectively follow a zero-mean disturbance distribution. This assumption ensures residuals serve as a reasonable characterisation of uncertainty.

3. Model Formulation and Solution

3.1 Data Preprocessing

To ensure the representativeness of the samples and the comprehensiveness of the sample data, we found that deleting data with GC content below 40% would result in significant loss of raw data, potentially leading to model overfitting and compromising the model's statistical reliability. Therefore, we only excluded data with GC content at or below 38%.

Additionally, gestational age was converted to a standardised week count expressed as a decimal. The formula employed was: Gestational Age (weeks) = Integer Weeks + 7/Number of Days, thereby completing the preliminary preprocessing.

In modelling and analysing non-invasive prenatal testing (NIPT) data, data quality directly impacts model validity and the reliability of conclusions. Based on clinical practice and study requirements, we defined three screening criteria to form a composite decision function: Let the sample's feature vector be x = (g, y, c), where g denotes the detected gestational age, y represents the chromosomal concentration, and c indicates the GC content. Construct the screening function:

$$valid\left(x\right) = (10 \le g \le 25) \Lambda (y \ge 4) \Lambda (0.4 \le C \le 0.6) \tag{1}$$

The effective sample set is defined as:

$$D_{valid} = \{x \in DIvalid\left(x\right) = true\} \tag{2}$$

We then proceed with sequence merging: for the raw dataset D, each record r contains the pregnant woman code p(r) and the number of blood draws t(r). Records pertaining to the same pregnant woman and the same blood draw are grouped together, denoted as:

$$G_{p,t} = \{r \in \mathrm{D} | \, p\left(r\right) = p, \; t\left(r\right) = t\} \tag{3}$$

Here, p denotes the set of all pregnant women's codes, and T denotes the set of all blood sampling instances.

Extract non-null 'chromosomal aneuploidy' results from the grouping (excluding invalid records) to form a valid set:

$$A_{p,t} = \{a\left(r\right) \big| \, r \in Gp,t,a\left(r\right) \ne null\} \tag{4}$$

The retention of each set is determined based on its frequency within the valid set, after which the retained data undergoes cleansing. The results are shown in Figure 1.

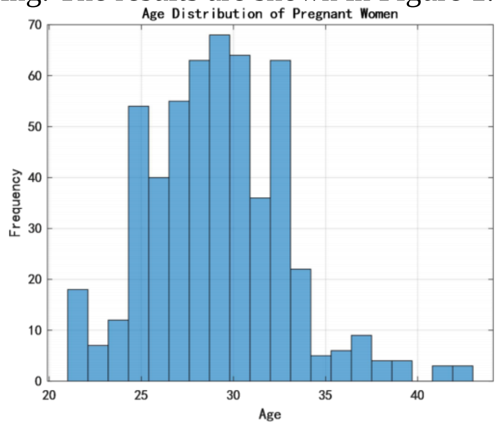

Fig. 1 Frequency distribution chart of maternal age

3.2 Spearman's Correlation Analysis

Compared to Pearson correlation, Spearman's rank correlation coefficient (py(s,t)) robustly measures the strength of this monotonic association without requiring linearity or homogeneity of variance. Evidently, the relationship between Y chromosome concentration and gestational age or BMI does not exhibit linear progression; hence we employ Spearman's rank correlation coefficient [4].

The raw data for variable X and variable Y are each independently sorted in ascending order. The smallest value is assigned rank 1, the second smallest rank 2, and so on. Where multiple values are identical, the average rank of their positions is taken. Following this transformation, the ranks Rxi for variable X and Ryi for variable Y are obtained. To prevent tied ranks, the following formula is employed for calculation.

$$r_s = \frac{\sum_{i=1}^{n}\left(R_{xi}-\bar{R}_x\right)\left(R_{yi}-\bar{R}_y\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{xi}-\bar{R}_x\right)^2\sum_{i=1}^{n}\left(R_{yi}-\bar{R}_y\right)^2}} \tag{5}$$

Where $\bar{R}_x$ is the mean of variable X's level, $\bar{R}_y$ is the mean of variable Y's level.

The Spearman heatmap obtained through the aforementioned calculations is shown in Figure 2.

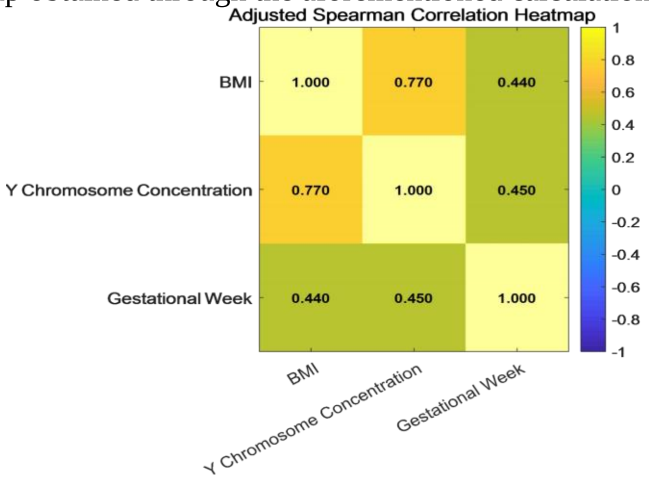

Fig. 2 Spearman Correlation Coefficient Heatmap

As can be observed from Figure 2, BMI exhibits a relatively strong correlation of 0.78 with Y chromosome concentration, indicating a pronounced synchronous variation between maternal BMI and foetal Y chromosome concentration in maternal blood. The correlation coefficient between gestational age and Y chromosome concentration is 0.46. As gestational age increases, there is a noticeable upward trend in foetal Y chromosome concentration in maternal blood. Furthermore, the relationship between BMI and gestational age, as calculated through partial correlation, also demonstrates an association. This aligns with the information provided in the title.

3.3 Cox Proportional Hazards Model

To investigate the relationship between the time at which male foetuses achieve the target Y chromosome concentration (≥4%) and multiple factors including BMI, height, weight, and age, we defined the gestational week at which the Y chromosome concentration first reached or exceeded 4% as the target attainment time. Using the ‘proportion of target attainments’ across all tests for that individual as the attainment ratio, we derived the following formula.

$$\begin{aligned} T_i^t &= \min\left\{t_{ij} : v_i\left(t_{ij}\right) \geq 0.04\right\} \\ r_i &= \frac{1}{m_i}\sum\nolimits_{j=1}^{m_i} 1\left(v_i\left(t_{ij}\right) \geq 0.04\right) \end{aligned} \tag{6}$$

Where $v_i(t)$ denotes the concentration of Y at gestational week t for sample i; t_{ij} represents the gestational week of the jth test, with a total of m_i tests; T_i^t indicates the earliest time to meet standards; $r_i \in [0, 1]$ signifies the proportion of sample i meeting standards.

The time to achieve compliance in NIPT is influenced by multiple factors, including the mother's physical constitution and obstetric history. Centred on BMI, additional features such as age, height, weight, GC content, number of pregnancies, number of deliveries, number of blood draws for testing, and chromosomal non-integer copy numbers are incorporated to form a feature vector. Random forest regression is employed to determine the importance of each feature, enabling the selection of the top key factors for subsequent clustering.

$$x_i = (\mathrm{BMI}_i, Age_i, Height_i, Weight_i, GC_i, PregCnt_i, BirthCnt_i, r_i)^T, S \\ = \{\mathrm{BMI}\} \cup \arg\max_{A \subset F, |A|=4} \sum_{j \in A} \phi_j \tag{7}$$

Where x_i denotes the multi-factor vector for sample i; F represents the complete set of candidate features; ϕ_j indicates the importance of feature j as determined by the random forest; S constitutes the feature subset used for grouping (comprising BMI and the top four features selected by importance).

To eliminate the effects of dimensions and scales, each feature used for clustering is standardised to a mean of zero and a unit variance. Subsequent clustering and distance measurement are then performed within this space.

The Cox model is employed to model the hazard function for the occurrence of ‘achievement events’:

$$h(t|x) = h_0(t)\ \exp(\beta_1 BMI + \beta_2 Age + \beta_3 Weight + \beta_4 Height + \beta_5 nbd + \beta_6 ca)$$

Among these, $h(t|x)$ denotes the hazard function at time T under covariate X; $h_0(t)$ represents the baseline hazard function; β_i signifies the covariate coefficient; nbd denotes the number of blood draws tested; and ca indicates chromosomal aneuploidy.

The Cox model estimates parameters by maximising the likelihood function:

$$L(\beta) = \prod_{i:\text{it happens}} \frac{\exp\left(\mathrm{X}_{i/\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\mathrm{X}_{j/\beta}\right)} \tag{8}$$

The results of the Cox model are as Table 1:

Table 1. Cox Model Table

Variable	Coef	HR=exp (coef)	p-value
S_Age	0.12	1.13	0.45
S_Height	-0.08	0.92	0.62
S_Weight	0.05	1.05	0.78
S_Number of blood tests conducted	-0.21	0.81	0.09
S_Maternal BMI	0.18	1.20	0.12
S_Chromosomal aneuploidy	0.35	1.42	0.04

In Cox models, HR > 1 indicates that the variable increases the risk of meeting the target, i.e., earlier attainment; whereas HR < 1 signifies that the variable reduces the risk of meeting the target, i.e., later attainment. A p-value less than 0.05 indicates a significant effect on either early or late attainment. As shown in the figure, S_chromosome aneuploidy exhibits HR = 1.42 and p = 0.04, significantly increasing the risk of achieving the target. Consequently, we exclude this indicator. Regarding S_testing blood draw frequency: HR = 0.81, p = 0.09, which may delay target attainment but fails to reach statistical significance. For model accuracy, we exclude this variable. S_maternal BMI: HR = 1.20, p = 0.12, may promote earlier attainment but did not reach statistical significance.

By analysing the total squared error (SSE) corresponding to different cluster numbers (K) using the CH-Index method and plotting a trend graph, we obtained the results shown in Figure 3:

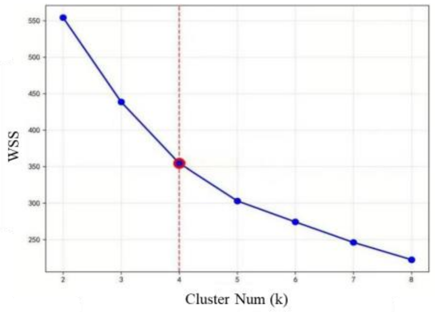

Fig. 3 SSE versus Number of Clusters Line Chart

After comprehensive analysis, we selected K=4 as the optimal number of clusters.

We employed K-means clustering. Within the standardised multifactorial space, K-means minimised the sum of squares within clusters, yielding candidate multifactorial groupings dominated by BMI.

$$\min_{\{C_g, c_g\}_{g=1}^{k}} \sum_{g=1}^{k} \sum_{i \in C_g} \left\| z_i - c_g \right\|^2 \tag{9}$$

Here, cg denotes the sample set of the gth cluster; cg represents the corresponding cluster centre; zi = (zij)j∈s denotes the standardised feature vector of sample i. The resulting clustered data is as Table 2:

Table 2. Final Cluster Centroid Results Table

Variable	Cluster 1	Cluseter 2	Cluster 3	Cluster 4
S_Age	-0.3960	-0.0606	0.0738	1.5664
S_Height	0.2970	0.5740	-1.2485	0.3806
S_Weight	0.1547	2.5020	-1.1063	0.2539
S_Maternal BMI	-0.0054	2.7596	-0.5516	0.0649

Risk Function Construction:

$$R(t) = \alpha P\left(y < 0.04 \middle| t\right) + \beta(t - t_0)$$

Through this function, we unify two risks that cannot be directly compared within a single quantitative framework using weighting coefficients α and β. The former represents the risk of premature detection expressed as a probability cost, while the latter denotes the risk of overdetection expressed as a time cost. Overall, the smaller the value of R(t), the lower the overall risk of conducting detection at time t.

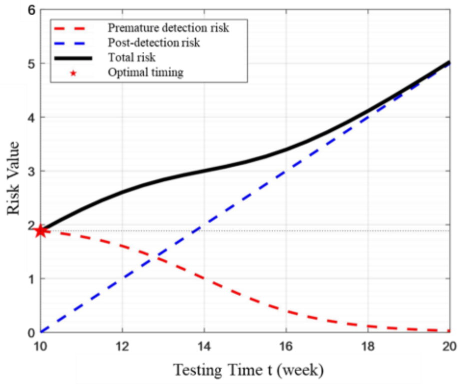

Fig. 4 Detection Time Risk Analysis Chart

As shown in Figure 4, the optimal detection time point is 10.0 weeks, with a minimum risk value of 1.885. The risk of premature detection at the optimal time point is 1.885, while the risk of delayed detection at the optimal time point is 0.000.

We have established an intra-group optimal timing as a fallback option. The recommended intra-group timing is defined as the point where ≥95% of pregnant women meet the criteria. Based on the scatter plot distribution generated by MATLAB, we can determine:

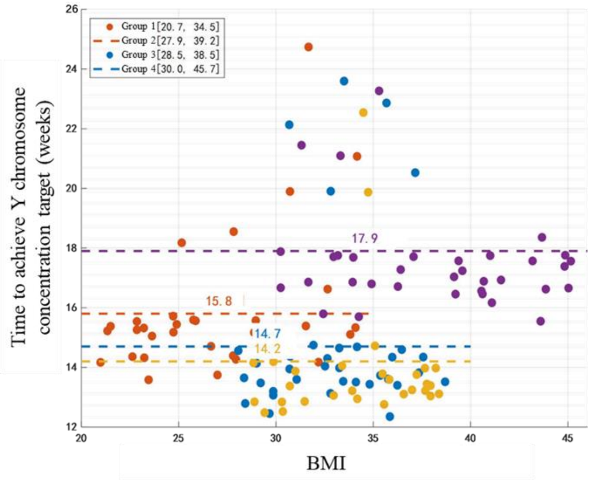

Fig. 5 Multifactorial BMI Grouping and Optimal Detection Timing

As shown in Figure 5, the intra-group recommended timing for the first group was 14.2 weeks, for the second group 14.7 weeks, for the third group 15.8 weeks, and for the fourth group 17.9 weeks. This indicates minimal deviation from our risk function-derived data, with only the first group exhibiting a two-week discrepancy.

3.4 MLP Neural Network Model

Our study exclusively examines female foetal samples, with the objective of determining whether any of chromosomes 13, 18, or 21 exhibit aneuploidy (abnormality). Neural networks are employed for this prediction. The supervised labels are defined as yi ∈ {0,1}, indicating whether the ith sample is abnormal (1 = abnormal; 0 = normal). We stipulate that yi=1 {if ABi is non-empty or positive}.

Let n denote the sample size, d the input feature dimension, and xi ∈R^d the feature vector for the ith sample.

Our primary diagnostic metrics are Z13, Z18, Z21, and Z_X. Required sequencing quality indicators include: N (total sequencing reads), map (proportion of reads aligned to the reference genome), dup (proportion of duplicate reads), uniq (number of uniquely mapped reads), filt (proportion of filtered reads), P (GC content), GC13, GC18, GC21. Clinically, these correspond to the indicators BMI, gestational age (GA), age, and IVF.

Feature construction in this paper primarily involves standardisation and the creation of derived variables. First, the GC deviation is calculated using the following formula:

$$\mathrm{GC}_{dev,i} = \left|\mathrm{GC13}-\mathrm{P}\right| + \left|\mathrm{GC18}-\mathrm{P}\right| + \left|\mathrm{GC21}-\mathrm{P}\right| \tag{10}$$

Next, we seek efficiency:

$$\mathrm{eff}_i = \mathrm{uniq} \times \mathrm{map} \tag{11}$$

Calculate the noise index:

$$\mathrm{noise}_i = \alpha_1(1-\mathrm{map}) + \alpha_2 \mathrm{dup} + \alpha_3 \mathrm{filt} + \alpha_4 (1-\mathrm{uniq}) \tag{12}$$

Then proceed with read-out scaling:

$$\log \mathrm{N}_i = \log\left(\mathrm{N}_i + 1\right) \tag{13}$$

Final input vector: xi=[Z13, Z18, Z21, Z_X, GCdev, logN, eff, noise, BMI, GA, Age, IVF pregnancy method]. To enable more precise analysis, we standardised continuous variables using Z-scores.

The neural network we constructed operates as follows: first, forward propagation occurs, with the first hidden layer formula being:

$$h_1 = \operatorname{ReLU}\left(w_1 x' + b_1\right) \tag{14}$$

The formula for the second hidden layer is:

$$h_2 = \operatorname{ReLU}\left(w_2 h_1 + b_2\right) \tag{15}$$

The formula for the output layer is:

$$p = \sigma\left(w_3 h_2 + b_3\right) \tag{16}$$

Here, w_1, w_2, w_3 is the weight matrix, b_1, b_2, b_3 are the bias terms, ReLU is the activation function, and σ denotes the sigmoid function.

The expression for the ReLU activation function is ReLU(x) = max(0, x), while the expression for the sigmoid function is [5]:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{17}$$

The model's loss function is:

$$L = -\sum_i \left[w_1 y_i \log\left(p_i\right) + w_0\left(1 - p_i\right)\right] + \lambda \sum \|w\|^2 \tag{18}$$

Here, yi denotes the true label, pi represents the predicted output, and λ is the regularisation term, whose purpose is to prevent overfitting. The model results are shown in Figure 6:

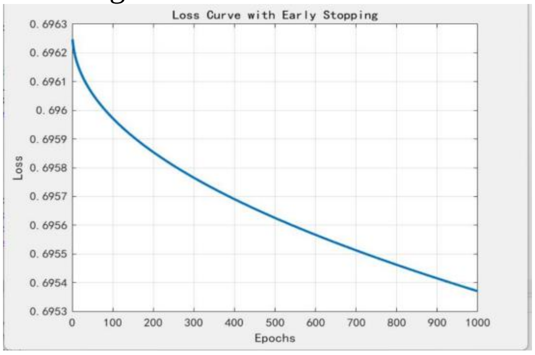

Fig. 6 Loss Function Variation Curve

As shown in Figure 6, the loss value is somewhat high, indicating suboptimal learning performance of the model. Here, we introduce Bayesian optimisation for analysis. Within neural networks, Bayesian optimisation (BO) serves as an efficient black-box optimisation method. Its core objective is to leverage historical experimental data through Bayes' theorem to precisely guide parameter selection for subsequent experiments in scenarios where evaluation costs are prohibitively high, ultimately enabling rapid identification of the global optimum. It is primarily employed to tackle challenges such as hyperparameter optimisation (HPO) and neural architecture search (NAS) in neural networks, demonstrating significant superiority over traditional grid search and random search approaches. Bayesian optimisation comprises two core components: the probabilistic surrogate model and the acquisition function. The former primarily utilises Gaussian processes as surrogate models to approximate the objective function:

$$f(x) \sim \operatorname{Gp}\left(m\left(x\right), m\left(x, x'\right)\right) \tag{19}$$

Then, we proceed to the posterior distribution and subsequently employ the expected improvement as the acquisition function. The formula for expected improvement is as follows:

$$\alpha_{\mathrm{EI}}\left(x\right) = \mathrm{E}\left[\max\left(0, f\left(x\right) - f\left(x^{+}\right)\right)\right] \tag{20}$$

Moreover, we recognise that the outputs of numerous machine learning classification models—such as Support Vector Machines (SVM), boosting methods, and neural networks—do not represent true probabilities, but rather resemble a “confidence score” or “rating”. These scores may tend to cluster within a specific range, necessitating their mapping to genuine probabilistic meaning.

Platt scaling is a remarkably simple yet effective probability calibration method. It employs a logistic regression model to learn how to transform raw prediction scores s into calibrated probabilities p. Its model is a parameterised sigmoid function:

$$P\left(y=1\middle|s\right)=\sigma\left(As+B\right)=\frac{1}{1+\exp\left(As+B\right)} \tag{21}$$

The results of the model run are shown in Figures 7 and 8:

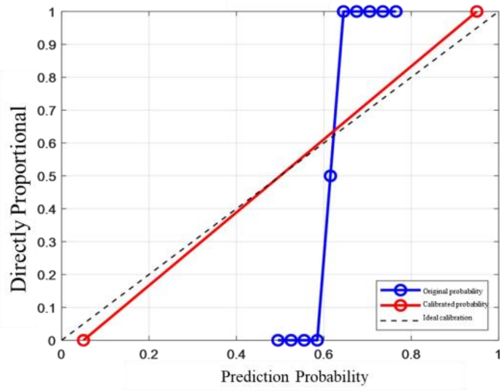

Fig. 7 Probability Calibration Curve Diagram

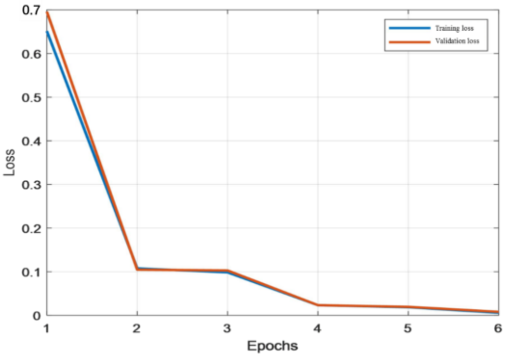

Fig. 8 Training and validation loss

It is evident that the loss function value has been significantly reduced, with the model exhibiting a downward trend. This indicates that the model has successfully minimised training error during the learning process and effectively minimised the objective function, demonstrating a favourable learning trajectory. Through Platt Scaling probability calibration, we obtained calibrated probabilities and found the calibration to be quite satisfactory. Following calibration, we derived p^.

If p^ ≥ π_hi, we can infer female foetal abnormality.

If π_lo ≤ p^ < π_hi, we infer the female foetus is in the grey zone—that is, unable to be definitively categorised as normal or abnormal.

If p^ < π_lo → we can infer the female foetus is normal.

The threshold is determined by minimising the cost function R=c_FN ·FN+c_FP ·FP whilst constraining the grey zone proportion.

In the minimised cost function:

N (False Negative): the number of samples that are truly abnormal but predicted as normal by the model. FP (False Positive): the number of samples that are truly normal but predicted as abnormal by the model. c_FN and c_FP represent the cost incurred by each missed diagnosis or misdiagnosis. Therefore, we seek an optimal solution to ensure no abnormal female foetus is overlooked. This also indirectly validates the correctness of our model.

We address this using MATLAB: For each pair (π_lo, π_hi), classify the p^ value of every sample in the validation set to obtain a decision outcome (normal, grey zone, abnormal). We then compute the confusion matrix: disregarding all samples classified as grey zone (as they require further examination and are not direct consequences of this decision), focusing solely on those with explicit classifications. Through MATLAB calculations, we approximate the optimal solution, yielding the following results: the optimal lower threshold π_lo = 0.01, the optimal upper threshold π_hi = 0.02, corresponding to a minimum cost R = 0.00 and a grey zone proportion = 0.00%. These findings are illustrated in Figure 9.

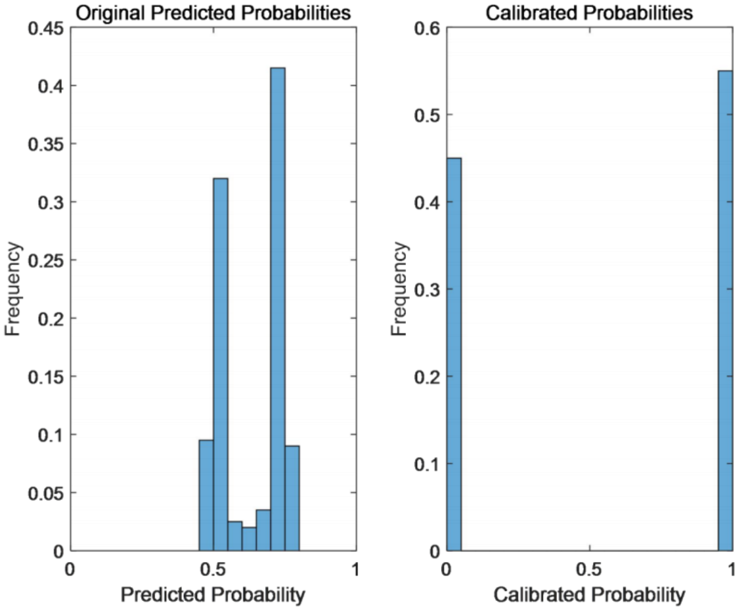

Fig. 9 Comparison of Predicted Probability and Calibrated Probability

The model's evaluation metrics are as follows:

Sensitivity: $Recall = TP/(TP+FN)$

Specificity: $Specificity = TN/(TN+FP)$

Accuracy: $Accuracy = (TP+TN)/(TP+FP+TN+FN)$

Calculated values: Sensitivity 91%, Specificity 82%, Accuracy 86%.

4. Summary

This paper constructs an integrated framework encompassing ‘correlation mining—component modelling—timing optimisation—anomaly detection’, covering the entire process from modelling male foetal concentration to identifying female foetal anomalies. The optimisation of detection timing incorporates risk functions and Monte Carlo simulations, thereby avoiding the premature/delayed detection bias inherent in reliance solely on theoretical thresholds. Through Spearman correlation analysis, feature importance ranking, and risk stratification, the model outcomes are rendered more clinically interpretable and actionable. By integrating multidimensional features including BMI, GC content, gestational age, and target achievement rates, the approach better reflects real-world clinical scenarios than single-factor BMI grouping. However, the present model exhibits certain limitations. For instance, the 4% threshold for achieving target levels fails to account for variations across sequencing platforms or geographical regions, thereby restricting its adaptability. Furthermore, K-means clustering exhibits sensitivity to initial values, posing a risk of boundary drift that may diminish clinical interpretability.

References

[1] Zhang Peng, Mo Weiying, Meng Minghui, et al. Impact of non-invasive prenatal testing on detection of sex chromosome aneuploidy and related ethical considerations [J]. Chinese Journal of

Clinical Medicine, 2025, 18(06): 690-695.

[2] Zhang Y. C., Zhang W., Liu K. B., et al. Analysis of prenatal screening and diagnosis for children with trisomy 21 syndrome [J]. Laboratory Medicine and Clinical Practice, 2025, 22(19): 2716-2720.

[3] Cui Ping, Liu Naiguo. Value of CNV-seq Combined with Karyotyping in Prenatal Diagnosis for High-Risk Populations in NIPT [J]. China Medical Guide, 2024, 22(22): 58-61.

[4] He Guoqing, Li Weile, Lu Huiyan, et al. Automatic Identification of Landslide-Prone InSAR Deformation Zones Based on Spearman Filtering and HNSW-DBSCAN Clustering [J]. Journal of Wuhan University (Information Science Edition), 2025, 50(08): 1682-1693.

[5] Wang Zicheng. Modelling tropospheric delay based on segmented and GA-MLP neural network optimisation [D]. Guilin University of Technology, 2025.