Vol. 18 (2026): Proceedings of 2026 International Conference on Smart Healthcare and Biomedical Engineering (ICSHBE 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/shbe.202696118

Research on NIPT Timing Selection and Fetal Abnormality Determination Based on Optimized Classification Models

Zhi Lin*

College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361001, China

**Corresponding author: Zhi Lin.*

Abstract

With the growing demand for prenatal screening, selecting appropriate timing for non-invasive prenatal testing (NIPT) and improving the accuracy of abnormality determination have become key issues in clinical practice and public health. This study, based on clinical NIPT data from a specific region, constructs and compares a Beta regression-based proportional response model with a classification model based on Gradient Boosting Decision Trees (GBDT). The research process includes data cleaning and standardization, Shapiro–Wilk normality testing, Spearman correlation screening, and stepwise Beta regression modeling to quantify the relationship between Y chromosome concentration and factors such as gestational age and BMI. Subsequently, a multi-objective optimization classification model is designed for male fetus samples, using BMI grouping critical points and corresponding NIPT timing as decision variables, solved with fmincon to simultaneously minimize potential risks and maximize detection accuracy. This is then extended to a multi-factor comprehensive discrimination index incorporating age, height, weight, and other variables. Finally, a GBDT classifier is constructed for female fetuses, comparing three sample balancing strategies—random oversampling, random undersampling, and SMOTE—and optimizing hyperparameters through grid search. The main results show that the proposed methods can significantly improve detection accuracy under different settings (the model achieves notably higher accuracy in several configurations, with female fetus classification reaching a high level under the optimal balancing strategy). The study provides actionable quantitative support for developing stratified NIPT timing and abnormality determination protocols in clinical settings. The research discusses limitations such as the single data source and the lack of dynamic modeling, and proposes future directions including expansion to multi-center data and prospective validation.

Keywords

beta regression, optimized classification model, fmincon function, gradient boosting decision tree algorithm, random oversampling

1. Introduction

Non-invasive prenatal testing (NIPT) has become an important tool for assessing the risk of fetal chromosomal aneuploidies. However, its detection accuracy is significantly affected by testing timing and maternal individual differences. In particular, the influence of pregnant women's body mass index (BMI) and gestational age on Y chromosome concentration in male fetuses alters the time at which detection reaches standard levels, thereby affecting clinical decision-making and follow-up arrangements. In addition, abnormality determination for female fetuses often faces issues such as sample imbalance and sequencing quality fluctuations, leading to decreased classifier performance in real clinical data.

To address the above problems, this paper, based on clinical NIPT data from a specific region, first uses Beta regression to quantify the relationship between Y chromosome concentration and factors such as gestational age and BMI. It then constructs a multi-objective optimization model, taking BMI grouping critical points and corresponding NIPT timing as decision variables, and employs fmincon to solve it so as to simultaneously minimize potential risks and improve detection accuracy. Finally, for female fetal abnormality determination, it compares three balancing strategies—random oversampling, undersampling, and SMOTE—and evaluates performance using Gradient Boosting Decision Trees as the base classifier. The main contributions of this paper include: quantifying the relationship between Y chromosome concentration (Dy) and key influencing factors, proposing an actionable BMI grouping and timing optimization scheme, and providing best practices for female fetal abnormality determination under imbalanced sample conditions. The remainder of the article is organized as follows: Section 2 introduces the data and methods, Section 3 presents the models and results, Section 4 discusses limitations, and Section 5 provides conclusions and prospects.

1.1 Background

NIPT technology uses high-throughput sequencing and bioinformatics to detect and analyze cell-free fetal DNA in maternal peripheral blood, assessing the risk of fetal chromosomal aneuploidies. Its non-invasive nature, high accuracy, and wide applicability across gestational weeks have made it a core method for prenatal screening [1]. Clinical experience shows that abnormal concentrations of chromosomes 21, 18, and 13 correspond respectively to Down syndrome, Edwards syndrome, and Patau syndrome. Typically, fetal sex chromosome concentrations can be detected between 10 and 25 weeks of gestation. If the Y chromosome concentration in male fetuses reaches 4% or above, and there is no abnormality in X chromosome concentration in female fetuses, the NIPT results can be considered basically accurate. At the same time, unhealthy fetuses should be detected as early as possible in clinical practice; otherwise, the risk of a shortened treatment window increases: risk is lower when detected within 12 weeks, higher between 13–27 weeks, and extremely high after 28 weeks.

Practice has confirmed that Y chromosome concentration in male fetuses is closely associated with the pregnant woman's gestational age and body mass index (BMI). Because pregnant women differ individually in age, BMI, pregnancy conditions, etc., they are generally grouped according to BMI values, with relative NIPT timing set for each group during pregnancy. This improves detection accuracy, reduces the potential risks to pregnant women caused by shortened treatment windows due to fetal health issues, and provides key support for reducing birth defects related to chromosomal abnormalities.

1.2 Problem Statement

The attachment provides NIPT data for pregnant women from a specific region. In clinical practice, sequencing failures easily occur due to testing too early or other uncertain factors. To improve detection reliability, some pregnant women undergo multiple blood collections with multiple tests or multiple tests from a single blood collection. Based on actual conditions and the attached information, a mathematical model needs to be established to solve the following problems:

Problem 1: Analyze the correlation characteristics between fetal Y chromosome concentration and maternal gestational age, BMI, and other indicators; construct the corresponding relational model and test its significance.

Problem 2: Clinical evidence confirms that BMI in pregnant women carrying male fetuses is the main factor affecting the earliest time at which fetal Y chromosome concentration reaches standard levels. Perform

reasonable grouping of BMI for pregnant women with male fetuses, clarify the BMI intervals and optimal NIPT timing for each group, minimize the pregnant women's potential risks to the greatest extent, and analyze the impact of detection errors on the results.

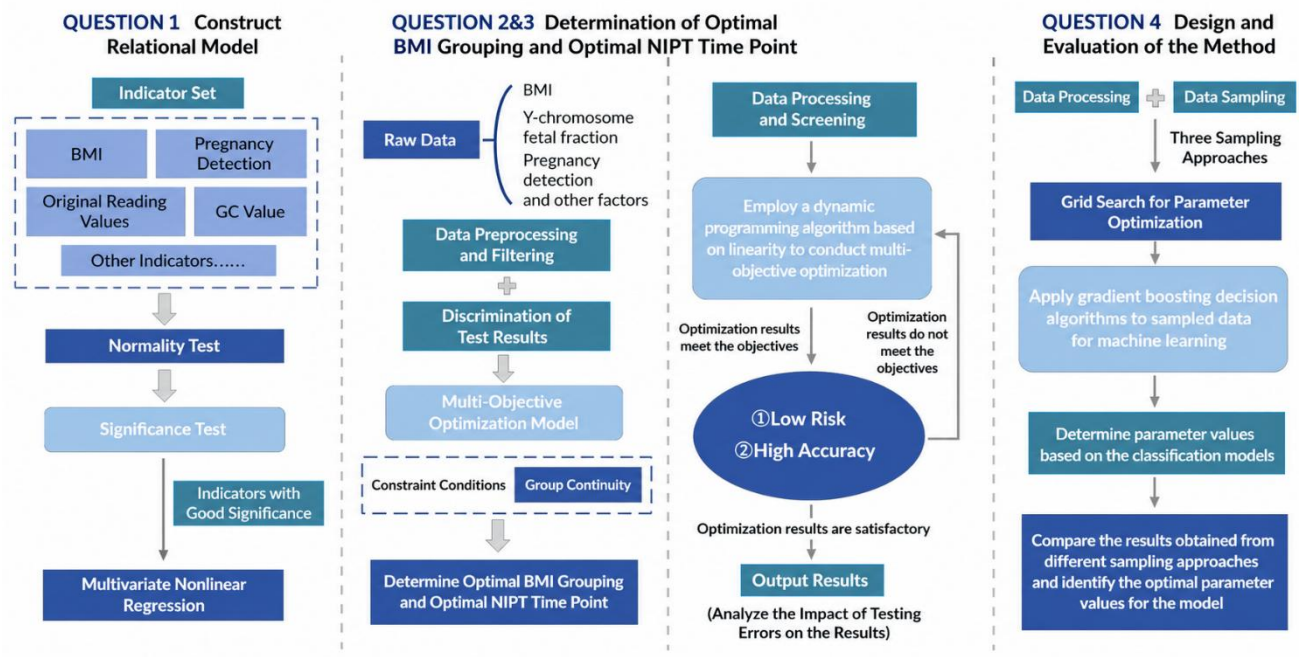
Problem 3: The time for male fetal Y chromosome concentration to reach standard levels is affected by multiple factors such as height, weight, and age. Comprehensively considering these factors, detection errors, and the proportion of fetal Y chromosome concentration reaching standard levels, perform reasonable BMI grouping for pregnant women with male fetuses, clarify the optimal NIPT timing for each group to minimize potential risks to the greatest extent, and analyze the impact of detection errors on the results.

Problem 4: Using chromosomal aneuploidy status for chromosomes 21, 18, and 13 (columns AB) in pregnant women carrying female fetuses as the determination basis, comprehensively consider Z-values of the X chromosome and the above chromosomes, GC content, read counts and related ratios, BMI, and other factors to clarify the method for determining abnormalities in female fetuses.

1.3 Solution Approach

To address the above problems, this paper conducts systematic research and solutions based on the following overall framework:

Figure 1: Overall Framework



2. Model Assumptions

(1) It is assumed that each test result for every pregnant woman is independent of the others and does not influence them. That is, fluctuations in routine blood indicators do not interfere with NIPT's determination of fetal development status, and the levels of various physiological indicators from the previous test do not directly determine the trend of results in subsequent tests.

(2) It is assumed that between testing intervals, the pregnant woman's various physiological indicators do not undergo significant changes due to external factors and remain stable within the normal reference range for pregnancy, thereby ensuring comparability between the two sets of test data.

(3) It is assumed that in non-invasive prenatal testing of pregnant women, there exists a quantifiable mathematical relationship between the Y chromosome concentration in maternal blood and indicators such as the pregnant woman's BMI (body mass index) and gestational age. The normal threshold of Y chromosome concentration for different BMI intervals can be calculated through nonlinear fitting formulas.

(4) It is assumed that feature data such as GC content provided in the attachment can effectively reflect the quality of sequencing data and can serve as one of the bases for determining whether chromosomes are normal.

(5) It is assumed that abnormal conditions among different chromosomes have no mutual influence; that is, an abnormality in one chromosome will not cause additional structural or numerical abnormalities in another chromosome.

3. Symbol Description

Table 1: Symbol Description

Symbol	Meaning	Unit
D_y	Y chromosome concentration	%
D_x	X chromosome concentration	%
BMI_i	Body Mass Index (BMI)	kg/m²
T	Gestational age at testing	weeks
age	Maternal age	years
$height$	Maternal height	cm
$weight$	Maternal weight	kg
D_{GC}	Total GC content	%
R	Risk level	–
s	Comprehensive discrimination index	–
N	Sample size	count
$b_1, b_2, ..., b_{19}$	BMI segmentation point	kg/m²
$l_1, l_2, ..., l_{20}$	Lower bound of NIPT testing time for the interval	weeks
$u_1, u_2, ..., u_{20}$	Upper bound of NIPT testing time for the interval	weeks
a_{age}	Age coefficient	–
a_{height}	Height coefficient	–
a_{weight}	Weight coefficient	–
a_y	Y chromosome concentration coefficient	–

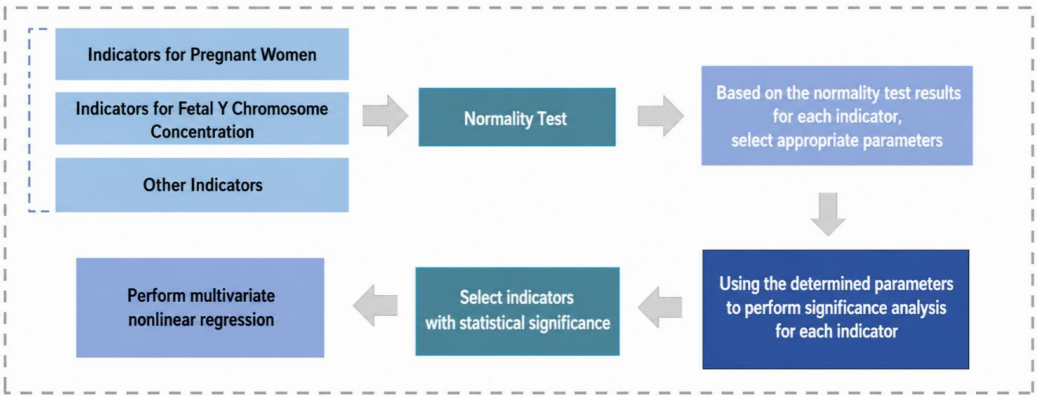
4. Model Establishment and Solution

4.1 Model Establishment and Solution for Problem 1

4.1.1 Problem Analysis

For Problem 1, the data in the attachment are first standardized and normalized to facilitate subsequent analysis. Descriptive statistics are performed on key indicator features (such as BMI, GC content, gestational age, etc.), and the results are visualized. Normality tests are then conducted to analyze the data distribution characteristics, based on which appropriate correlation coefficients (such as Pearson or Spearman) are selected for correlation analysis and significance testing. A relational model with strong correlations between each factor and Y chromosome concentration is thereby constructed. This paper adopts a stepwise Beta regression model to perform parameter estimation on this relationship and test its reasonableness.

Figure 2: Logical Framework for Question 1



4.1.2 Data Preprocessing

To ensure smooth analysis, some information columns in the attachment are preprocessed first. For example, the string "$aw+b$" in the "gestational age" data is converted to the floating-point number "$a+\frac{b}{7}$" Missing values in the data are also marked as blank for subsequent processing.

For linearly related data in the attachment such as "height," "weight," "BMI," "last menstrual period," "testing date," and "gestational age," including all of them simultaneously would cause multicollinearity and affect the model's explanatory power. Therefore, only one is retained, keeping the effective data. Data such as "serial number," "number of blood draws for testing," and "IVF pregnancy," which are either non-biological or have low sample proportions, are also excluded from model analysis. This reduces the dimensionality of the problem analysis and minimizes noise in the model, which is conducive to obtaining a more accurate model.

Descriptive statistics and visualizations of the attachment data are performed, as shown in the following figure:

Figure 3: Proportion of Male Fetuses with Fetal Y Chromosome Concentration at or Above 4%(≥4%)

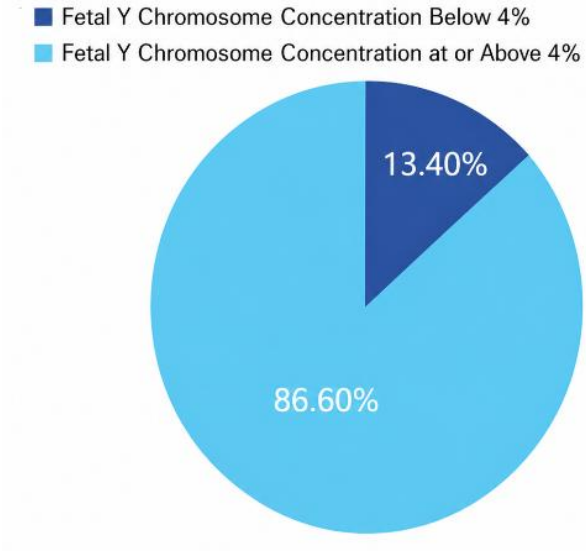


Figure 4: Scatter Plot of Maternal BMI in Male Pregnancies

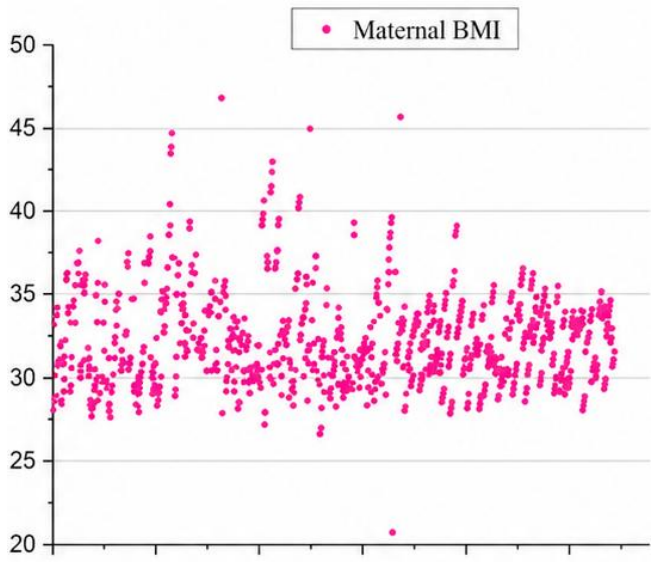


Figure 5: Scatter Plot of Gestational Age in Male Pregnancies

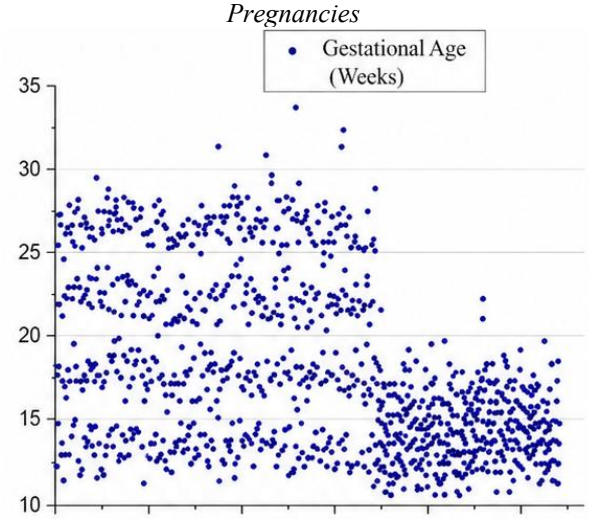
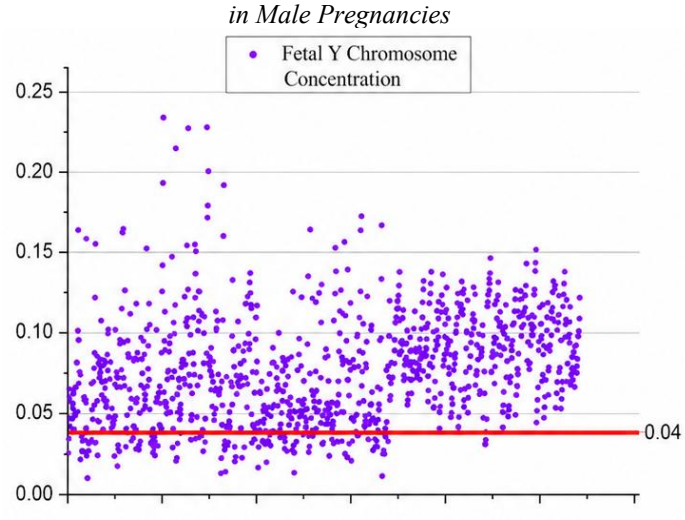


Figure 6: Scatter Plot of Fetal Y Chromosome Concentration in Male Pregnancies



4.1.3 Data Normality Testing

Next, the Shapiro-Wilk normality test is performed on each effective continuous variable X after preprocessing. The null hypothesis is that the sample comes from a normal distribution $(H_0: X \sim N(\mu, \sigma^2))$; the alternative hypothesis is that the sample does not follow a normal distribution. At a significance level of α=0.05, if the obtained $\rho < \alpha$, the null hypothesis is rejected and the variable X is considered not to follow a normal distribution; otherwise, it can be regarded as approximately normal.

Table 2: Normality Test Results

Variable Name	W Statistic	P-value	Significance	Normally Distributed?
Gestational age at testing	0.1321	0.001	***	No
Maternal BMI	0.0689	0.001	***	No
Raw read count	0.0413	0.001	***	No
Alignment ratio to reference genome	0.1119	0.001	***	No
Duplicate read ratio	0.0411	0.001	***	No
Unique aligned read count	0.0487	0.001	***	No
GC content	0.0367	0.0014	***	No
Z-value for chromosome 13	0.0291	0.0289	**	No
Z-value for chromosome 18	0.0323	0.0088	***	No
Z-value for chromosome 21	0.0188	0.4833	-	Yes
Z-value for X chromosome	0.0421	0.001	***	No
Z-value for Y chromosome	0.0524	0.001	***	No
X chromosome concentration	0.0382	0.001	***	No
Y chromosome concentration	0.0485	0.001	***	No
GC content of chromosome 13	0.064	0.001	***	No
GC content of chromosome 18	0.0484	0.001	***	No
GC content of chromosome 21	0.0398	0.001	***	No

According to the code calculation results, most variables show significant deviation from a normal distribution. Therefore, in subsequent correlation analysis, correlation coefficients that do not rely on the assumption of normal distribution should be used.

4.1.4 Correlation Analysis

Since the SW normality test above shows that the relevant indicator variables deviate from normality, this paper uses the non-parametric Spearman correlation coefficient for analysis.

(1) Principle of Spearman Correlation Analysis:

The core idea of the Spearman correlation coefficient is that it does not directly measure the linear relationship between two variables, but rather measures the consistency or monotonic relationship between the "ranks" (or "orderings") of the two variables. It therefore does not depend on the assumption of normal distribution of the data. The formula for the correlation coefficient is:

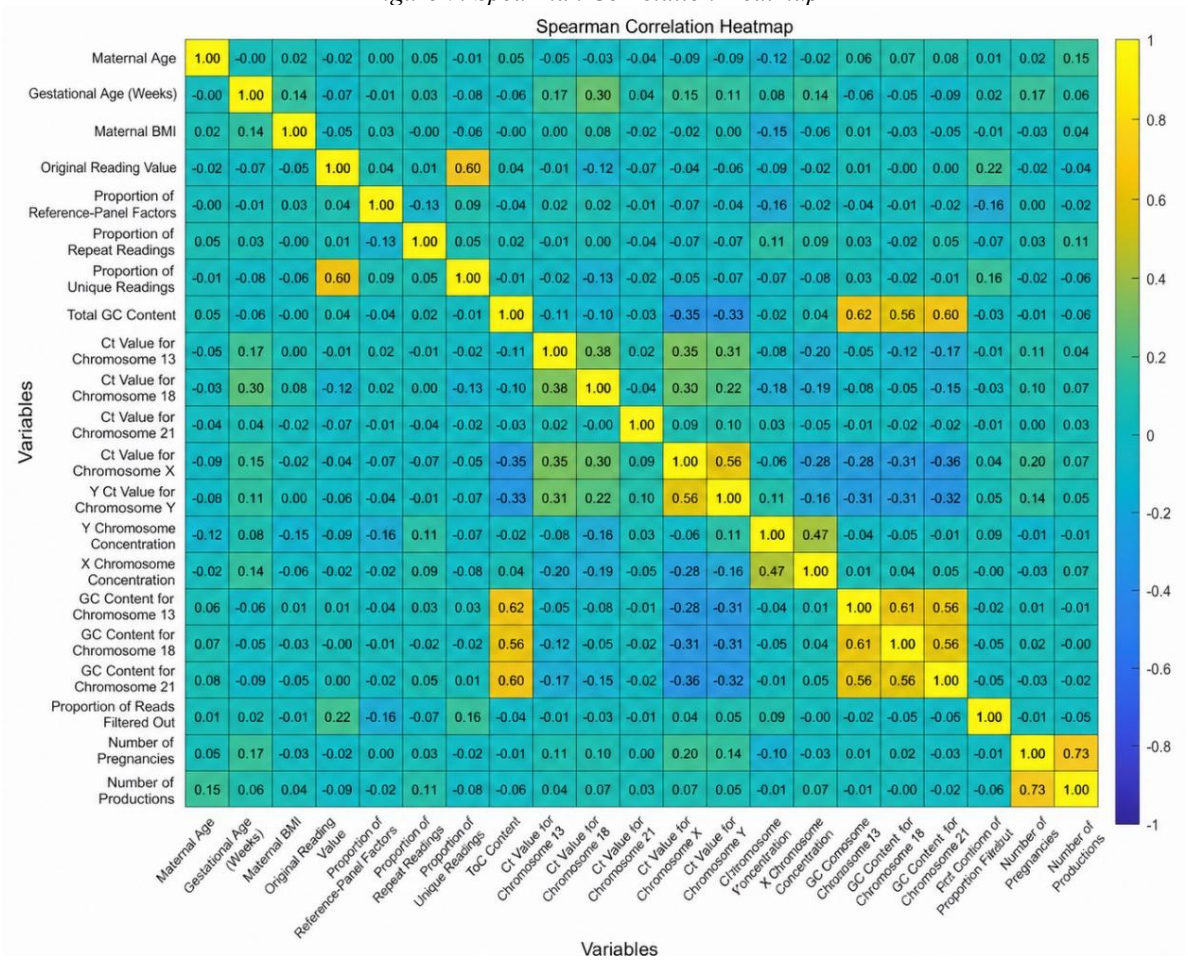
$$r_s = \frac{cov(R_g(X), R_g(Y))}{\sigma_{R_g(X)}\sigma_{R_g(Y)}} \tag{1}$$

That is, it calculates the Pearson correlation coefficient of the ranked data, where $R_g(X_i), R_g(Y_i)$ represent the ranks of the two variables X_i, Y_i in the data, respectively.

(2) Calculation Results:

Matlab was used to compute the Spearman correlation coefficients between variables, and the results were visualized as a heatmap of the Spearman correlation coefficient matrix (see figure below):

Figure 7: Spearman Correlation Heatmap



From the results, the following conclusions can be drawn: The variable with a relatively ideal correlation with Y chromosome concentration is X chromosome concentration (ρ=), which is consistent with biological principles. The correlations between the other variables and Y chromosome concentration are all weak (additional data table results can be added). This indicates that, when considering the influence of all these factors, Y chromosome concentration does not show strong correlations.

Therefore, regression analysis is needed to further control for variables other than the core ones. After performing significance tests on the correlation coefficients of these variables using Matlab, the results are as follows:

Table 3: Correlation Analysis Results

Variable Name	ρ	P-value	Significance
X chromosome concentration	0.4744	0.0000	***
Z-value for chromosome 18	-0.1782	0.0000	***
Alignment ratio to reference genome	-0.1594	0.0000	***
Maternal BMI	-0.1550	0.0000	***
Age	-0.1172	0.0001	***
Z-value for Y chromosome	0.1135	0.0002	***
Duplicate read ratio	0.1058	0.0005	***
Number of pregnancies	-0.1043	0.0035	***
Raw read count	-0.0860	0.0047	***
Filtered read ratio	0.0860	0.0046	***
Gestational age at testing	0.0843	0.0055	***
Z-value for chromosome 13	-0.0766	0.0117	**
Unique aligned read count	-0.0694	0.0224	**
Z-value for X chromosome	-0.0559	0.0661	*
GC content of chromosome 18	-0.0499	0.1007	-
GC content of chromosome 13	-0.0372	0.2213	-
Z-value for chromosome 21	0.0335	0.2704	-
GC content	-0.0166	0.5844	-
GC content of chromosome 21	-0.0107	0.7252	-
parity	-0.0072	0.8137	-

The results still show that only X chromosome concentration has a moderate positive correlation with Y chromosome concentration, while the other significant variables are all weakly correlated. This indicates that a multivariate model should be used for subsequent model building.

4.1.5 Establishment and Solution of the Beta Regression Model

Based on the analysis results above, a relational model between Y chromosome concentration and its influencing indicators is next established and solved. Given the proportional nature of Y chromosome concentration ($Y_i \in (0,1)$) and the fact that variables do not follow a normal distribution, this paper constructs a Beta regression model to describe the relationship affecting Y chromosome concentration.

(1) Principle of Beta Regression:

The Beta regression model is a regression model suitable for response variables $Y_i \in (0,1)$. It is specifically designed for continuous proportional data. The model assumes:

$$y_i \sim Beta(\alpha_i, \beta_i),\ \alpha_i = \mu_i \phi,\ \beta_i = (1-\mu_i)\phi \tag{2}$$

where $\mu_i \in (0,1)$ represents the conditional mean, and the precision parameter $\phi > 0$ reflects the variance. Using a link function (e.g., logit link) to map μ_i to the real line allows treatment similar to a linear model:

$$LOGIT(\mu_i) = \log \frac{\mu_i}{1-\mu_i} \tag{3}$$

All parameters in the model can be estimated by maximum likelihood, and their significance is tested using z-tests. Note that independent variables should be standardized before calculation.

(2) Regression Results:

Significant variables from the previous Spearman correlation analysis were selected as independent variables for the Beta regression. Stepwise regression results obtained via Matlab are shown in the following table:

Table 4: Stepwise Beta Regression Results for Y Chromosome Concentration (with Significance)

Variable Name	(1)	(2)	(3)	(4)	(5)
X chromosome concentration	0.2365	0.2364	0.2362	0.2333	0.2342

	(***)	(***)	(***)	(***)	(***)
Z-value for chromosome 18	-0.0690 (***)	-0.0690 (***)	-0.0691 (***)	-0.0618 (***)	-0.0620 (***)
Maternal BMI	-0.0658 (***)	-0.0659 (***)	-0.0660 (***)	-0.0672 (***)	-0.0676 (***)
Age	-0.0457 (***)	-0.0459 (***)	-0.0459 (***)	-0.0469 (***)	-0.0476 (***)
Z-value for Y chromosome	0.0581 (***)	0.0581 (***)	0.0581 (***)	0.0628 (***)	0.0626 (***)
Duplicate read ratio	0.0243 (*)	0.0242 (*)	0.0243 (*)	0.0241 (*)	0.0283 (*)
Raw read count	-0.0666 (***)	-0.0667 (***)	-0.0650 (***)	-0.0643 (***)	-0.0666 (***)
Filtered read ratio	0.0358 (**)	0.0359 (**)	0.0360 (**)	0.0352 (**)	0.0395 (***)
Gestational age at testing	0.0336 (**)	0.0337 (**)	0.0337 (**)	0.0357 (**)	0.0368 (**)
Alignment ratio to reference genome	-0.0191 (ns)	-0.0190 (ns)	-0.0188 (ns)	-0.0187 (ns)	-
Z-value for chromosome 13	0.0233 (ns)	0.0232 (ns)	0.0232 (ns)	-	-
Unique aligned read count	0.0024 (ns)	0.0025 (ns)	-	-	-
parity	-0.0012 (ns)	-	-	-	-
const	-2.5109 (***)	-2.5109 (***)	-2.5109 (***)	-2.5106 (***)	-2.5107 (***)

In these five stepwise Beta regression models, the conclusions are relatively stable and significant. During the stepwise process, the calculated F-value gradually increases, indicating that the model's significance level improves progressively as insignificant influencing factors are removed. Qualitatively, Y chromosome concentration has positive correlations with gestational age and sex chromosome-related indicators, and negative correlations with BMI and age. The best model estimated by Beta regression (the final step) has the following parameters:

Table 5: Beta Regression Results

Variable	Standardized Coefficient	Z -value	P -value	Significance
const	-2.5107	-213.44	0	***
X chromosome concentration	0.2342	19.87	0	***
Z-value for chromosome 18	-0.0620	-4.89	0	***
Maternal BMI	-0.0676	-5.99	0	***
Age	-0.0476	-4.18	0	***
Z-value for Y chromosome	0.0626	5.58	0	***
Duplicate read ratio	0.0283	2.53	0.0120	*
Raw read count	-0.0666	-5.69	0	***
Filtered read ratio	0.0395	3.48	0.0005	***
Gestational age at testing	0.0368	3.10	0.0020	**
precision	4.5859	106.15	0	***

The model expression is:

$$LOGIT(\mu) = \Sigma \alpha_i X_i + \alpha_0 \tag{4}$$

$$\hat{\mu} = \frac{1}{1 + e^{-LOGIT(\mu)}},\ Y|X \sim Beta\left(\hat{\mu}\phi, (1-\hat{\mu})\phi\right) \tag{5}$$

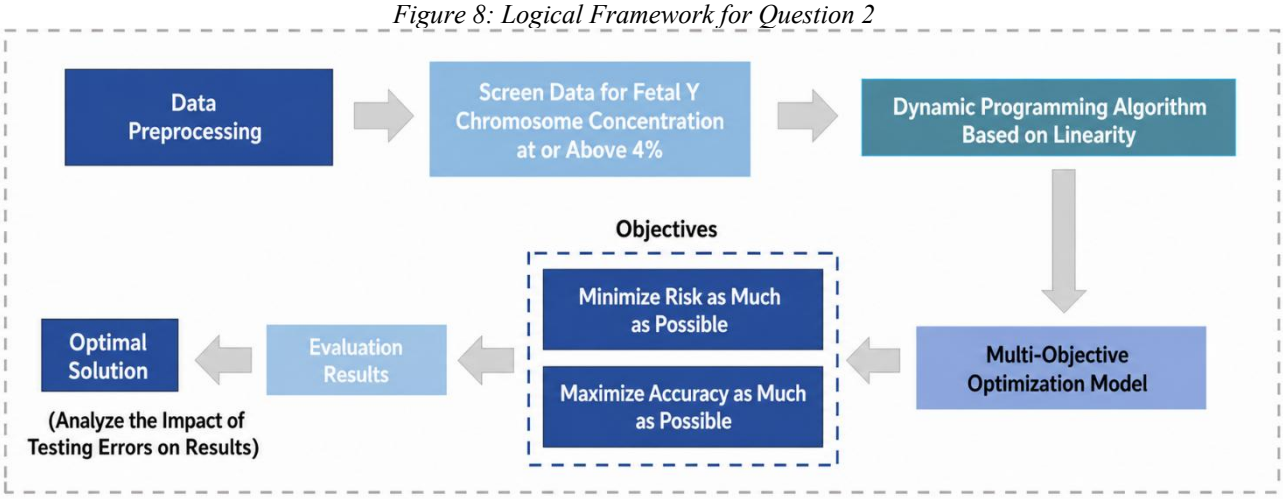
where X_i, α_i are the standardized coefficients from the table above, α_0 is the const term, $\phi = 4.5859$ is the Beta precision coefficient.

4.2 Model Establishment and Solution for Problem 2

4.2.1 Problem Analysis

Problem 2 requires grouping pregnant women carrying male fetuses according to their BMI and determining the optimal NIPT testing time point for each group to minimize the potential risk of an unhealthy fetus. The problem states that BMI of pregnant women with male fetuses is the main factor affecting the earliest time at which fetal Y chromosome concentration reaches the standard (i.e., the earliest time the concentration reaches or exceeds 4%). This question aims to establish an optimized classification model and construct an NIPT testing scheme for pregnant women carrying male fetuses. Accurate prediction data are first filtered from all male-fetus samples in the attachment, and then the NIPT timing with the minimum risk is derived.

Figure 8: Logical Framework for Question 2



4.2.2 Data Preprocessing

Building upon Problem 1, a new column "Risk" R is added according to the relationship between gestational weeks and risk given in the problem. It can be expressed using a piecewise function as follows:

$$R=\begin{cases}1, & T_i\in(0,12]\\ 2, & T_i\in(12,28]\\ 3, & T_i\in[28,+\infty)\end{cases} \tag{6}$$

The information in "chromosomal aneuploidy" is marked as 0 (predicted unaffected) or 1 (predicted affected) to create the column "Prediction." Similarly, "fetal health status" is converted into a 0/1 variable as the column "Actual." A new column "Prediction Correctness" is added by comparing the "Prediction" and "Actual" columns (1 if the same, 0 otherwise).

Indicators with relatively strong correlations with Y chromosome concentration are retained for subsequent analysis.

4.2.3 Decision Variables

According to the problem requirements, the optimization model aims to group pregnant women carrying male fetuses by BMI and provide the optimal NIPT time points to minimize potential risk. The decision variables include BMI grouping critical points and the corresponding upper and lower bounds of NIPT testing times. The problem does not specify a fixed number of groups. Considering that treating the number of groups as a decision variable would introduce coupling issues and turn the problem into mixed-integer optimization (increasing solution difficulty), a relatively large number of groups (20) is adopted. In the subsequent solving process, adjacent narrow BMI intervals can be merged to obtain a more practical grouping scheme. The decision variables are shown in the table below:

Table 6: Decision Variables for Problem 2

Decision Variable	Meaning	Quantity
$b_1, b_2, ..., b_{19}$	BMI segmentation points	19
$l_1, l_2, ..., l_{20}$	Lower bound of corresponding NIPT testing time for the interval	20
$u_1, u_2, ..., u_{20}$	Upper bound of corresponding NIPT testing time for the interval	20

4.2.4 Multi-Objective Optimization

(1) Constraints:

The constraints of this model are clear. The segmentation points must satisfy:

$$b_1 \leqslant b_2 \leqslant ... \leqslant b_{19} \tag{7}$$

The time bounds must satisfy:

$$l_1 < u_1, l_2 < u_2, ..., l_{20} < u_{20} \tag{8}$$

There are a total of 38 constraints, all linear inequality constraints.

(2) Objective Functions:

As stated earlier, the primary goal is to minimize the "pregnant woman's potential risk." Considering measurement errors, detection accuracy is also included as an optimization objective. These two objectives are combined to form the overall objective function for optimization.

Objective 1: Minimization of Potential Risk

The problem states: "Typically, fetal sex chromosome concentration can be detected between 10 and 25 weeks of gestation. If the Y chromosome concentration in a male fetus reaches or exceeds 4%, or the X chromosome concentration in a female fetus shows no abnormality, the NIPT result can be considered basically accurate. Otherwise, the accuracy requirement is difficult to guarantee." Therefore, the following screening process is applied to the data: For each test record, the pregnant woman's BMI value is used to determine which BMI interval (Min_i, Max_i) it belongs to, and it is checked whether the gestational age *Time* falls within the corresponding optimal NIPT testing time window for that BMI group. Only records that satisfy this condition are retained. Additionally, records must further satisfy the condition that the Y chromosome concentration ≥ 4% $D_y \geqslant 0.04$ when $T_i \in [10, 25]$ and $D_{GC} \in [0.4, 0.6]$. These screened records are used as input data. Let the number of retained records be N_{real}. The potential risk objective function is then defined as the average risk across these retained records: $O_1 = \frac{\Sigma Risk_i}{N_{real}}$

Objective 2: Maximizaton of Detection Accuracy

This objective quantifies the accuracy of all testing records. After the above filtering, the "Prediction Correctness" column added during data preprocessing is used. The accuracy objective function $O_2 = \frac{\Sigma C_i}{N_{real}}$ is the proportion of correct predictions among retained records.

(3) Comprehensive Objective Function:

Combining the two objective functions above, and considering that $O_1 \in [1, 3]$ and $O_2 \in [0, 1]$, both of which are dimensionless quantities, the comprehensive objective function is defined as: $O = O_1 - O_2$. According to the meanings of the two individual objectives, this combined objective is a minimization problem.

4.2.5 Solution of the Model for Problem 2

This is a continuous variable optimization model. Due to data screening requirements, it is difficult to express through a simple linear form and involves complex logical relationships. It can be solved using the fmincon function in Matlab.

By consolidating the previously mentioned decision variables and constraints, the solution can be obtained through the code implementation.

The aforementioned optimization model can be solved via the Matlab fmincon function. The attached file provides the optimization results for all decision variables: the first 19 variables represent the BMI cut-off points, while the remaining 40 values represent the optimal upper and lower bounds for the NIPT testing time intervals.

The BMI grouping for pregnant women carrying male fetuses and the corresponding optimal NIPT time windows are as follows:

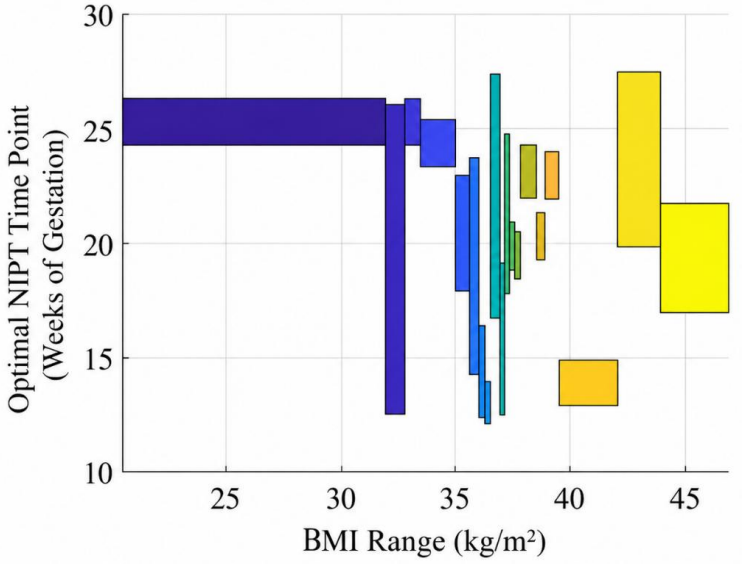
Table 7: BMI Grouping and Optimal NIPT Timing for Male-Fetus Pregnancies

Group No.	BMI Range for Pregnant Women with Male Fetus	Optimal NIPT Testing Time Interval (weeks)
1	[20.00, 31.98)	[24.2795, 26.3084]
2	[31.98, 32.83)	[12.5344, 26.0507]
3	[32.83, 33.50)	[24.2586, 26.2884]
4	[33.50, 35.06)	[23.3130, 25.3723]
5	[35.06, 35.66)	[17.8941, 22.9355]
6	[35.66, 36.07)	[14.2403, 23.7037]
7	[36.07, 36.32)	[12.3834, 16.3878]
8	[36.32, 36.56)	[12.1269, 13.9477]
9	[36.56, 36.57)	[23.4617, 25.5175]
10	[36.57, 36.98)	[16.7068, 27.3821]
11	[36.98, 37.16)	[12.4692, 19.0936]
12	[37.16, 37.37)	[17.7722, 24.7569]
13	[37.37, 37.60)	[18.7979, 20.8883]
14	[37.60, 37.85)	[18.3893, 20.4793]
15	[37.85, 38.55)	[21.9362, 24.2606]
16	[38.55, 38.92)	[19.2236, 21.3139]
17	[38.92, 39.54)	[21.8800, 23.9595]
18	[39.54, 42.07)	[12.9080, 14.8801]
19	[42.07, 43.91)	[19.8275, 27.4816]
20	above 43.91	[16.9781, 21.7280]

4.2.6 Analysis of Results

The 20 BMI groups and their optimal NIPT times are visualized in the figure below:

Figure 9: Optimal NIPT Time Points Corresponding to BMI Groups in Male Pregnancies



The figure above illustrates the optimal NIPT testing windows for different BMI groups, similar to those discussed in Problem 2. As observed from the figure, there is no direct correlation between BMI and the optimal NIPT testing time. Given that several narrow adjacent groups can be merged, it is unnecessary to categorize BMI into as many as 20 distinct classes in practical clinical applications.

The following section analyzes the impact of measurement errors on the results. In this study, the primary sources of error are Y-chromosome concentration and GC content.

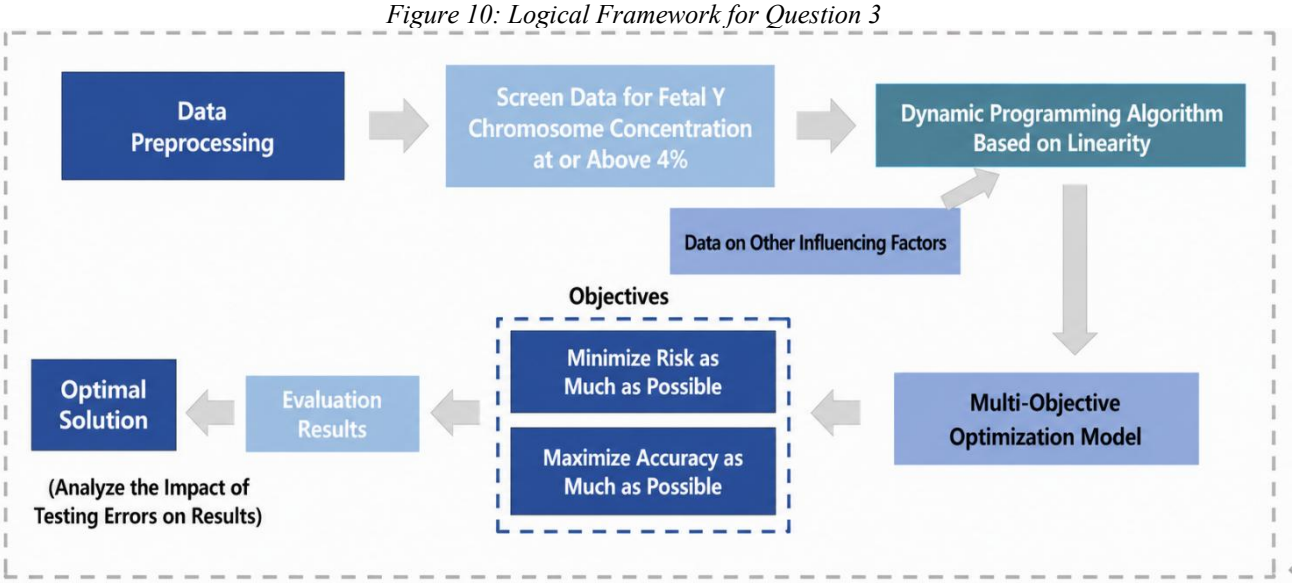
According to the computational results, after applying the BMI–NIPT interval filtering, 200 out of the total 1,082 records passed the initial screening. Subsequently, the impact of measurement errors was analyzed within these 200 records. After controlling for Y-chromosome concentration and GC content as defined in this problem, 97 records passed the final screening, indicating that the model's control criteria are relatively stringent. The samples that passed the screening exhibited a high overall accuracy (87.63%) and robust specificity (89.67%). In contrast, the precision of the data for samples that failed the screening was only 8.7%, reflecting poor predictive reliability. These results demonstrate that the model possesses effective classification and screening performance.

4.3 Model Establishment and Solution for Problem 3

4.3.1 Problem Analysis

Problem 3 can be viewed as an extension of Problem 2. It also requires minimization of potential risk to pregnant women. However, instead of using only the Y chromosome reaching-standard proportion as the accuracy criterion, it incorporates additional factors such as height, weight, and age. The data screening step is modified from a single "gestational age + Y chromosome" judgment to a comprehensive multi-factor judgment.

Figure 10: Logical Framework for Question 3



4.3.2 Data Processing for Problem 3

Building on Problem 2, the columns "height," "weight," and "age" are additionally retained for multi-factor analysis.

4.3.3 Decision Variables

Building upon the model from Problem 2, it is necessary to incorporate additional decision variables related to other factors. New decision variables are introduced: the age factor a_{age} , height factor a_{height} , and weight factor a_{weight} . Since the Y-chromosome concentration is no longer treated as an independent control indicator, it is also regarded as one of the decision variables influenced by multiple factors; thus, the Y-chromosome factor a_y is introduced. The complete set of decision variables is as follows:

Table 8: Decision Variables for Problem 3

Decision Variable	Meaning	Quantity
$b_1, b_2, \ldots, b_{19}$	BMI segmentation points	19
$l_1, l_2, \ldots, l_{20}$	Lower NIPT time bounds	20
$u_1, u_2, \ldots, u_{20}$	Upper NIPT time bounds	20
a_{age}	Age coefficient	1
a_{height}	Height coefficient	1
a_{weight}	Weight coefficient	1
a_y	Y chromosome concentration coefficient	1

4.3.4 Multi-Objective Optimization

(1) Constraints:

The constraints for this problem are consistent with those in Problem 2; the newly introduced decision variables do not affect the existing constraints of the model.

(2) Objective Functions:

The screening condition is updated to a comprehensive multi-factor score s, mapped via the Sigmoid function to [0,1]. Records are retained based on a threshold on this score combined with GC content control.

As previously noted, this problem requires a modification of the data screening criteria based on the model from Problem 2.Specifically, $D_y \geqslant 0.04$ is no longer treated as an individual filtering condition when $T_i \in [10, 25]$; instead, a more comprehensive multi-factor assessment approach is adopted:

$$s = a_{age} \cdot age + a_{weight} \cdot weight + a_{height} \cdot height + a_y \cdot D_y \tag{9}$$

After mapping the value of s to the range $(0, 1)$ using the Sigmoid function: $S(s) = \frac{1}{1 + e^{-s}}$, we obtain. This value serves as the criterion for data retention: if $S(s) < 0.5$, the data is excluded; otherwise, it is included. This decision mechanism, combined with the concentration-based assessment, constitutes the data screening criteria for Problem 3.

4.3.5 Solution of the Model for Problem 3

The aforementioned optimization model can be solved using the fmincon function in Matlab. The optimization results for all decision variables are provided in the attachment. The definitions of the first 59 decision variables remain identical to those in Problem 2, representing the BMI cut-off points and the optimal NIPT testing intervals.

The results obtained are as follows:

Table 9: BMI Grouping and Optimal NIPT Timing for Male-Fetus Pregnancies (Problem 3)

Group No.	BMI Range for Pregnant Women with Male Fetus	Optimal NIPT Testing Time Interval (weeks)
1	[20.00, 32.02)	[18.2600, 19.5671]
2	[32.02, 32.09)	[23.7081, 25.5530]
3	[32.09, 32.15)	[23.3181, 24.6171]
4	[32.15, 32.19)	[16.3689, 17.6722]
5	[32.19, 32.27)	[22.5653, 23.8677]
6	[32.27, 32.56)	[14.9296, 23.9086]
7	[32.56, 32.86)	[21.5754, 22.8807]
8	[32.86, 33.10)	[21.7581, 23.0630]
9	[33.10, 33.43)	[23.7849, 27.7034]
10	[33.43, 34.12)	[15.8340, 17.1352]
11	[34.12, 35.05)	[18.2058, 24.4419]
12	[35.05, 35.97)	[20.5080, 21.8150]
13	[35.97, 36.64)	[23.2678, 24.5670]
14	[36.64, 37.25)	[17.6938, 23.7899]
15	[37.25, 37.84)	[12.6187, 18.2478]

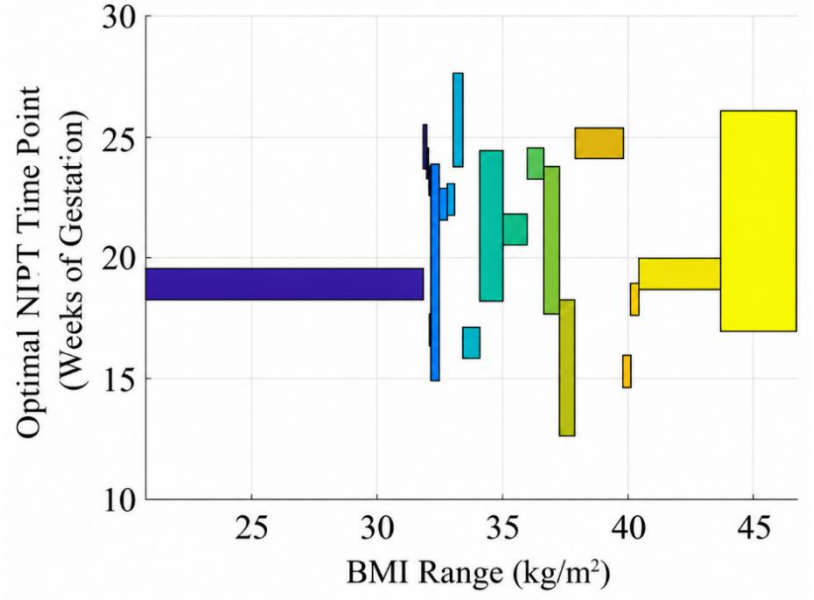
16	[37.84, 39.82)	[24.1171, 25.4104]
17	[39.82, 40.11)	[14.6689, 15.9629]
18	[40.11, 40.44)	[17.6168, 18.9231]
19	[40.44, 43.69)	[18.6667, 19.9742]
20	above 43.69	[16.9512, 26.1366]

The last four decision variables are the coefficients for age, height, weight, and Y chromosome concentration. The optimized comprehensive judgment expression is:

$$s = 0.050 \cdot age + 0.048 \cdot height + 0.047 \cdot weight + 83.34 \cdot D_y \tag{10}$$

The visualization of the results is shown below:

Figure 11: Optimal NIPT Time Points Corresponding to BMI Groups in Male Pregnancies



The figure above illustrates the optimal NIPT testing windows for different BMI groups, similar to the findings in Problem 2. It can be observed from the figure that several narrow adjacent groups are eligible for merging. In practical clinical applications, it is unnecessary to categorize BMI into as many as 20 distinct classes.

The following section analyzes the impact of measurement errors on the results. In this problem, the sources of error include not only Y-chromosome concentration and GC content but also the age, height, and weight factors incorporated into the model analysis.

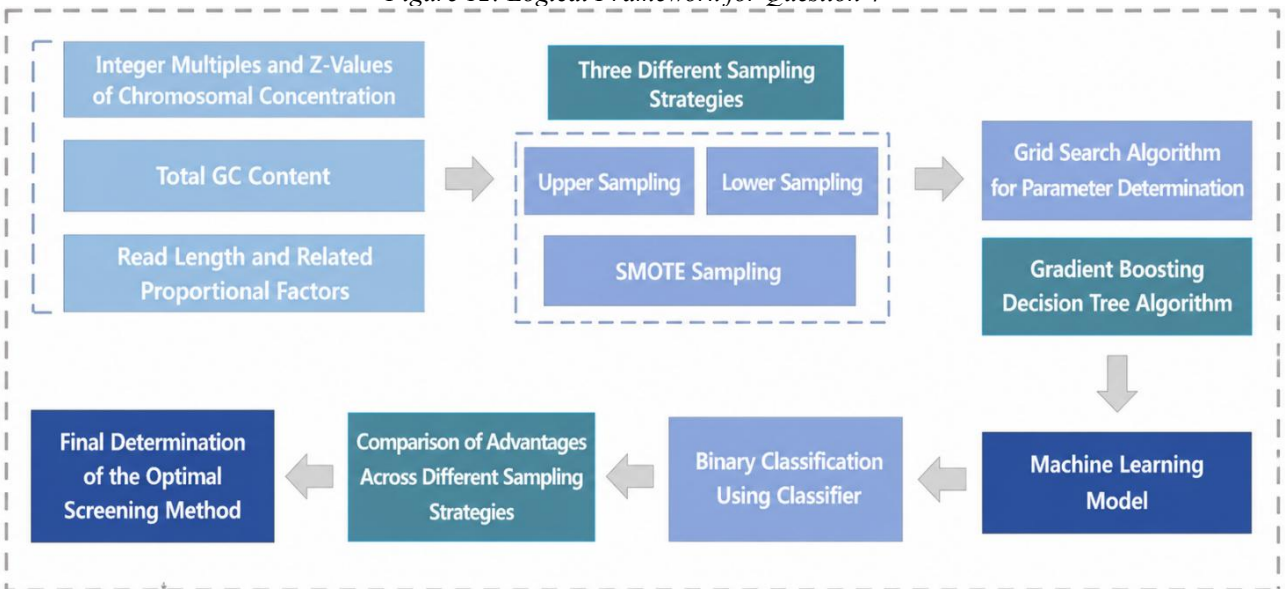
According to the computational results, after applying the BMI–NIPT interval filtering, 119 out of the total 1,082 records passed the initial screening. Subsequently, the impact of measurement errors was analyzed within these 119 records. After controlling for the multi-factor judgment metric s and the GC content as defined in this problem, 59 records passed the final screening, indicating that the model's control criteria are relatively stringent. The samples that passed the screening exhibited a high overall accuracy (93.22%) and robust specificity (76.67%). In contrast, the precision of the data for samples that failed the screening was only 6.67%, reflecting poor predictive reliability. these results demonstrate that the model possesses effective classification and screening performance.

4.4 Model Establishment and Solution for Problem 4

4.4.1 Problem Analysis

Problem 4 requires the establishment of a determination method for fetal abnormalities in female fetuses. This task can be framed as a binary classification problem, for which a classification model will be constructed for analysis.

Figure 12: Logical Framework for Question 4



4.4.2 Data Processing

The preprocessing for all types of indicators remains consistent with the methods used in the first three problems. Since the problem does not require the determination of specific chromosomal aneuploidy types, any record containing an abnormality in T13, T18, or T21 can be labeled as 1, while records with no abnormalities are labeled as 0.

Figure 13: Scatter Plot of GC Content Distribution in Female Pregnancies

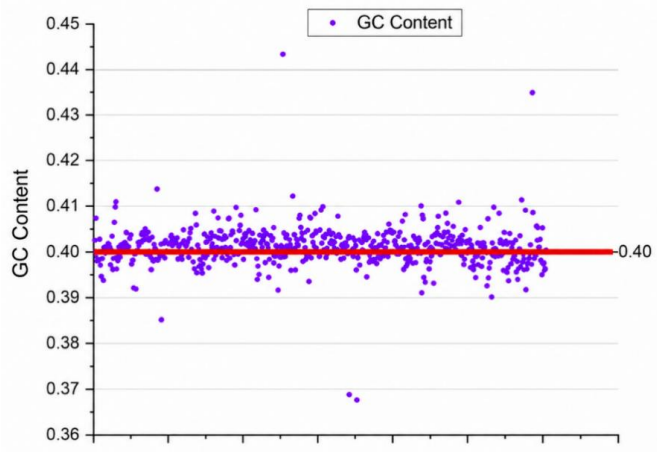
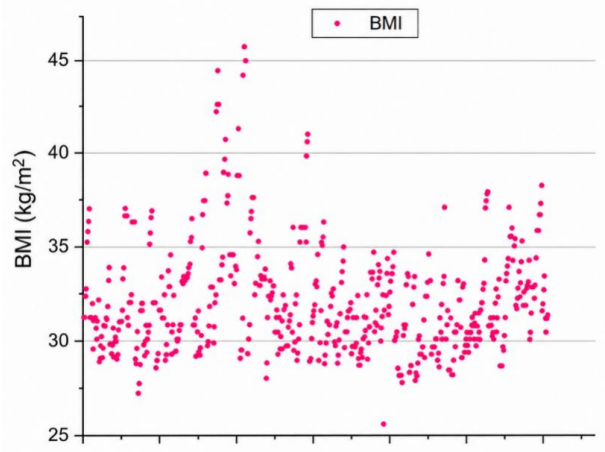


Figure 14: Scatter Plot of BMI in Female Pregnancies



4.4.3 Handling Imbalanced Samples

In this study, three sampling methods—Random Over-sampling, Random Under-sampling, and Synthetic Minority Over-sampling Technique (SMOTE)—are employed to process the data, with their performance analyzed during the subsequent model-solving phase. Each method possesses distinct characteristics (to be elaborated). Sampling is executed exclusively within the cross-validation process or the training set; data in the test set remain unprocessed to prevent the occurrence of data leakage associated with re-sampling.

(1) Principles:

Random Oversampling: This method increases the number of samples in the minority class to match the majority class. While keeping the size of the majority class constant, $(n_0^{'} = n_0)$, the minority class size

$n_1^{'} = \left\lceil \frac{\eta}{1-\eta} n_0 \right\rceil$ is increased by adding duplicated copies until it reaches that value. Over-sampling is theoretically equivalent to increasing the weight of the minority class samples.

Random Undersampling: This method reduces the number of samples in the majority class to match the minority class. Samples are randomly removed from the majority class without replacement until the size $n_0^{'}$ satisfies $n_0^{'} = \left\lfloor \frac{1-\eta}{\eta} n_1 \right\rfloor$. Under-sampling is theoretically equivalent to decreasing the weight of the majority class samples.

SMOTE: As a synthetic over-sampling approach, its core principle involves generating new samples by interpolating between existing minority class data points. For each minority class sample $x_j \in N_k(x_i)$ its k-nearest neighbor set $\hat{x} = x_i + \lambda(x_j - x_i)$ is first calculated. In each iteration, a neighbor is randomly selected from and a interpolation factor $\lambda \sim u(0,1)$ is chosen. A new synthetic sample is then generated as: x_i . This process is repeated until the sample size reaches the required threshold, thereby completing the sampling procedure.

(2) Sampling Results:

After processing the data using the three aforementioned sampling methods, the resulting sampling outcomes are summarized in the table below:

Table 10: Comparison of Three Sampling Methods

Data set	Normal (0)	Abnormal (1)	Total N	Positive Rate (%)	Class Ratio (1:0)
Original data set	538	67	605	11.07	67/538
Random Oversampling	538	538	1076	50.00	538/538
Random Undersampling	67	67	134	50.00	67/67
SMOTE	538	538	1076	50.00	538/538

4.4.4 Model Establishment for Problem 4

In this problem, a classification model for determining fetal abnormalities in female fetuses will be established using the Gradient Boosting Decision Tree (GBDT) as the core algorithm. The model takes the various influential indicators as inputs and chromosomal aneuploidies as outputs. All data are partitioned into a training set and a test set with a ratio of 7:3. On the training side, Grid Search is combined with K-fold Cross-Validation to jointly optimize hyperparameters—including the learning rate, number of base learners, maximum depth, and minimum samples per leaf. The objective is to maximize accuracy while simultaneously constraining metrics such as precision, recall, and the F1-score.

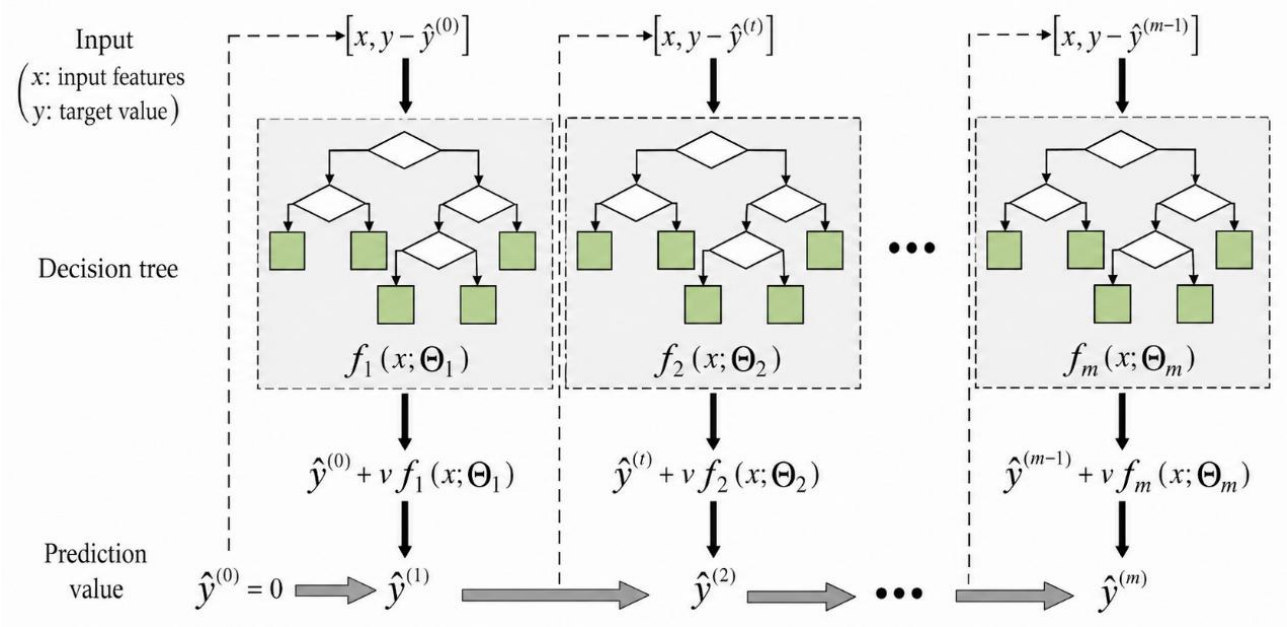
Principles of the GBDT Algorithm:

The Gradient Boosting Decision Tree (GBDT) is an additive model based on the boosting ensemble learning framework. It progressively approximates the true classification results by constructing a series of decision tree models. In each iteration, a new decision tree is trained based on the classification error of the previous round, thereby continuously reducing the model's objective loss. GBDT demonstrates robust performance in both regression and classification tasks. The optimization expression for its final classifier is given by:

$$GBDT(\omega) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \gamma_m h_m(x) + F_0(x) \tag{11}$$

where $F_0(x)$ is the initial model, serving as the initialization function for model prediction. In the m-th iteration, $h_m(x)$ is a function trained in the direction of the negative gradient of the loss function based on the errors from the preceding round. $\gamma_m(x)$ represents the scaling coefficient (or step size) for each round's decision tree, which is determined during optimization to adjust the weight and influence of each individual tree on the overall model. The schematic diagram of the algorithm is shown below:

Figure 15: Schematic Diagram of the GBDT Algorithm



By utilizing this iterative algorithm, classification errors are progressively reduced, enabling the construction of a binary classification model that meets the requirements of this study.

Hyperparameter Grid Search for the GBDT Algorithm:

The following figures illustrate the cross-validation performance distribution of the GBDT model across the hyperparameter grid, based on four distinct training sets: the original dataset and the three different sampling methods. These result plots highlight the "advantageous regions" of each strategy, providing a basis for selecting the optimal parameters for each approach:

Figure 16: Original Dataset Search Results

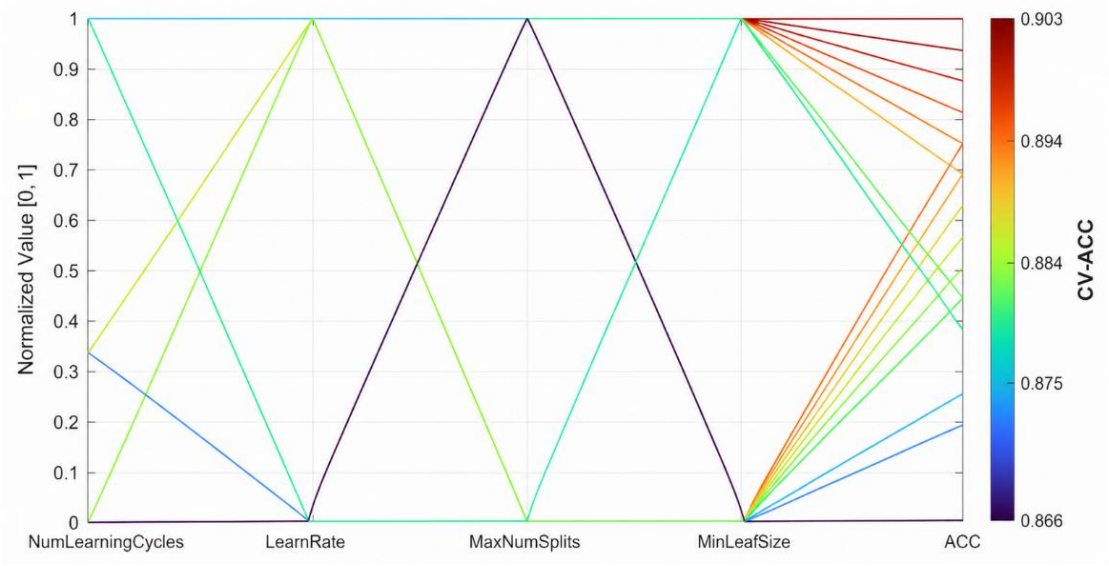


Figure 17: Search Results for the Training Set

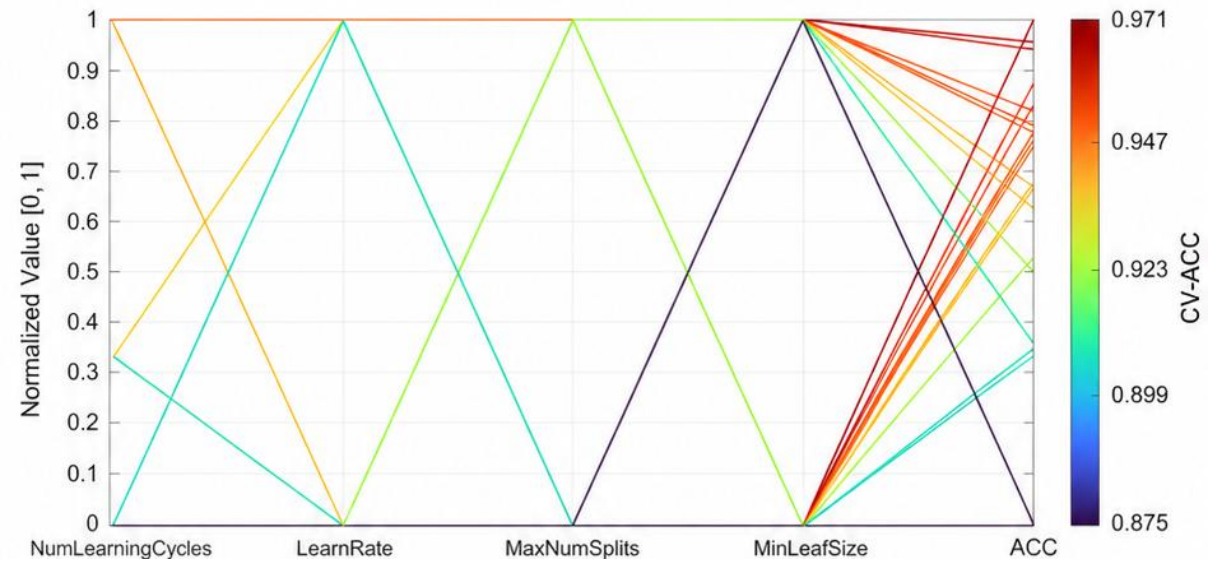


Figure 18: Search Results for the Testing Set

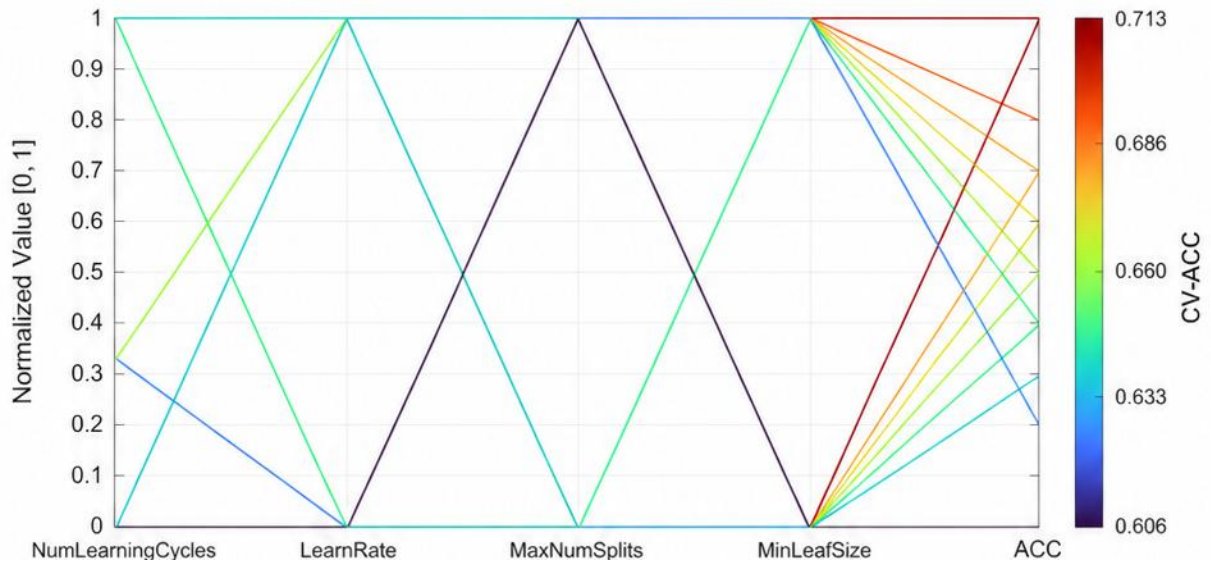
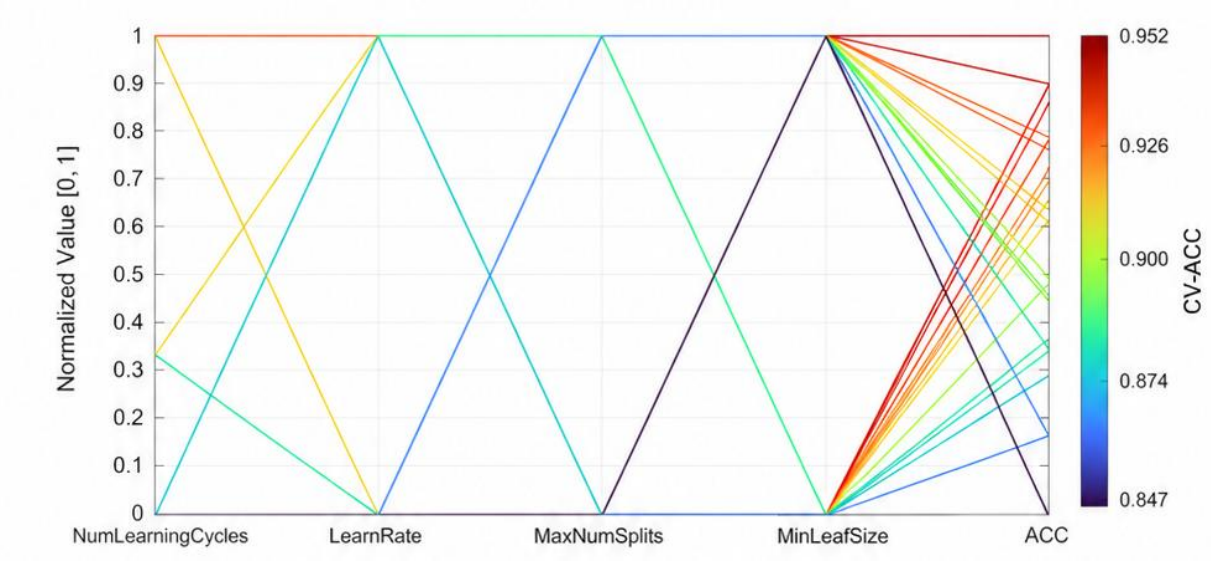


Figure 19: Search Results for the SMOTE Dataset



The figures above provide a clear comparison of the effectiveness of the four training schemes: Random Over-sampling achieves the highest cross-validation accuracy (approximately 0.97) and exhibits a continuous high-performance region, indicating superior robustness. SMOTE also yields high accuracy (appro ximately 0.95), though its overall peak values are slightly lower than those of the over-sampling strategy. The Original Data, due to inherent class imbalance, results in lower accuracy (approximately 0.90) and is more susceptible to noise or interference. Under-sampling achieves an accuracy of only about 0.71 due to excessive information loss, making it unsuitable for this model. Consequently, the Over-sampling + GBDT modeling approach is prioritized for its superior performance. The specific parameters adopted are listed in the table below:

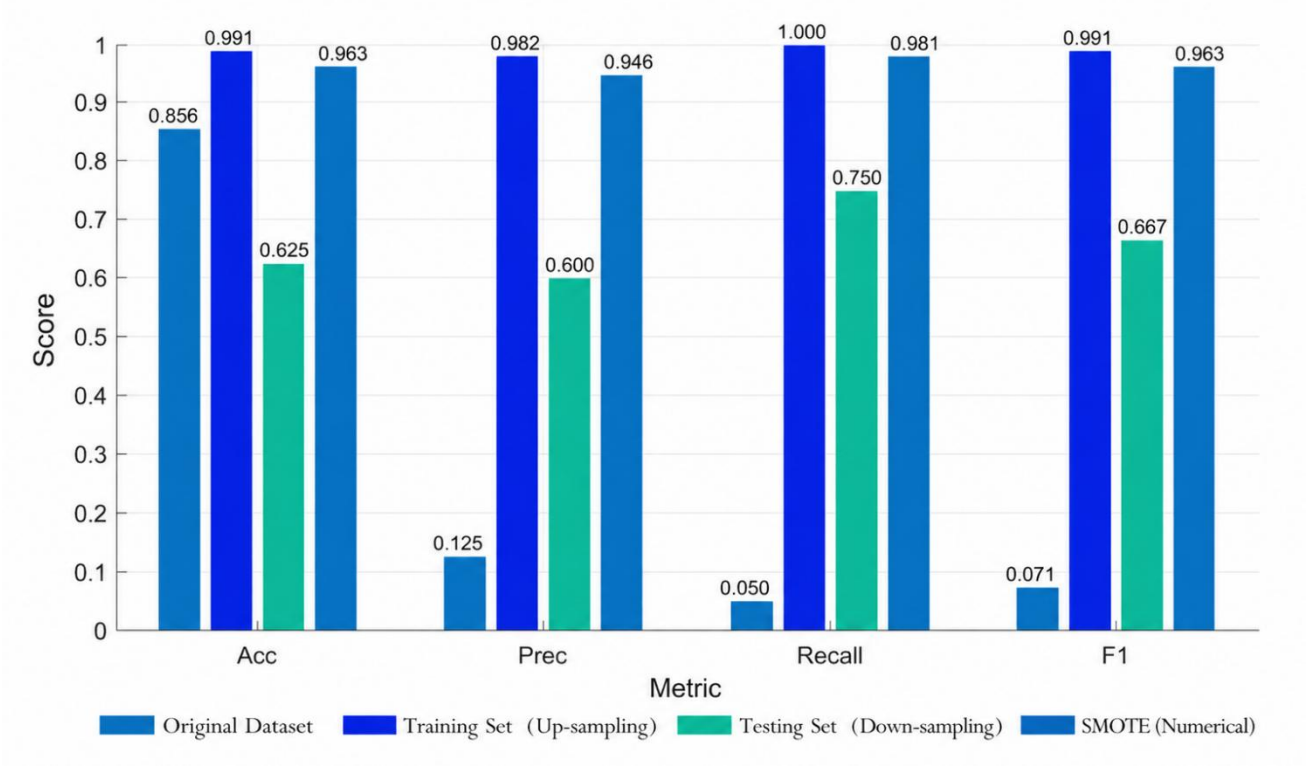
Table 11: Optimal Parameters for Different Datasets

Data set	NumLearningCycles	LearnRate	MaxSplits	MinLeaf
Original data set	100	0.05	30	5
Random Oversampling	200	0.10	30	1
Random Undersampling	200	0.05	10	1
SMOTE	200	0.10	30	5

4.4.5 Determination of Abnormalities in Female Fetuses

Upon completing the grid search and determining the optimal parameter combinations for each sampling method, the GBDT models were retrained using the training set and evaluated on the test set. A comparative analysis of accuracy, precision, recall, and F1-score across the different combinations was conducted, yielding the following results:

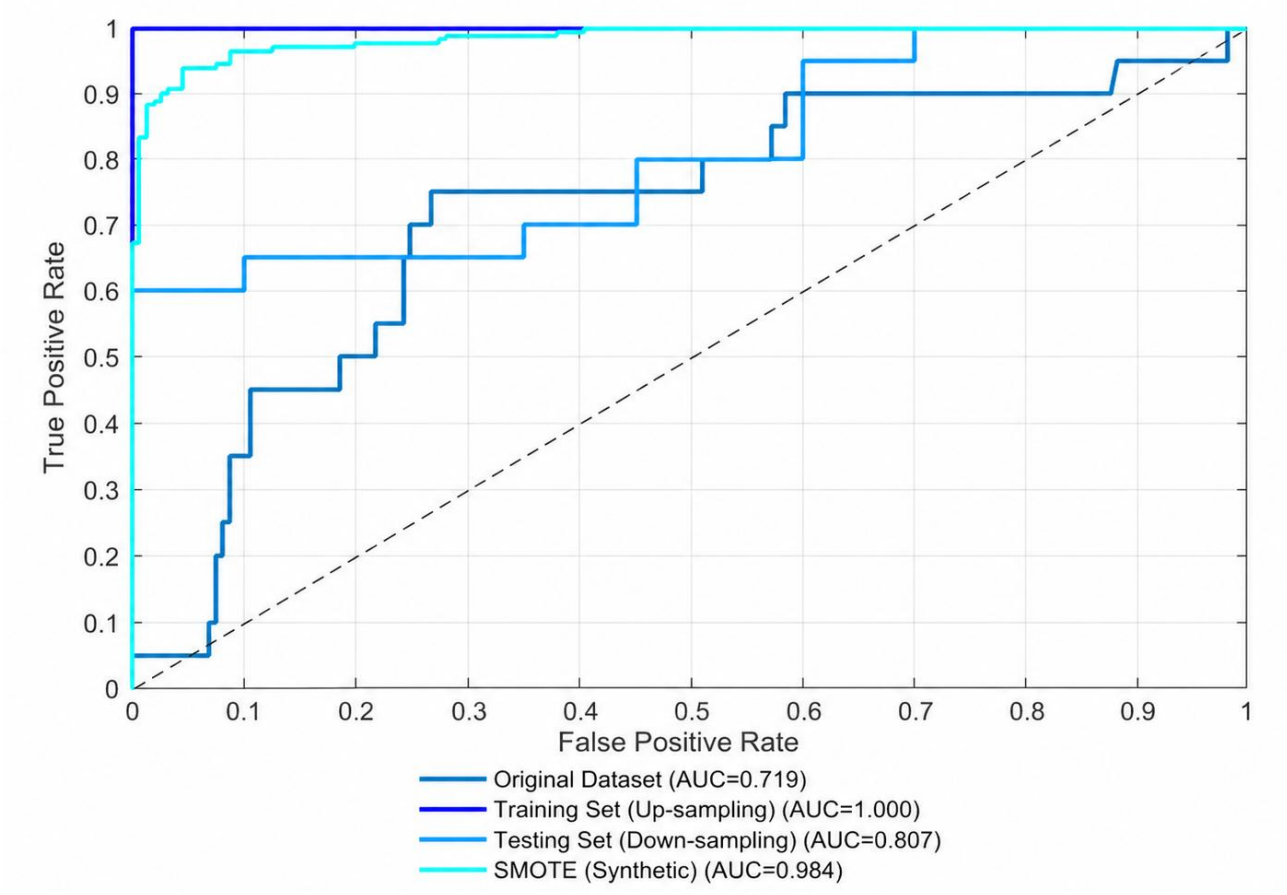
Figure 20: Comparison of Accuracy, Precision, Recall, and F1 Score for Four Combinations of Datasets



The results indicate that the over-sampling scheme consistently outperforms others across all metrics, demonstrating the best overall performance. SMOTE also exhibits stable predictive capabilities, though slightly inferior to the over-sampling approach. Under-sampling suffers from low overall performance due to significant data loss. The results for the original data align with expectations: because of the extreme disparity between minority and majority class sizes, the learning process is biased toward the majority class. This leads to a seemingly respectable accuracy but extremely low precision and recall. Therefore, the over-sampling approach is determined to be optimal and serves as the primary determination method for this problem.

The ROC curves below illustrate the performance of the three sampling methods on the training set:

Figure 21: Comparison of Accuracy, Precision, Recall, and F1 Score for Four Combinations of Datasets



As observed in the figure, over-sampling maintains a substantial predictive accuracy and the strongest diagnostic capability under all conditions. The evaluations for the remaining three training schemes are consistent with previous findings, further validating the dominant advantage of over-sampling.

In summary, based on the female fetal detection data provided in the attachment and considering multiple factors such as X-chromosome content and GC content, a classification model was constructed using the GBDT algorithm to predict chromosomal aneuploidies. After determining the optimal parameters via grid search and comparing performance across three sampling methods, the over-sampling approach was identified as having the best comprehensive effect. For practical application, an "Over-sampling + GBDT" configuration—characterized by a low learning rate, a high number of base learners, and moderate tree depth—can be employed to construct the abnormality determination model. This model provides high accuracy, precision, and recall, offering significant reliability for clinical use.

5. Model Evaluation, Improvement, and Promotion

5.1 Advantages of the Model

(1) The model system is relatively comprehensive, integrating multiple algorithmic models and constructing a complete framework ranging from statistical analysis to machine learning, demonstrating strong adaptability to the data. (2) Clinical practical significance was considered in the quantification of the data. Core indicators with meaningful clinical value were incorporated into the model analysis, giving the study strong practical relevance.

(3) The model employs various techniques such as significance testing and cross-validation to ensure credibility. Standardized corrections and practically grounded screening measures were applied during data processing, guaranteeing the reliability of the results.

5.2 Limitations of the Model

(1) The model does not account for dynamic changes in individual pregnant women, such as fluctuations in estrogen and progesterone levels from early to late pregnancy, or physiological adjustments in maternal blood volume and metabolic rate as the fetus develops. This makes it difficult for the model to precisely adapt to differences in health status across different stages of pregnancy.

(2) The model relies on data from a single source with uneven sample distribution, which may introduce bias in model selection.

5.3 Model Improvements

(1) Multi-dimensional data can be added to supplement the existing dataset.

(2) Based on the expanded data, more relevant indicators that reflect individual characteristics of pregnant women can be constructed to make the model more accurate.

(3) Decision variables can be further optimized, parameters adjusted, model solving complexity reduced, and result stability improved.

5.4 Model Promotion and Extension

(1) The NIPT timing selection model proposed in this paper can not only effectively detect T21, T18, and T13 aneuploidies, but also provide indications and important reference value for sex chromosome abnormalities, rare chromosomal copy number trisomies, other chromosomal microdeletions, and particularly excess sex chromosome material [2].

(2) The application of optimal NIPT timing selection is not limited to fetal health monitoring. Its potential in identifying and managing maternal complications and comorbidities, such as pregnancy-induced hypertension, gestational diabetes mellitus (GDM), and fetal growth restriction (FGR), is also receiving increasing attention [3].

6. Conclusions

This paper, based on regional clinical NIPT data, constructed Beta regression and multi-objective optimization grouping models, and combined them with GBDT to complete binary classification for female fetal abnormality determination. The main findings include: 1) Y chromosome concentration shows a significant positive correlation with gestational age and a significant negative correlation with BMI; 2) BMI-based grouping and optimized timing can significantly improve the detection accuracy of NIPT for male fetuses; 3) Incorporating multiple factors such as age, height, and weight further enhances model accuracy; 4) In addressing sample imbalance, random oversampling combined with GBDT delivers the best performance. The limitations of the study include the single data source, the lack of modeling for dynamic physiological changes during pregnancy, and some overly fine-grained groupings. Future work is recommended to expand to multi-center data, introduce time-series dynamic models, and conduct prospective validation in real clinical workflows.

References

[1] Zhang, L. L., Zhuo, Z. Z., Huang, S. W., et al. (2025). Comparative analysis of the application value of NIPT and NIPT-plus in prenatal screening. *Guizhou Medical Journal*, 49(06), 942-944.

[2] Zhang, Y. M., Han, Y., & Qiu, H. G. (2025). Application value of non-invasive prenatal screening technology in the detection of chromosomal abnormalities and microdeletions. *Journal of Medical Science of Yanbian University*, 48(06), 59-61. https://doi.org/10.16068/j.1000-1824.2025.06.018.

[3] Shi, W. H., & Xu, C. M. (2025). Application value of non-invasive prenatal testing in the diagnosis of maternal complications and comorbidities in obstetrics. *Journal of Practical Obstetrics and Gynecology*, 41(08), 617-620.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access