A Multi-Model Fusion and Risk Optimization Approach for Non-Invasive Prenatal Testing Based on Random Forest and K-means Clustering

Pengyu Yang *
Nanjing Forestry University, Nanjing, China
* Corresponding Author Email: 1158430736@qq.com

Abstract. This paper focuses on the complex relationship between Y-chromosome concentration and multiple factors in non-invasive prenatal testing, and constructs a methodological framework that integrates multiple models. First, to address the significant nonlinear characteristics among variables, a random forest model is introduced to assess feature importance, identify key influencing factors, and enhance model robustness. Building on this foundation, segmented polynomial regression is employed to effectively improve overall fitting performance through interval partitioning and local modeling. Furthermore, the concentration threshold is transformed into a risk minimization problem to construct an optimization model. K-means clustering is utilized to achieve joint optimization of grouping and testing timing, thereby reducing overall risk while ensuring structural stability. Additionally, to address class imbalance, SMOTE is introduced for data augmentation, and a classification model is established using Gradient Boosted Decision Trees (GBDT) to effectively identify abnormal samples. Overall, this method integrates Random Forests, piecewise regression, clustering, and ensemble learning models, demonstrating strong generalization capabilities and stability, and holds significant practical value in complex data modeling and risk control.

Keywords: Random Forest, K-means Clustering, Gradient Boosting Decision Tree.

1. Introduction

Non-invasive prenatal testing (NIPT) has emerged as a significant development in the field of medical diagnostics in recent years, playing a crucial role in enhancing the safety and accuracy of testing. However, in practical applications, test results are often influenced by a combination of factors, such as gestational age, maternal clinical indicators, and chromosomal characteristics. These variables typically exhibit complex nonlinear relationships, making it difficult for traditional analytical methods based on single statistical models to yield stable and reliable results.

Existing studies largely rely on linear regression or simple correlation analysis to characterize variable relationships. However, when dealing with high-dimensional, multi-source, and unevenly distributed data, such methods are easily constrained by model assumptions and struggle to adapt to the nonlinear structures and local variations present in real-world data. Additionally, regarding testing time optimization and risk control, traditional methods often address these issues in isolation, lacking a unified modeling framework, which limits their overall effectiveness. Furthermore, the issue of data imbalance further impacts the model’s performance in identifying anomalous samples.

To address these shortcomings, this paper proposes a multi-stage collaborative analysis framework based on the integration of algorithmic models. This framework first utilizes random forests to extract key features and combines them with segmented regression methods to characterize complex relationships; subsequently, it introduces clustering algorithms and optimization models to achieve joint optimization of detection time and risk; finally, it enhances classification discrimination capabilities through ensemble learning models [1][2]. By organically integrating multiple algorithms to form a unified and flexible modeling process, this approach not only enhances the model’s adaptability to complex data but also provides a solution with broad applicability for similar medical testing problems.

2. Nonlinear Relationship Modeling Based on Random Forest and Piecewise Regression

2.1. Feature Importance Evaluation and Key Variable Selection Using Random Forest

Y-chromosome concentration is related to many factors such as gestational age and BMI. However, the relationships between each variable and Y-chromosome concentration show no obvious linearity, nor can they be regarded as having a stable monotonic trend. Therefore, it is not appropriate to directly use Pearson correlation coefficient or rank correlation coefficient to measure such correlations.

Under such circumstances, random forest is adopted to evaluate the importance of each variable. Compared with traditional correlation analysis methods, random forest can produce more robust results under nonlinear and nonmonotonic conditions.

In the model, Y-chromosome concentration is set as the target variable, and the input features include maternal age, time of last menstrual period, gestational age, BMI, number of pregnancies and number of deliveries. Considering that BMI has comprehensively reflected the information of height and weight, these two indicators are no longer separately included in the model. The number of decision trees in the random forest is set as $n_{\text{tree}} = 100$ to strike a balance between computational efficiency and model stability [3].

By training the model and calculating feature importance, the following results are obtained:

Table 1. Feature importance

Feature Index	Importance
Maternal gestational age	0.294714
Maternal BMI	0.269218
Maternal time of last menstrual period	0.182070
Maternal age	0.161636
Number of pregnancies	0.054493
Number of deliveries	0.037869

From the results as shown in the Table 1, the importance of the number of pregnancies and the number of deliveries is obviously low, and their influence on Y-chromosome concentration is relatively limited, so these two features are ignored in subsequent modeling. In contrast, gestational age, BMI, time of last menstrual period and age make more significant contributions and can be regarded as the main influencing factors.

From the overall distribution of variable relationships, the variation trends differ significantly across different intervals, and fitting with a single function does not yield satisfactory results. A more natural approach is to divide the data into several intervals and build regression models within each interval separately. After such processing, the fitting performance is usually more stable.

Based on this idea, samples are divided into multiple intervals according to age, gestational age and time of last menstrual period, and polynomial regression models are constructed in each interval. The general form of the model can be expressed as $C_Y = f(t_{\text{gestational week}}, t_{\text{last menstrual period}}, d_{\text{age}}, b_{\text{BMI}}) + \varepsilon$, where ε is the random error term and is assumed to follow a normal distribution.

This piecewise modeling method compensates to some extent for the difficulty of a single model in describing complex relationships, and also provides a more reliable foundation for subsequent optimization analysis.

2.2. Model Solution

Since the scatter distribution is relatively complex, it is difficult to directly determine the specific form of the regression model. Therefore, a fixed functional structure is not adopted here; instead, an exhaustive search is used to select expressions with good fitting performance from candidate models.

Simply put, different polynomial combinations are tested within each segment, and the final model is determined under the principle of minimum error.

In the solution process, the fitting performance of the model is mainly evaluated by the residual sum of squares, whose basic form is

$$Q=\sum_{i=1}^{n}(C_{Y,i}-\hat{C}_{Y,i})^2 \tag{1}$$

The corresponding optimal regression coefficients are obtained when Q reaches the minimum value.

To further evaluate model quality, the coefficient of determination R^2 is introduced as an evaluation metric, expressed as

$$R^2=1-\frac{\sum_{i=1}^{n}(C_{Y,i}-\hat{C}_{Y,i})^2}{\sum_{i=1}^{n}(C_{Y,i}-\bar{C}_{Y})^2} \tag{2}$$

The value range of R^2 is [0,1], and a larger value indicates a stronger ability of the model to explain the data.

Within each segment, the model with the maximum R^2 is selected as the final result for that interval. This process is performed independently between different segments, and the optimal regression model corresponding to each interval is finally obtained.

2.3. Result Analysis

From the overall results, the fitting performance of each piecewise model varies to some extent. The coefficient of determination for most segments is above 0.5, indicating that the model has certain explanatory ability within these intervals. Meanwhile, the R^2 value of some segments exceeds 0.8, representing better fitting performance and a more stable relationship between variables in these ranges.

The fitting effect is relatively weak in a few individual intervals. This is caused by the uneven distribution of samples across different intervals, together with the inherently complex relationships among variables, so fluctuations in local models are unavoidable. Nevertheless, compared with the single model, piecewise modeling shows a significant improvement overall.

In summary, this piecewise polynomial regression method can properly characterize the relationship between Y-chromosome concentration and key variables, and also provides a foundation for subsequent optimization based on this relationship.

2.4. Test Analysis

After the models were established, significance tests were conducted on each piecewise regression model. The F-test is used to judge whether the overall model is statistically significant, and its basic form is:

$$F=\frac{SSR/p}{SSE/(n-p-1)} \tag{3}$$

The null hypothesis is that all regression coefficients are zero, namely $H_0:\beta_0=\beta_1=\cdots=\beta_p$. If the null hypothesis is rejected, the overall model is considered to have explanatory power.

According to the test results, the significance varies among different segments. Some segments have small p-values, indicating strong statistical significance of the models. A few segments show slightly weaker significance, but the overall trend still reflects the basic relationship between variables. Such results are acceptable considering the fluctuations inherent in the data.

3. Risk-Minimization-Based Optimization Model for Detection Timing

3.1. Optimization Model Establishment

In NIPT testing, detection can be performed when the Y-chromosome concentration reaches 4%. In practice, if the concentration only slightly exceeds this threshold, the impact on the detection time is relatively limited. However, a large excess may lead to an obvious shift in detection time, thereby increasing potential risks. Therefore, a more reasonable approach is to conduct analysis around the critical state of "reaching exactly 4%".

Based on the relational model established in the previous section, the corresponding gestational age at which the Y-chromosome concentration reaches 4% can be predicted for each sample, so as to obtain the corresponding relationship between BMI and gestational age under the critical state. It should be noted that the purpose of this step is not simple prediction, but to provide a relatively stable reference basis for subsequent grouping and optimization.

On this basis, the problem is transformed into an optimization problem aiming at risk minimization. Intuitively, the optimal detection time points corresponding to different BMI intervals vary. Improper grouping may lead to premature or delayed detection for some samples, thus increasing the overall risk. Therefore, it is necessary to determine the grouping scheme and the corresponding detection time points for each group simultaneously to minimize the total risk.

The model mainly involves the following variables: the number of intervals k, the boundaries of each interval m_i, the sample set s_i in each interval, and the corresponding optimal detection time point t_{s_i}. Here, t_{s_i} represents the representative detection time in the i-th interval.

To ensure the rationality of the model, several basic constraints are set for the variables. First, the number of intervals should be controlled within a certain range. Too few intervals will lead to excessive intra-group differences, while too many will increase computational complexity, so it is set as $2 \le k \le 20$.

Second, the boundaries of each interval should satisfy the order relationship and be restricted in combination with the data range, namely $28 \le m_1 \le m_2 \le \cdots \le m_{k+1} \le 38$.

A relatively robust definition is adopted for the detection time point of each interval. Specifically, after sorting the detection time points of all samples in the interval, the 95th percentile is taken as the representative value of the interval, which can avoid the influence of extreme values to a certain extent. The corresponding expression is:

$$t_{s_i} = \begin{cases} t_{0.95|s_i|}, & 0.95|s_i| \notin \mathbb{N} \\ \dfrac{t_{0.95|s_i|} + t_{0.95|s_i|+1}}{2}, & 0.95|s_i| \in \mathbb{N} \end{cases} \tag{4}$$

The setting of the risk function incorporates the medical significance of different gestational age stages. Generally, early detection is associated with low risk, medium-term detection shows a gradual increase in risk, and late detection carries high risk. Thus, the risk function can be expressed piecewise as:

$$R(t) = \begin{cases} 1, & t \le 12 \\ 1 + \dfrac{t-12}{15}, & 12 < t \le 27 \\ 3, & t > 27 \end{cases} \tag{5}$$

On this basis, the total risk can be expressed as the weighted sum of risks in each interval:

$$R_{\text{total}} = \sum_{i=1}^{k} R(t_{s_i}) \cdot |s_i| \tag{6}$$

Finally, the problem is transformed into the following optimization model:

$$\min R_{\text{total}} = \sum_{i=1}^{k} R(t_{s_i}) \cdot |s_i| \tag{7}$$

The core of this model is to simultaneously determine the optimal grouping scheme and the corresponding detection time points for each group to minimize the total risk. A clustering method will be used to solve this optimization problem in the subsequent work, and the influence of the number of groups on the results will be further analyzed.

3.2. Model Solution

On the basis of the above optimization model, the specific grouping scheme and corresponding detection time points need to be further determined. Since grouping is carried out based on the single feature of BMI, while ensuring intra-group similarity and inter-group difference, the clustering method is a more natural choice. After comprehensive consideration, the K-means clustering algorithm is adopted to complete the grouping.

Specifically, given the number of groups k, clustering is performed on the BMI data to divide the samples into k intervals. The sample mean within each interval is taken as the representative to obtain the corresponding detection time point, and the total risk under this grouping scheme is further calculated. For the update of clustering centers, the form can be expressed as

$$h_i = \frac{1}{n} \sum_{j=1}^{n} b_{\text{BMI},j} \tag{8}$$

where h_i represents the center position of the i-th cluster.

As different numbers of groups have a great impact on the results, calculations are carried out one by one for $k \in [2, 20]$. For each value of k, the clustering process is repeated and the corresponding total risk is calculated to ensure the stability of the results.

In terms of grouping performance evaluation, in addition to direct comparison of total risk, the silhouette coefficient and Calinski-Harabasz index are introduced for auxiliary judgment. The silhouette coefficient is defined as

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

which measures the compactness within clusters and the separation between clusters. The variance ratio index is expressed as

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \tag{10}$$

and a larger value indicates a clearer clustering structure.

Based on the above indicators, the model is compared under different numbers of groups to determine a better grouping scheme.

3.3. Result Analysis

First, from the perspective of clustering structure, the silhouette coefficient and Calinski-Harabasz index corresponding to each number of groups are shown in Fig. 1. It can be seen that when the number of groups is between 8 and 13, the overall silhouette coefficient changes relatively smoothly, indicating that the clustering structure is relatively stable at this time. After further increasing the number of groups, the data begin to scatter and the interval characteristics gradually weaken, so larger numbers of groups are no longer considered.

On this basis, the changes in total risk under different numbers of groups are further compared, as shown in Fig. 2. It can be seen that the total risk shows a downward trend with the increase of the number of groups, but the decline rate slows down significantly after reaching a certain scale. When the number of groups is 13, the total risk reaches the minimum value, while the clustering structure still maintains good stability, so it is taken as the final scheme.

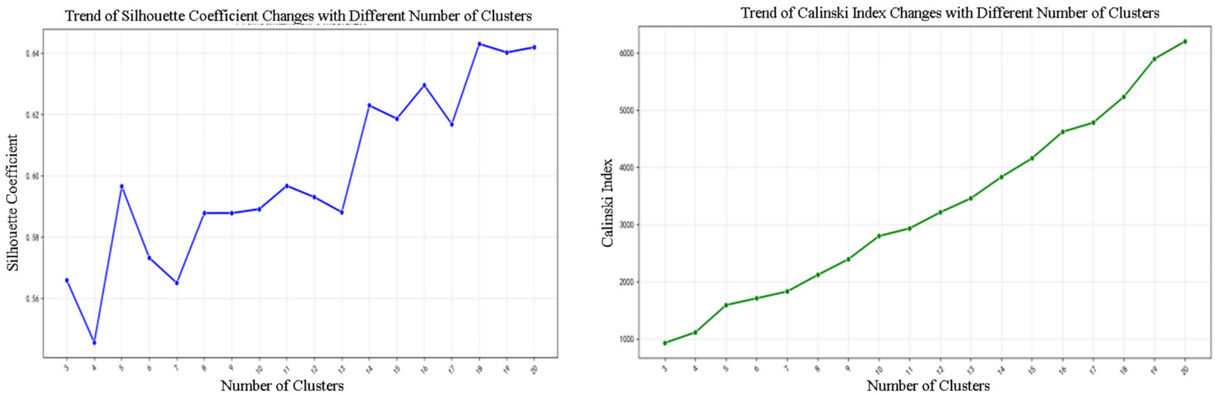

Fig. 1 Plot of the overall shape factor and variance ratio for each group

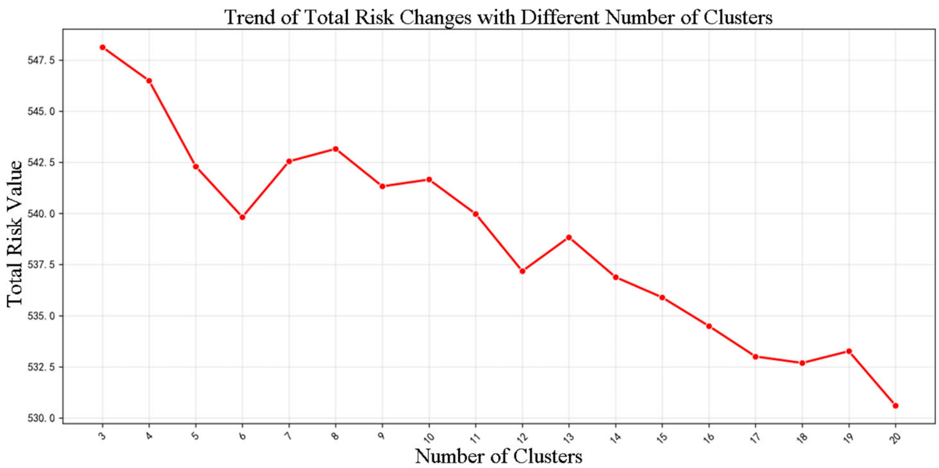

Fig. 2 Minimum total risk values for each group

When $k = 13$, the division results of each BMI interval, as well as the corresponding detection time points and risk values, are shown in Table 2.

Table 2. Information of each interval

Cluster Label	BMI Range	Sample Size	Optimal NIPT Time Point	Risk Value
0	28.0–28.7	18	17.02	1.6693
1	28.8–29.6	37	17.08	1.6779
2	29.6–30.3	32	17.36	1.7153
3	30.4–30.9	51	18.1	1.814
4	31.1–31.4	15	16.5	1.6004
5	31.6–32.2	37	17	1.6667
6	32.3–32.9	35	18.01	1.8013
7	33.0–33.6	23	16.94	1.6587
8	33.7–34.3	21	15.1	1.4133
9	34.4–34.7	17	16.92	1.6555
10	35.2–36.1	20	17.3	1.7067

It can be seen from the table that the optimal detection time points corresponding to different BMI intervals are somewhat different, which also indirectly shows that it is unreasonable to simply adopt a unified time point. Furthermore, the interval [30.4,30.9] has the largest number of samples, while the interval [36.3,37.6] has the least, indicating a certain imbalance in data distribution.

Based on the comprehensive calculation results, the total risk value is 537.184, with a corresponding per capita risk of approximately 1.695. Overall, this grouping scheme not only controls the risk but also takes into account the stability of the data structure, which can be used as a reasonable reference result.

3.4. Influence of Detection Error on Results

To investigate the influence of detection error on grouping results and risk assessment, unscreened raw data are further used for fitting, and the minimum gestational age corresponding to a Y-chromosome concentration of 0.04 is predicted for each sample. Compared with screened data, the predicted results derived from error-containing data show a more dispersed distribution, indicating that detection error increases sample fluctuation and reduces the stability of estimated critical detection time points.

On this basis, the K-means method is still adopted to group BMI values, and the total risk is calculated using the same risk function. Results show that the optimal grouping obtained from error-containing data differs significantly from that from screened data, and both the total risk and average risk increase accordingly. Further comparison reveals that the average risk derived from error-containing data is consistently higher than that from error-free data under different grouping schemes.

This indicates that detection error not only affects the boundary division of BMI groups but also shifts the optimal NIPT detection time points to a later stage overall. In other words, error amplifies the model's bias toward conservative time points, thereby raising the total risk. Therefore, in practical applications, data screening and quality control remain important prerequisites affecting the reliability of results.

4. Classification Modeling and Discrimination Based on GBDT and SMOTE

4.1. Classification Model Establishment

The goal of this section is to determine whether a female fetus is abnormal, which is essentially a binary classification problem. Since abnormal conditions are usually affected by multiple factors and there may be strong nonlinear relationships among variables, the gradient boosting decision tree (GBDT) is chosen as the classification model. Compared with traditional methods, this type of model is generally more stable when dealing with complex feature relationships and noisy data [4].

However, a prominent problem needs to be addressed before modeling. The number of abnormal samples in the original data is significantly smaller than that of normal samples (67 vs. 538), resulting in imbalanced class distribution. Direct model training would easily lead to bias toward predicting the normal class. Therefore, the SMOTE method is introduced to augment abnormal samples [5].

Specifically, new samples are generated by interpolation between minority-class samples to increase the number of abnormal cases. Its basic form can be expressed as $x_{\text{new}} = x_i + \lambda(x_i - x_{i,j})$, where $\lambda \in [0,1]$ is a random coefficient. After augmentation, the number of abnormal samples is increased to 300, making the data distribution more balanced.

After data preprocessing, a GBDT classification model is constructed. The number of decision trees is set to 300 with a maximum depth of 8, and the data are split into training and test sets at a ratio of 7:3. Let the training data be (x_i, y_i), where $y_i \in \{0,1\}$ denotes the sample class [6].

The model is trained in a stepwise additive manner, expressed as:

$$f_m(x) = f_{m-1}(x) + p \cdot h_m(x) \tag{11}$$

where $h_m(x)$ represents the fitting result of the m-th tree to the current residual, and p is the learning rate.

After multiple iterations, the final model is obtained:

$$f(x)=\sum_{m=1}^{300} p\cdot h_m(x) \tag{12}$$

To obtain classification results, the model output is transformed into probability via the sigmoid function:

$$p(x)=\frac{1}{1+e^{-f(x)}} \tag{13}$$

A sample is classified as abnormal if $p(x)>0.5$, and normal otherwise.

Overall, this model accounts for nonlinear relationships while alleviating the class imbalance problem through data augmentation, thus improving the ability to identify abnormal samples.

4.2. Model Solution

Evaluation results of the model on the test set are as follows:

Table 3. Test set evaluation results

Metric	Accuracy	Precision	Recall	F1-score	AUC-ROC	AUPRC
Score	0.6446	0.4115	0.8462	0.5537	0.7108	0.2737

According to the Table 3, the recall rate of the model reaches 0.8462, indicating that most abnormal samples can be identified, which is critical in practical applications. In contrast, the precision is 0.4115, with a certain proportion of misjudgments, meaning some normal samples are classified as abnormal.

Overall, the model tends to detect more positives rather than reduce false positives. Such a result is understandable, as the cost of missed detection is usually higher in medical testing scenarios, so appropriately raising the recall rate is acceptable.

Further analysis of feature importance yields the following results:

Table 4. Top 10 feature importance ranking

Feature	Importance
X-chromosome concentration	0.1077
GC content of chromosome 13	0.0948
GC content of chromosome 21	0.0583
Maternal BMI	0.0557
Gestational week at testing	0.0503
Number of uniquely mapped reads	0.0479
Z-score ratio (18/21)	0.0441
Z-score of chromosome 13	0.0430
Z-score of chromosome 18	0.0410
GC content	0.0378

It can be seen from the Table 4 that chromosome-related features (such as GC content, Z-score, etc.) account for relatively high weights, and maternal characteristics (such as BMI and gestational age) also exert certain influences on the results. This shows that the model integrates multiple types of information in the discrimination process rather than relying on a single indicator.

4.3. Result Analysis

Combined with the model evaluation results, this method is considered reliable in identifying abnormal samples. Especially with a high recall rate, it can effectively reduce the risk of missed detection, which is of great significance in actual testing.

From the perspective of features, X-chromosome concentration and related indicators of chromosomes 13 and 21 have considerable impacts on classification results, consistent with the focus on chromosomal abnormalities in practical testing. In addition, factors such as BMI and gestational age, although with slightly lower weights, still assist model discrimination.

For common chromosomal abnormality detection, the Z-score can be used for auxiliary judgment, whose expression is

$$Z = \frac{X - \mu}{\sigma} \tag{14}$$

In general, $|Z| \leq 3$ is regarded as the normal range. This criterion can complement the model output to improve the stability of judgment.

5. Conclusion

This paper focuses on key decision-making issues in non-invasive prenatal testing and constructs an integrated methodological framework comprising feature selection, relationship modeling, risk optimization, and classification. The results indicate that, under multifactorial nonlinear conditions, the combination of random forests and piecewise regression effectively improves fitting stability. The goodness-of-fit for most intervals exceeds 0.5, with some intervals surpassing 0.8, demonstrating the model's strong interpretability; simultaneously, the grouping scheme derived from K-means clustering and risk function optimization kept the overall risk at a low level (e.g., total risk of approximately 537.184), demonstrating excellent practical effectiveness. In the classification task, the introduction of SMOTE and GBDT models achieved a recall rate of 0.8462, significantly reducing the risk of false negatives, which holds significant implications for practical medical and public health governance scenarios. From a broader social science perspective, this study not only enhances detection efficiency at the technical level but also provides data support and methodological references for public health decision-making. Overall, the proposed model framework demonstrates good robustness and practical value in complex data environments. Future research could further explore multi-source data fusion and model adaptive optimization to enhance the method's generalization ability and application depth.

References

[1] Xing Hong, Wei Yiqiang, Li Chenlong. Handling Imbalanced Data Using the G-Mean Weighted Random Forest Algorithm [J]. Advances in Applied Mathematics, 2022, 11: 2071.

[2] Wu Yan, Zhang Huang. Assessment of IoT Network Security Status Based on the K-SMOTE Random Forest Algorithm [J]. IoT Technology, 2025, 15(24): 21-23. DOI:10.16667/j.issn.2095-1302.2025.24.004.

[3] Liu Zhenwen, He Peng. Composition Analysis and Authentication of Glass Artifacts Based on Random Forest and K-means++ [J]. Journal of Changchun University of Technology, 2025, 46(06): 545-551. DOI: 10.15923/j.cnki.cn22-1382/t.2025.6.09.

[4] Liu Siyi. Research on Improved Gradient Boosting Decision Trees and Their Interpretability [D]. Changchun University of Technology, 2025. DOI:10.27805/d.cnki.gccgy.2025.000561.

[5] Li, A., Han, M., Mu, D., et al. A Review of Classification Methods for Multi-class Imbalanced Data [J]. Research on Computer Applications, 2022, 39(12): 3534–3545. DOI: 10.19734/j.issn.1001-3695.2022.03.0198.

[6] Liu Ruiting, Xie Suli, Wang Ke, et al. Construction of a Postoperative Deep Vein Thrombosis Risk Prediction Model for Patients with Lower Limb Traumatic Fractures Based on Gradient Boosted Decision Trees [J]. Journal of Trauma Surgery, 2025, 27(07): 523-531.