Research on a Personalised NIPT Screening Strategy Based on Multi-factor Collaborative Optimisation and a Multi-level Bayesian Network

Huan Li[1], Zhiguo Hu[1], Chenyang Hu[1], Xiang Li[1], Mingming Gong[2]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China;
[2]iFLYTEK Co., Ltd., Hefei, Anhui, China

Abstract: This paper proposes an integrated optimisation framework based on hierarchical decision trees and multi-layer Bayesian diagnosis. Firstly, a B-spline regression model is employed to analyse the non-linear interaction between foetal cell-free DNA concentration, gestational age and BMI, and robust BMI stratification is achieved via a hierarchical Bayesian model. Secondly, an extended hierarchical decision tree model (HDT-Extension) is established to dynamically adjust the testing thresholds for different subgroups by incorporating factors such as maternal age and body constitution. Finally, with a focus on abnormal conditions in female foetuses, a dual-model collaborative algorithm combining Multi-Hierarchical Bayesian Diagnosis (MHBD) and Feature Engineering Integrated Decision Tree (FEIDT) is developed for comprehensive evaluation. Experiments reveal that the optimal testing timing differs significantly across various BMI subgroups (P<0.0001), with the optimal timing set at 14.3 gestational weeks for the low-BMI cohort and 17.8 gestational weeks for the obese cohort; following strategic optimisation, the overall qualification rate rises from 86.6% to 93.8%. Furthermore, the complementary effect of the two models strikes a balance between high recall and high specificity. The proposed framework effectively reduces the risk of missed diagnoses in high-BMI populations and provides intelligent decision-making support for personalised prenatal screening.

Keywords: Non-Invasive Prenatal Testing (NIPT); Hierarchical Decision Tree; Bayesian Hierarchical Model; Dynamic Threshold; Ensemble Learning; Medical Decision Support System.

1. Introduction

Non-invasive prenatal testing (NIPT) analyses cell-free foetal DNA (cffDNA) in the peripheral plasma of pregnant women via high-throughput sequencing and is currently the primary method for screening for foetal chromosomal aneuploidy [1-5]. The accuracy of this test is highly dependent on the effective concentration of cffDNA (clinically, a concentration of ≥4% is typically required). However, the increase in maternal blood volume and changes in fat metabolism caused by a high body mass index (BMI) can result in a significant 'dilution effect' on foetal DNA [6-8] .Currently, a fixed testing window of 12–22 weeks is widely adopted in clinical practice, overlooking individual heterogeneity such as maternal age and body type. This 'one-size-fits-all' static strategy results in higher rates of testing failure among groups with high BMI, creating an urgent need to establish a dynamic optimisation mechanism for individualised testing timings.

Furthermore, for female foetuses lacking the Y chromosome as a concentration marker, the determination of abnormalities relies heavily on standardised Z-scores derived from autosomes; this metric is highly susceptible to interference from sequencing noise and GC content bias [9]. Furthermore, the proportion of genuinely clinically abnormal samples is extremely low, exhibiting typical characteristics of class imbalance [10,11]. Traditional classification rules based on a single fixed threshold have limited generalisation capabilities, making it difficult to meet the dual clinical requirements of high sensitivity in initial screening and high specificity in confirmatory testing.

2. Related Work and Research Framework

Existing research on NIPT data analysis has largely focused on linear statistical association analyses or conventional machine

learning models. These methods struggle to accurately capture the complex non-linear interactive effects between multidimensional physiological factors; furthermore, when dealing with highly imbalanced data on foetal abnormalities in female foetuses, they lack a composite decision-making mechanism that balances clinical interpretability with high accuracy [9-12].

To overcome these technical bottlenecks, this study proposes a framework for intelligent NIPT decision-making and personalised strategy optimisation based on multidimensional feature fusion. The core research framework comprises three progressive modules: firstly, B-spline regression is introduced to precisely analyse the non-linear coupling relationship between cffDNA concentration, gestational age and BMI, whilst a Bayesian hierarchical model is utilised to achieve robust population risk stratification;secondly, a Hierarchical Decision Tree Extension (HDT-Extension) model is constructed, integrating features such as age and body type to dynamically optimise the optimal testing time for different BMI subgroups; Finally, addressing the data imbalance associated with female foetal abnormalities, we innovatively propose a dual-model collaborative algorithm combining Multi-layer Bayesian Diagnosis (MHBD) and Feature Engineering Integrated Decision Tree (FEIDT). The MHBD focuses on probabilistic inference to ensure high recall, whilst the FEIDT relies on ensemble learning to enhance specificity. Through the interdisciplinary integration of statistical inference and machine learning, this study where p denotes the number of features, X_{il} denotes the sequencing quality control features selected via Lasso regularisation (LassoCV), andϵ_i denotes the random error.

provides reliable methodological support for the development of precise antenatal screening protocols in clinical practice.

3. Research Methods

3.1 B-spline Non-Linear Regression Model for Cell-Free DNA Concentration

B-spline regression is a non-parametric modelling method based on piecewise polynomials, capable of flexibly and smoothly fitting complex, high-dimensional non-linear curves through local support characteristics [13]. In NIPT testing, the concentration of cell-free foetal DNA (cffDNA) is subject to a non-linear interaction between gestational age and body mass index (BMI).Given that the cffDNA concentration Y_i is ratio data defined on the interval $[0,1]$, and to avoid out-of-bounds predictions in conventional linear regression, this study first applies an arcsine square root transformation to it:

$$Y_i^* = \arcsin(\sqrt{Y_i}) \quad (1)$$

To finely characterise the non-linear effects of the covariates, B-spline basis functions were introduced. Let$f(W_i)$ and$g(B_i)$ denote the main effect functions for gestational week and BMI, respectively; the synergistic interaction between the two was captured using tensor product splines $h(W_i, B_i)$ to construct a comprehensive non-linear regression model:

$$Y_i^* = \alpha + f(W_i) + g(B_i) + h(W_i, B_i) + \sum_{l=1}^{p} \delta_l X_{il} + \epsilon_i \quad (2)$$

3.2 Robust Population Grouping Based on Bayesian Hierarchical Inference

The Bayesian hierarchical model, by constructing a probabilistic directed acyclic graph (DAG) of data, parameters and hyperparameters, is able to fully utilise prior information and effectively quantify the statistical uncertainty of multi-level data [14]. Having established the significant inhibitory effect of BMI on cffDNA, it is necessary to stratify the pregnant women's population according to risk. However, traditional rigid threshold-based grouping often leads to the misclassification of borderline samples. To this end, this study constructed a Bayesian hierarchical model to infer the optimal detection time boundaries for different BMI groups. Let the observed detection time for the kth group be t_{ki} ; its probabilistic graph structure is defined as follows:

$$t_{ki}|\mu_k, \sigma_k^2 \sim N(\mu_k, \sigma_k^2) \quad (3)$$

$$\mu_k|\mu_0, \tau^2 \sim N(\mu_0, \tau^2) \quad (4)$$

The model employs the non-parametric Bootstrap resampling method to approximate the posterior distribution of the mean parameter $p(\mu_k|t_k)$. Based on the posterior inference results, robust quartile risk boundaries are extracted, strictly dividing the sample into four risk categories (low BMI, normal, overweight and obese groups), thereby providing a reliable foundation for subsequent stratification algorithms.

3.3 Multifactorial Hierarchical Decision Tree

Extension Model (HDT-Extension)

As a classic top-down supervised learning algorithm, the decision tree possesses a strong ability to capture feature interactions and offers interpretability of clinical rules [15] .Building upon the Bayesian BMI primary grouping, and to address the impact of complex heterogeneity in clinical practice—such as the age and body shape distribution of pregnant women (e.g. abdominal obesity)—this study proposes and designs the Hierarchical Decision Tree Extension Model (HDT-Extension). This model dynamically and adaptively adjusts the optimal screening time point thresholds for each subgroup via multi-level decision nodes.For the i-th individual, the decision function for their personalised optimal screening time point$T_{optimal}(i)$ is defined as:

$$T_{optimal}(i)=T_{base}(i)+\alpha_{age}\cdot Age(i)+\beta_{body}\cdot Body(i) \quad (5)$$

where $T_{base}(i)$ is the baseline threshold for the BMI group to which the individual belongs;α_{age} is the discretised age correction coefficient; andβ_{body} is the continuous body type correction coefficient. To solve for the optimal decision parameters$\theta=[\alpha_{age},\beta_{body}]$, a multi-objective loss function $J(\theta)$ was constructed, incorporating screening compliance rate, screening timeliness and false-negative penalty.The model employs Projected Gradient Descent with Momentum for optimisation, with the following iterative update where the prior probability `$P(Abnormal)$` is derived from the pregnant woman's BMI risk classification; the likelihood function `$P(D|Abnormal)$` is constructed based on the distribution of Z-scores for autosomal DNA. Additionally, a sequencing quality correction factor and an individual BMI penalty term are introduced to output the final calibrated confidence level for abnormalities.

(2) FEIDT ensemble learning model (focusing on high specificity)

After constructing derived features such as the Z-score ratio and quality-weighted density, the model combines Random Forest (to reduce variance) with AdaBoost (to reduce bias). The RF component outputs via Bootstrap aggregation:

rule for parameters:

$$\theta_{t+1}=\Pi_{\Theta}\left[\theta_t-\eta\nabla J(\theta_t)+\mu(\theta_t-\theta_{t-1})\right] \quad (6)$$

where η is the learning rate,Π_{Θ} is the projection onto the clinically safe gestational age range, and μ is the momentum coefficient.

3.4 MHBD and FEIDT Dual-Model Joint Diagnostic Algorithm

Faced with the class imbalance and high-dimensional noisy data typical of medical settings, a single classifier often encounters severe performance bottlenecks; however, a dual architecture combining probabilistic graphical models and ensemble learning is an effective paradigm for overcoming this bottleneck [15,16] .To address the technical challenges posed by the lack of Y-chromosome references for female foetal samples and the extreme imbalance in abnormal data, this study has designed a dual-model collaborative architecture comprising Multi-Hierarchical Bayesian Diagnosis (MHBD) and Feature Engineering Integrated Decision Trees (FEIDT).

(1) MHBD probabilistic inference model (focusing on high recall)

The model calculates the posterior probability of an abnormal condition based on Bayes' theorem:

$$P(Abnormal|D)=\frac{P(D|Abnormal)P(Abnormal)}{P(D)} \quad (7)$$

$$P_{RF}=\frac{1}{N_{trees}}\sum_{t=1}^{N_{trees}}h_t(x) \quad (8)$$

The component enhances the discrimination of a small number of anomalous samples by dynamically adjusting the weights of misclassified samples $H(x)=sign(\sum_{t=1}^{T}\alpha_t h_t(x))$. Finally, the system outputs decision results via a weighted fusion mechanism$P_{ensemble}=w_{RF}\cdot P_{RF}+w_{AB}\cdot P_{AB}$, achieving complementary advantages in sensitivity and specificity.

4. Data Description and Processing

4.1 Data Description

This study utilised real-world clinical NIPT data provided by a medical institution. Given the fundamental differences in the mechanisms of chromosomal abnormality detection between male and female foetuses, the original dataset was divided into two independently analysed subsets:

(1) Male foetal sample subset (1,082 cases): As the Y-chromosome concentration in male foetuses is a direct biomarker for assessing test accuracy (clinically, an effective concentration of ≥4% is typically required) [1,2], this study first screened this subset for core mechanism analysis, non-linear modelling and dynamic optimisation of individualised testing

timepoints.

(2) Female foetal sample subset (over 2,000 cases): Owing to the lack of a Y chromosome reference, determining abnormalities is highly challenging; this subset was primarily used for the construction and validation of combined diagnostic models for complex chromosomal abnormalities. It is worth noting that this cohort exhibits typical characteristics of class imbalance, with normal samples accounting for approximately 98 per cent, whilst abnormal samples affected by trisomy 13, 18 and 21 account for only about 2 per cent [2,3] .

The data characteristics encompass multi-source information covering the entire process from blood collection to sequencing, primarily comprising the following three dimensions:

(1) Basic information on pregnant women: including age, height, weight, gestational age at testing (primarily concentrated within the clinically recommended window of 12–20 weeks), body mass index (BMI, covering the full spectrum from underweight to severely obese), and other physiological indicators.

(2) Technical parameters: These include sequencing-related and quality control parameters such as the number of raw sequencing reads, GC content (consistently distributed within the range of 0.38–0.42), mapping rate and duplication rate.

(3) Chromosomal characteristics: These include information such as the concentration ratios and standardised scores (Z-scores) for each autosome (with a focus on chromosomes 13, 18 and 21) and the sex chromosomes (X and Y).

4.2 Data Pre-processing

Given the inevitable presence of noise or distortion in real-world clinical data, this study carried out rigorous data cleaning and feature transformation prior to the core algorithmic process:

Outlier removal and data cleaning: Based on clinical medical common sense, samples with physiological indicators exceeding reasonable physiological thresholds (such as extremely abnormal records for gestational age or BMI) and those with severe deficiencies in key sequencing quality control indicators were excluded to ensure the robustness of subsequent statistical inference and model training.

Transformation of the Target Variable's Constrained Domain: Given that the regression prediction target (Y-chromosome concentration) in the male foetal subset is a proportional variable strictly constrained within the interval $[0,1]$, the direct application of ordinary non-linear regression is prone to causing out-of-bounds prediction errors at the boundaries. To address this, this study applied an arcsine square root transformation to the response variable prior to modelling, smoothly mapping it to the real number domain, thereby effectively enhancing the reliability of predictions at the boundaries.

Feature standardisation: To eliminate differences in units of measurement between multi-dimensional, multi-source features and to accelerate the convergence rate of machine learning models (such as projection gradient descent with momentum and ensemble tree models), all continuous covariates (e.g. BMI, age, number of sequencing reads, etc.) were uniformly standardised using Z-scores.

5. Feature Engineering and Analysis of Association Mechanisms

5.1 Correlation Analysis of Core Covariates

To identify key factors influencing NIPT detection quality and elucidate their physiological mechanisms, this study first analysed the linear relationships between each raw feature and the target variable (foetal Y-chromosome concentration) using Pearson's correlation coefficient.

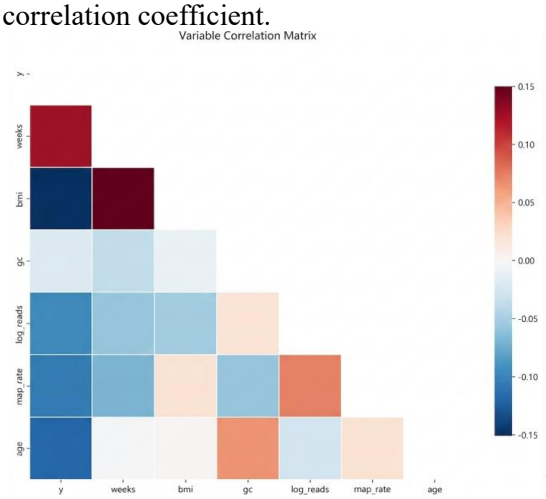

Figure 1. Correlation Heatmap between Variables

As shown in the correlation heatmap in Figure 1, gestational age was significantly positively correlated with Y-chromosome concentration ($r = 0.42$, $p < 0.001$). This indicates that, as gestational age increases, the total amount of cell-free DNA released through apoptosis of

foetal trophoblast cells tends to rise, consistent with the physiological patterns of development [6,8,17] .

Maternal BMI was significantly negatively correlated with Y-chromosome concentration ($r = -0.18$, $p < 0.01$): this provides statistical confirmation of the strong 'dilution effect' exerted on cffDNA concentration by high levels of maternal adipose tissue and increased blood volume [6,18] .

Correlations with other sequencing quality parameters were weaker: no clear direct linear relationship was observed between GC content and the number of raw reads and Y-chromosome concentration; however, as these are sequencing environmental parameters, they must still be included as control variables in modelling to eliminate systematic bias. Meanwhile, the multicollinearity between gestational age and BMI was extremely low ($r = 0.08$), allowing them to be treated as independent predictors.

5.2 Extraction and Visualisation of Non-linear Interaction Effects

A simple linear correlation analysis cannot effectively capture the complex interaction between the 'gestational age-promoting' and 'BMI-inhibiting' effects [6,7]. To further explore the underlying mechanisms, this study plotted a multidimensional scatter plot matrix of key variables.

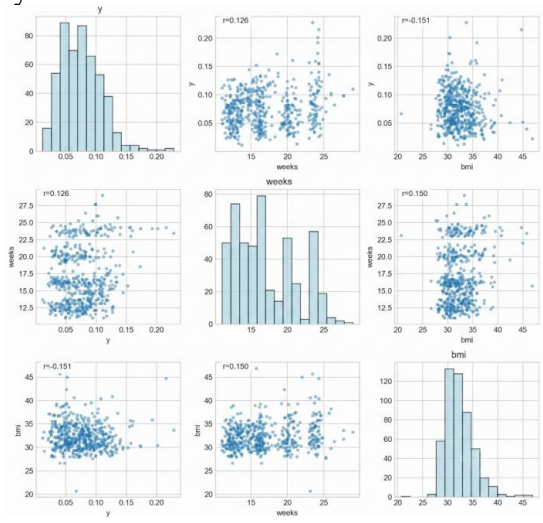

Figure2. Scatter Plot Matrix of Key Variables

As shown in Figure 2, the scatter plot matrix intuitively illustrates the underlying complex interactive characteristics:

Non-linear diminishing marginal returns: Y-chromosome concentration does not follow a simple linear pattern with increasing gestational age; rather, the rate of increase gradually slows after 16–18 weeks, exhibiting a distinct 'S-shaped' non-linear growth trend.

Spatially coupled interaction effects: The inhibitory effect of BMI on DNA concentration exhibits markedly different slopes at different gestational stages, with the two variables displaying complex non-linear spatial surface characteristics.

This visualisation suggests that traditional linear regression or simple quadratic polynomials cannot accurately capture these patterns. This provides a solid data basis for the subsequent introduction of B-spline tensor products in the methodology of this paper to extract high-dimensional non-linear interaction features.

5.3 Derived Feature Construction for Class Imbalance

Following the extraction of non-linear features from the male foetal subset, the model faced the severe challenge of class imbalance in the task of identifying chromosomal abnormalities in female foetuses, characterised by an extreme scarcity of abnormal samples (accounting for only approximately 2%).If the raw autosomal Z-values were fed directly into the classifier, the model would be highly prone to getting stuck in local optima. To address this, this study combined bioinformatics mechanisms with the manual construction of six highly discriminative higher-order derived features to amplify the decision boundary of the extremely small number of abnormal samples:

(1) Inter-chromosomal Z-score ratio feature ($R_{21/18}$,$R_{21/13}$): To eliminate the interference caused by maternal heterogeneity and local sequencing noise on the Z-score of a single chromosome, the relative concentration deviation ratio between chromosomes was constructed by introducing a smoothing factorϵ :

$$R_{21/18}=\frac{Z_{21}}{Z_{18}+\epsilon} \quad (9)$$

$$R_{21/13}=\frac{Z_{21}}{Z_{13}+\epsilon} \quad (10)$$

(2) Sequencing environment bias correction factor (F_{GC})

To quantify the extent to which GC content drift compromises diagnostic reliability, the ideal state (GC = 0.4) is mapped to 1, and a linear penalty weight is constructed:

$$F_{GC}=1-\frac{|GC-0.4|}{0.1} \quad (11)$$

(3) Sequencing quality-weighted density feature

($D_{quality}$)

A dimensionless composite of read count, alignment rate and duplication rate, used to identify poor-quality samples resulting from insufficient sequencing depth or excessive library amplification:

$$D_{quality}=reads\times map\_rate\times(1-dup\_rate) \quad (12)$$

(4) Spatially Complex Trisomy Composite Z-score (Z_{sum})

Utilising the concept of Euclidean distance, this captures combined abnormal variations in trisomies 21, 18 and 13 within a multidimensional space:

$$Z_{sum}=Z_{13}^2+Z_{18}^2+Z_{21}^2 \quad (13)$$

(5) Discrete Sequencing Read Adequacy Index (I_{reads})

Introduction of a threshold binary feature reflecting sequencing saturation:

$$I_{reads}=\begin{cases}1, & reads>4\times10^6\\ 0, & otherwise\end{cases} \quad (14)$$

Individualised BMI Risk Scaling Factor (F_{BMI})

This enables the classification model to explicitly account for the risk of missed diagnoses due to obesity by applying a range-standardised transformation to continuous BMI values:

$$F_{BMI}=\frac{BMI-18.5}{50-18.5} \quad (15)$$

Through the comprehensive feature engineering process described above—ranging from non-linear interactions to higher-order composite features—the original low-dimensional sparse information has been successfully transformed into a highly discriminative feature space, providing high-quality input for the subsequent FEIDT ensemble learning model.

6. Experiments and Results Analysis

6.1 Fitting of Cell-Free DNA Concentrations and Robust Risk Stratification

Firstly, a B-spline non-linear fit was performed on the cell-free DNA (cffDNA) concentrations of the male foetal subset [13]. The joint test for the entire model was highly significant ($F=21.589$, $P<1.89\times10^{-39}$), with both gestational age and BMI main effects passing the test ($P<0.001$).The coefficient of determination for the model on the training set was $R^2=0.2108$, and for cross-validation it was $R_{CV}^2=0.1314$. Furthermore, residual diagnostics (Shapiro-Wilk test) and Cook's distance (with a maximum value of only 0.08 across the entire sample) confirmed the statistical robustness of this regression framework.

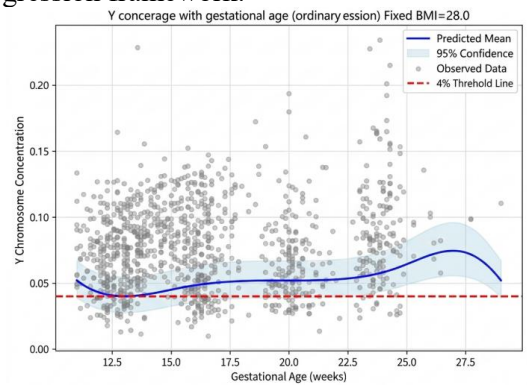

Figure 3. Predicted Relationship between Y-Chromosome Concentration and Gestational Age

As shown in Figure 3, the prediction surface indicates that cffDNA exhibits a typical 'S-shaped' marginally decreasing growth pattern with gestational age; simultaneously, concentrations in the high-BMI group (red curve) are consistently at the lowest levels, directly confirming the 'dilution effect' of body fat on cell-free DNA [6.7] .

To overcome the shortcomings of traditional 'one-size-fits-all' categorisation, a Bayesian hierarchical model was further utilised to extract the posterior distribution [14] , establishing four robust BMI quartile risk boundaries with balanced sample sizes: the low BMI group (G1), the normal group (G2), the overweight group (G3) and the obese group (G4).

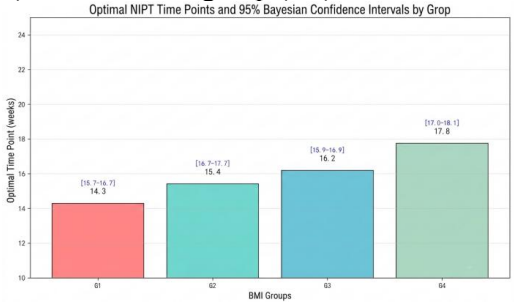

Figure 4. Confidence Interval Plot for the Optimal NIPT Timing

Combined with the Bayesian posterior estimates in Figure 4, the baseline optimal testing time points for different BMI subgroups show an ascending pattern: 14.3 weeks for group G1, 15.4 weeks for group G2, 16.2 weeks for group G3, whilst the high-risk obese group G4 requires testing to be deferred to 17.8 weeks. ANOVA analysis indicated extremely significant differences between groups ($F=11267.799$,$P<0.0001$), with Cohen's d as high as 2.749, demonstrating that this

stratification strategy possesses extremely strong discriminatory power.

6.2 Evaluation of the Optimisation Effect of Multifactorial Dynamic Time Points

Building upon the primary BMI grouping, the HDT-Extension decision tree model was introduced, utilising the projected gradient descent algorithm to comprehensively optimise the derived features of maternal age and body type. The converged multi-factor personalised optimal timing recommendation matrix is shown in Table 1 [15,16] .The optimisation results indicate that, due to their high metabolic rate, women in the young age group may undergo screening slightly earlier, whilst those in the older age group and those with abdominal obesity require an adaptive delay in the screening timing, thereby overcoming the bias associated with static, fixed screening windows.

As shown in the comparison of target attainment rates in Figure 5, following evaluation via 5-fold cross-validation, the introduction of the HDT-Extension framework resulted in a significant increase in the overall NIPT target attainment rate from the conventional baseline of 86.6% to 93.8%.Notably, the compliance rate for the obese, older-age group—which previously had the lowest detection success rate—increased steadily from 79.3% to 87.3%, reducing the overall clinical risk of missed detection by 53.7%. Further sensitivity analyses indicate that even under the most extreme simulated conditions—such as sequencing concentration errors (±0.3%) and BMI measurement bias ($\pm 1.5 kg/m^2$)—the overall compliance rate remains within the range of92.1%±1.2% , validating the algorithm's high noise-resistance and robustness.

Table1. HDT-Extension Comprehensive Multi-Factor Optimal Screening Timing Decision Matrix

BMI range	Young group (≤28 years)	Standard group (29–34 years)	Older adults (≥35)	Compliance rate following optimisation (%)
[20.7, 30.2)	19.8	20.6	21.0	100
[30.2, 31.8)	21.4	22.2	22.6	100
[31.8, 33.9)	19.2	20.0	20.4	87.8
[33.9, +∞)	22.1	22.9	23.3	87.3

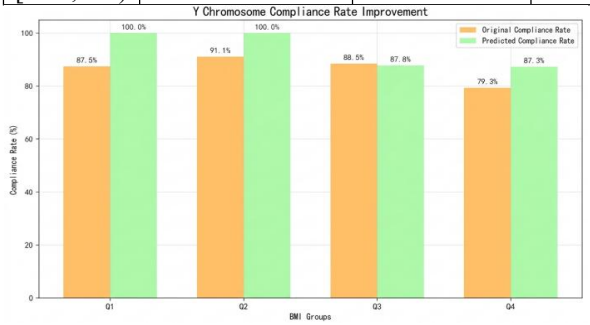

Figure 5. Comparison of Improvements in Compliance Rates

6.3 Combined Diagnosis of Female Foetal Anomalies and Clinical Application

Given the absence of the Y chromosome as a concentration marker and the extreme scarcity of chromosomal abnormality samples—where normal and abnormal cases exhibit an extremely class-imbalanced distribution of approximately 98%:2% [10,11]), this study focused on evaluating the joint diagnostic performance of the MHBD probabilistic inference and FEIDT ensemble learning models under complex real-world noise [16] .

In medical data scenarios characterised by extreme class imbalance and high-dimensional sequencing noise, traditional classification evaluation metrics (such as precision and F1 score) often suffer from severe numerical penalties due to the large number of negative cases, making it difficult to fully reflect the model's true clinical value. To address this, this study conducted an in-depth analysis by directly combining global ROC curves with confusion matrices.

Although the absolute values of individual models on the global ROC curve (Figure 6) are limited, an in-depth analysis combining the confusion matrix (Figure 7) revealed that the two models demonstrate strong complementary value in a 'primary screening–review' workflow in clinical practice:

MHBD model (clinical initial screening/recall-priority): The statistically probability-driven MHBD model demonstrates a significant advantage in sensitivity (recall), reaching 0.385.Leveraging Bayesian prior constraints and a dual calibration mechanism, this model is exceptionally adept at detecting faint variant signals in autosomal Z-scores, thereby minimising the risk of missing foetuses with potential trisomy variants; it is an ideal algorithm for the first stage of clinical screening [14] .

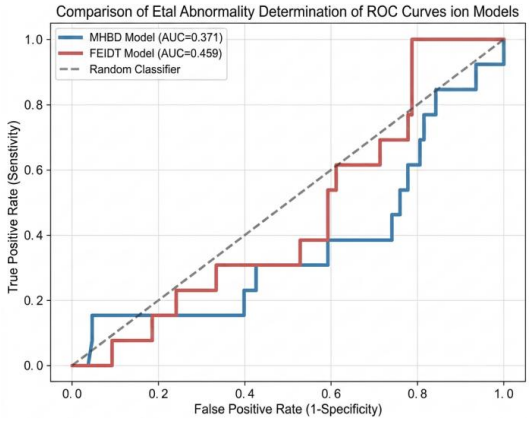

Figure 6. Comparison of ROC Curves for the Dual Models in the Diagnosis of Foetal Female Abnormalities

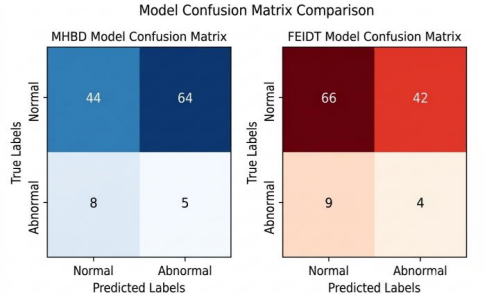

Figure 7. Comparison of Confusion Matrices for the MHBD and FEIDT Models in Classification Diagnosis

FEIDT Model (Expert Review/Specificity-Prioritised): Compared with the MHBD model, the FEIDT model—which is based on higher-order feature integration—exhibits superior specificity (reaching 0.611), with the number of false-positive (misdiagnosed) samples significantly reduced. By tracking the non-linear relationship between the spatial composite score and the GC correction factor, this model can accurately identify and eliminate false-positive abnormalities caused by maternal physiological biases or sequencing environmental noise, making it highly suitable as a second-stage expert review (Verification) algorithm.

Based on the complementary strengths and weaknesses described above, this study has deeply integrated the two approaches, overcoming the drawbacks of black-box models and providing obstetricians with a highly interpretable 'IF-THEN' clinical tiered diagnostic decision pathway [12]:

High-risk mutation alert: When the test reveals a complex spatial trisomy composite score of $Z_{sum}>4.5$, or when $Z_{sum}>3.0$ and the inter-chromosomal Z-score ratio is abnormal, the system immediately triggers a high-risk alert, prompting obstetricians to immediately perform gold-standard invasive diagnostic procedures such as amniocentesis.

Technical Re-testing: When the environmental bias correction factor is $F_{GC}<0.5$ and the sequencing density is extremely low, the system automatically determines that the sample's sequencing quality is substandard, triggering a technical re-sampling mechanism to curb misdiagnoses and missed diagnoses caused by noise at the source of the data.

7. Conclusions

This study addresses two major technical challenges faced by non-invasive prenatal testing (NIPT) in real-world clinical applications: 'difficulty in achieving effective concentrations in high-risk populations' and 'extreme imbalance in rare chromosomal abnormalities'. It constructs a multi-dimensional, data-driven intelligent decision-support framework, providing a full-chain solution for clinical prenatal screening, ranging from 'dynamic time-point optimisation' to 'complex anomaly determination'. The core contributions of this study are primarily reflected in the following three aspects:

Firstly, by deeply analysing multidimensional features, the study revealed the non-linear coupling mechanism between physiological indicators and free DNA concentration. The research confirmed that the concentration of cell-free foetal DNA (cffDNA) does not increase in a simple linear fashion with gestational age. Compared with traditional linear or quadratic polynomial regression, this study innovatively introduced B-spline tensor products to accurately characterise the complex non-linear diminishing marginal returns between 'gestational age promotion' and 'BMI suppression', thereby thoroughly analysing, from a mathematical perspective, the fundamental mechanism underlying the high incidence of failed initial screening in pregnant women with high BMI.

Secondly, breaking away from the 'one-size-fits-all' approach, the study proposes a multi-factor-driven, personalised strategy for optimising the timing of testing. Addressing the high rate of missed detection among overweight individuals caused by traditional, uniform fixed testing timepoints (such as the conventionally recommended 12 weeks), this study integrates Bayesian hierarchical inference with the

HDT-Extension hierarchical decision tree model. This model comprehensively accounts for heterogeneity in pregnant women's body type and age, dynamically generating personalised screening windows accurate to the 'week' (e.g. for severely obese individuals, the screening window is adaptively extended to 17.8 weeks or later). In backtesting, this strategy significantly increased the overall screening compliance rate from 86.6% to 93.8%, thereby substantially reducing the wastage of clinical resources.

Thirdly, addressing the challenge of class imbalance, we constructed a highly interpretable closed-loop joint diagnostic decision-making system. In the assessment of foetal chromosomal abnormalities—where the proportion of abnormal samples is extremely low—this study moved away from a single 'black-box' model and established a two-tiered collaborative diagnostic architecture combining MHBD probabilistic inference with FEIDT ensemble learning. Through a complementary mechanism whereby 'MHBD ensures high sensitivity in initial screening, whilst FEIDT ensures high specificity in review', the system generates highly interpretable IF-THEN clinical decision rules, including 'high-risk alerts' and 'technical rejections for retesting', thereby significantly reducing diagnostic uncertainty for clinicians.

In summary, this study not only quantifies the key influencing factors of NIPT testing in theory but also provides, in practice, a scientific, robust and intelligent clinical decision-support tool with a low risk of missed diagnoses. Future research will focus on the following two directions: firstly, expanding and enriching the heterogeneity of the sample repository, particularly by incorporating more complex special clinical samples such as twin or multiple pregnancies, to further calibrate population risk thresholds and validate the model's generalisability; secondly, to explore the integration of deep neural networks and multi-omics technologies, attempting to incorporate multi-source, heterogeneous data—such as the pregnant woman's medical history and deep sequencing profiles—into the feature space, thereby advancing non-invasive prenatal screening towards a more comprehensive form of precision medicine.

References

[1] Jiang L Y, Lu S K, Du J, et al. Development and Application of Non-invasive Prenatal Testing Technology. Clinical Medicine Research and Practice, 2025, 10(23): 191–194. DOI: 10.19347/j.cnki.2096-1413.202523047.

[2] Zhang M, Li Q, Xu L, et al. The clinical utility of NIPT in the screening for foetal chromosomal aneuploidy. International Journal of Laboratory Medicine, 2025, 46(23): 2896–2901.

[3] Geng Z X, He F J, Xu J J. Analysis of the Screening Value of CMA Combined with NIPT for Foetal Chromosomal Abnormalities in the Second Trimester. Chinese Journal of Family Planning, 2025, 33(02): 473–478.

[4] Ye Y, Liu C L. An Analysis of the Clinical Value of Non-invasive DNA Prenatal Testing in Pregnant Women at High Risk of Trisomy 21. Primary Care Forum, 2026, 30(10): 56–58. DOI: 10.19435/j.1672-1721.2026.10.019.

[5] Kamath V, Sivakumar B M, Callado Y G, et al. Non-invasive prenatal testing as a prenatal screening test in pregnancies following assisted reproductive technologies: a systematic review and meta-analysis of diagnostic test accuracy. Fertility and Sterility, 2026, DOI:10.1016/J.FERTNSTERT.2026.04.017

[6] Zhou M Y, Chen X, Lin L, et al. The influence of maternal age, gestational age and body mass index on the proportion of foetal cell-free DNA in maternal peripheral blood. Chinese Journal of Prenatal Diagnosis (Electronic Edition), 2021, 13(03): 21–25.

[7] Wu Z Y, Jiang S Q, Huang M Z. Construction of a predictive model for factors influencing the proportion of foetal cell-free DNA in maternal peripheral blood based on the random forest algorithm. Medical Theory and Practice, 2026, 39(11): 1908–1911. DOI: 10.19381/j.issn.1001-7585.2026.11.028.

[8] Zeng M, Zhang Y, Lu H L, et al. Investigation into the correlation between foetal cell-free DNA concentration in maternal peripheral blood and adverse perinatal outcomes. Chinese Journal of Perinatal Medicine, 2023, 26(05): 409–414.

[9] Liscovitch-Brauer N, Mesika R, Rabinowitz T, et al. Machine

learning-enhanced non-invasive prenatal testing for monogenic disorders. Prenatal Diagnosis, 2024, 44(5): 657–665.

[10] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357.

[11] Zhu K D. A Study on Classification for Multi-Class Imbalanced Data and Its Applications. Inner Mongolia University, 2025. DOI:10.27224/d.cnki.gnmdu.2025.001057.

[12] Rahimzadeh F, Khanli M L, Salehpour P, et al. From Heuristics to Deep Generative Models: A Critical Review of Protein Sequence Analysis Architectures for Clinical Decision Support Systems (CDSS). Expert Systems with Applications, 2026, 324132559–132559. DOI: 10.1016/J.ESWA.2026.132559.

[13] Liang F, Pan G T, Wang Q Q, et al. A Robust Regression Algorithm Based on B-Splines. Journal of the Armoured Forces Engineering Academy, 2013, 27(1): 97–100.

[14] Zhang J N, Zou Y H, Liu H Y. Application of Bayesian Hierarchical Models to Estimating Historical Fertility Rates in China. Mathematical Statistics and Management, 1–13 [10 June 2026]. https://doi.org/10.13860/j.cnki.sltj.20260424-002.

[15] Zhang Y. A Method for Predicting lncRNA-Disease Associations Based on Ensemble Decision Trees. Yunnan University, 2020. DOI:10.27456/d.cnki.gyndu.2020.000180.

[16] Zheng H, Gao X, Yang X, et al. A study on TFT-LCD manufacturing quality prediction methods using SHAP feature selection and ensemble learning. International Journal of Computer Integrated Manufacturing, 2026, 39(6): 825–841. DOI: 10.1080/0951192X.2025.2498083.

[17] Dutta S, Bhattacharya S, Saha M M, et al. Oxidative Stress–Mediated Endothelial Dysfunction and Cell-Free Fetal DNA Release in Pre-eclampsia. Indian Journal of Clinical Biochemistry, 2026, (pre-publication): 1–10. DOI: 10.1007/S12291-026-01393-W.

[18] Rong H H, Qiu J Y, Bao Z M, et al. Evaluation of the impact of pre-pregnancy body mass index and weight gain during pregnancy on pregnancy outcomes. Contemporary Medical Collection, 2026, 24(6): 43–45.

