Research on Optimal NIPT Timing Decision and Intelligent Fetal Abnormality Detection Based on Machine Learning and Optimization Algorithms

Houfu Zhou *

School of Artificial Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, China, 400065

* Corresponding Author Email: 2023212097@stu.cqupt.edu.cn

Abstract. With the popularization of non-invasive prenatal testing (NIPT) technology, determining a personalized optimal testing time for pregnant women with different characteristics. To address this issue, this paper proposes a comprehensive intelligent decision-making framework that integrates prediction, optimization, and classification. First, the framework constructs a quadratic multivariate nonlinear regression model to quantify the complex relationships between key factors such as maternal BMI and gestational age and the concentration of the Y chromosome in male fetuses. On this basis, it designs an optimal timing decision model aimed at minimizing potential risks for pregnant women, which is solved using the particle swarm optimization (PSO) algorithm, thus generating NIPT gestational age recommendations for groups of pregnant women in different BMI ranges. Finally, addressing the issue of detecting chromosomal aneuploidy in female fetuses, this study establishes and trains a support vector machine (SVM) binary classification model. Experimental results show that this framework can effectively formulate personalized testing plans, with the SVM classifier demonstrating superior discriminative ability on independent test sets, achieving an area under the receiver operating characteristic curve (AUC) of 0.924, thus enabling efficient and precise identification of abnormal conditions. The findings of this study provide a scientific, data-driven theoretical basis and methodological support for NIPT clinical decision-making, showcasing the application value of integrating “prediction” with “decision-making.”

Keywords: NIPT, Non-linear Regression, Particle Swarm Optimization, Random Forest, Support Vector Machine.

1. Introduction

With the shift in modern reproductive attitudes and the increasing number of advanced maternal age pregnancies, non-invasive prenatal testing (NIPT) has shown promise as a simple and low-risk screening test, leading various governments and private organizations worldwide to dedicate significant resources towards its integration into national healthcare initiatives as well as the formation of consortia and research studies aimed at standardizing its implementation[1]. In multiple countries worldwide, non-invasive prenatal screening testing (NIPT) is offered to detect chromosomal aberrations by analyzing cell-free DNA (cfDNA) circulating in the maternal bloodstream[2]. And the effectiveness of the test is closely related to the concentration of fetal cell-free DNA, particularly the concentration of the Y chromosome in male fetuses. However, clinical practice has revealed that individual factors such as maternal body mass index (BMI) and gestational age can significantly affect when this concentration reaches the detection threshold. A uniform testing window cannot meet personalized needs, which may lead to invalid results if the test is done too early or missed optimal intervention opportunities if done too late, thereby increasing potential risks for pregnant women. Therefore, constructing accurate mathematical models to determine personalized optimal timings for NIPT is a major challenge facing the field today.

Even though much research has been published on NIPT, unbiased figures concerning NIPT and first trimester screening efficacy are yet not available[3]. At present, most research on the timing of NIPT relies on clinical experience or fixed gestational age ranges, lacking refined and dynamic decision models based on large-scale data. Existing statistical analysis methods can reveal correlations between certain physiological indicators and test results, but are often limited to linear

relationship assumptions, making it difficult to capture the complex nonlinear interactions among multiple factors. In addition, most studies focus on “prediction” of risk factors and fail to effectively integrate prediction results with the “decision optimization” process—meaning they do not answer how to plan test timing for specific populations to minimize overall risk under a certain success rate. So adequate genetic pre- and post-test counseling is mandatory to ensure an informed consent of the patients as the basis for all genetic testing strategies and furthermore, to respect the right not to know if desired[4]. These limitations and important notes make current NIPT strategies noticeably insufficient in terms of individualization and precision, making it difficult to meet the growing demands for refined pregnancy management.

To address these research gaps, the core contribution of this paper lies in constructing a closed-loop intelligent analysis framework that spans from prediction to decision-making. The key feature of this framework is that it does not handle prediction or classification tasks in isolation; rather, it organically integrates a multivariate nonlinear regression prediction model, a particle swarm optimization-based risk minimization decision model, and a support vector machine-based anomaly classification model. In this way, for the first time, this study achieves a data-driven, full-process decision-making system: quantifying the influence of critical physiological indicators, dynamically optimizing testing time points, and intelligently identifying abnormal states. This methodological innovation not only enhances the scientific basis and level of individualization in choosing NIPT timing but also provides an effective approach to tackling high-precision medical classification challenges involving imbalanced data.

2. Methods

2.1. Data Preprocessing and Feature Engineering

First of all, the data comes from https://www.mcm.edu.cn. To ensure the quality and effectiveness of model inputs, this study first conducted systematic preprocessing of the raw data. This process included identifying and handling missing values and outliers. Outlier detection is a widely used technique for identifying anomalous or exceptional events across various contexts[5]. Outliers were detected using the boxplot method (IQR method) and replaced with the mean value. To eliminate differences in scale among various indicators, all continuous numerical features were standardized. During the feature selection stage, considering that some physiological indicators may not follow a normal distribution, we used Spearman's correlation coefficient to measure the strength of the monotonic relationship between each candidate indicator and the target variable (e.g., Y chromosome concentration). By setting a significance threshold, we screened out the subset of features most strongly correlated with the target variable and removed redundant variables with high multicollinearity, thus constructing a streamlined and efficient feature space that lays a solid foundation for subsequent modeling.

2.2. Y-Chromosome Concentration Prediction Model

To accurately capture the complex nonlinear relationships between multiple independent variables such as maternal age, gestational weeks at testing, BMI, and the dependent variable Y chromosome concentration, this study constructed a quadratic multivariate nonlinear regression model. This model not only considers the linear term of each variable but also innovatively introduces a quadratic term of maternal BMI to capture its nonlinear impact on Y chromosome concentration. Additionally, to interpret possible interaction effects between different sequencing quality indicators, the model includes an interaction term between the "proportion of duplicate reads" and the "proportion aligned to the reference genome". The specific mathematical expression of this model corresponds to Equation (1) in the original text.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 {x_3}^2 + \beta_6 x_5 \cdot x_4 + \mu \quad (1)$$

With this model, we are able to quantitatively analyze the specific patterns and degrees of influence of key factors on Y chromosome concentration, providing essential predictive inputs for subsequent risk assessment and timing decisions.

2.3. Optimal NIPT Timing Decision Model

To recommend the optimal NIPT testing time point with minimal risk for pregnant women with different characteristics, this paper abstracted the issue as a complex optimization problem and constructed a risk-minimization decision model [6]. The model uses a comprehensive risk index as its core objective function, which takes into account both the risk arising from delayed testing and the uncertainty caused by the variability of optimal timing among groups of pregnant women. The decision variables of the model include the number of BMI categories, the interval boundaries of each BMI group, and the corresponding optimal NIPT testing time point for each interval. To ensure clinical feasibility, the model incorporates multiple constraints, such as guaranteeing a minimum sample size for each group, ensuring reasonable intervals between testing times for different groups, and securing a minimum qualifying rate for testing in all groups. Given the highly nonlinear and multi-constrained nature of this optimization problem, which makes it challenging for traditional solvers to handle efficiently, this study employs the Particle Swarm Optimization (PSO) algorithm for model solving [7]. The detailed mathematical formulation of this model is given in Equations (2) and (3) of the original text, with part of the objective function's weighting coefficients objectively assigned using the random forest algorithm to enhance the scientific rigor of the model [8].

$$\begin{gathered} v_i^d = \beta v_i^{d-1} + c_1 r_1 \left(qbest_i^d - x_i^d\right) + c_2 r_2 (gbest^d - x_i^d) \\ x_i^{d+1} = x_i^d + v_i^d \end{gathered} \tag{2}$$

$$\begin{gathered} \min f = f_1 + \theta_1 \cdot f_2 - \theta_2 \cdot f_3 + \theta_3 \cdot f_4 + f_5 \\ s.t. \begin{cases} b_1 < b_2 < \cdots < b_{m-1} \\ BMI_{min} \le b_1, b_2, \ldots b_{m-1} \le BMI_{max} \\ \sum_{j=1}^{m} x_{ij} = 1, \forall i = 1,2, \ldots, n \\ \sum_{n=1}^{n} x_{ij} \ge Q_{min}, \forall j = 1,2, \ldots, m \\ t_{j+1} - t_j \ge t_\Delta, \forall j = 1,2, \ldots, m-1 \\ \sum_{i=1}^{n} x_{ij} \cdot P(t_i \le t_j) \ge \sum_{i=1}^{n} x_{ij} \cdot a, \forall j = 1,2, \ldots, m \end{cases} \end{gathered} \tag{3}$$

2.4. Intelligent Classification Model for Female Fetal Chromosomal Abnormality

To address the problem of determining female fetal chromosomal aneuploidy abnormalities, this study constructed a binary classification machine learning model. The core task is to accurately distinguish between "abnormal" and "normal" samples based on physiological indicators of pregnant women, along with high-dimensional features such as the Z-scores for chromosomes 21, 18, 13, GC content, read count, and so on[9]. Given the high dimensionality and possible nonlinear separability of the data, we selected the Support Vector Machine (SVM) algorithm, which performs well with such problems. To handle nonlinear classification boundaries, the model uses kernel functions to map the original feature data into a higher-dimensional space, thereby achieving linear separability. The model's ultimate goal is to identify a maximum-margin hyperplane that optimally separates the two classes[10]. The detailed mathematical derivation and decision function for this model correspond to Equations (4) through (5) in the original text. With this model, automated, high-precision intelligent classification of female fetal abnormalities can be achieved.

$$max \sum_{i=1}^{n} \pi_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i \pi_j y_i y_j k(x_i, x_j) \tag{4}$$

$$f(x) = sign(\sum_{i=1}^{n} \pi_i y_i k(x_i, x) + b) \tag{5}$$

3. Results

3.1. Analysis of Influencing Factors and Validation of the Y-Chromosome Concentration Model

In order to quantify the influence of various factors on Y chromosome concentration and to validate the effectiveness of our predictive model, we conducted a detailed analysis of the model's fitting results. First, we used Spearman correlation analysis to identify the core predictive factors that are significantly correlated with Y chromosome concentration[11].

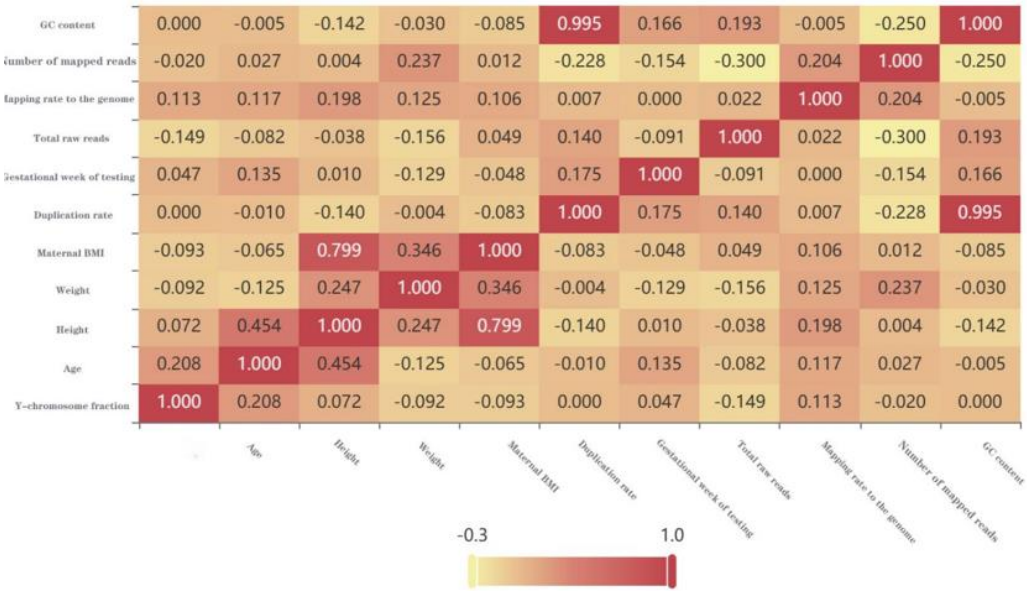

Figure 1. Heatmap of Spearman coefficients for each variable.

As shown in Figure 1, the heatmap visually presents the correlation matrix among the variables. The analysis indicates that physiological indicators such as maternal BMI and gestational age at testing show significant nonlinear monotonic relationships with Y chromosome concentration, whereas some sequencing quality indicators, such as GC content, have weaker correlations. This provides data support for building a nonlinear regression model and selecting key variables. Next, we conducted significance testing of the regression model [12].

Table 1. t-test results of the regression model.

Variable	Coefficient	Standard error	t-statistic	p-value	statistical significance
Constant term	0.34679	0.06068	5.71437	1.42491	***
x_1	-0.13450	0.03484	-3.85984	0.00012	***
x_2	0.10531	0.02588	4.06897	5.06731	***
x_3	0.44550	0.22980	1.93863	0.05280	None
x_4	0.19935	0.04424	4.50581	7.33544	***
${x_3}^2$	-0.77603	0.22812	-3.40179	0.00069	***
$x_3 \cdot x_4$	-0.18672	0.06134	-3.04367	0.00239	***

As shown in Table 1, the analysis shows that the overall F-test result of the model is highly significant statistically ($p < 0.001$), demonstrating the model's overall validity. From the results of the t-tests for each variable, the coefficients of gestational age at testing, both linear and quadratic terms of maternal BMI, as well as key interaction terms, are all significant at the 99% confidence level. This reveals that Y chromosome concentration increases with gestational age and exhibits a quadratic relationship with BMI—first increasing and then decreasing—confirming the necessity of including the quadratic term for BMI in the model. This finding profoundly reveals the complex mechanisms by which physiological factors influence detection indicators and serves as a theoretical basis for subsequent precise timing decisions.

3.2. Analysis of Optimal NIPT Timing Decision Based on Risk Minimization

Based on the established predictive model, we applied the particle swarm optimization algorithm to solve the risk minimization model and obtained the optimal NIPT testing time scheme for pregnant women in different BMI groups.

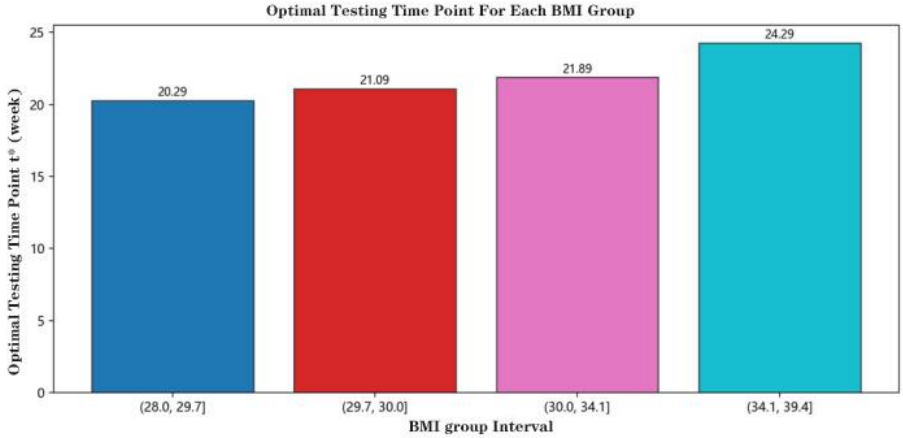

Figure 2. Bar chart of optimal testing times for each BMI group

As shown in Figure 2, the model divides the high BMI pregnant population into four distinct risk groups. Analysis of the chart clearly shows that as the BMI range increases, the recommended optimal NIPT testing time exhibits a noticeable delay. For example, for pregnant women with a BMI in the [28, 29.7] range, the optimal testing time is approximately 20.29 weeks, while for the ultra-high BMI range of [34.1, 39.4], the testing time is delayed to 24.29 weeks. This result deeply demonstrates that the model successfully captures the inhibitory effect of high BMI on the release of fetal cell-free DNA and compensates for this impact by postponing the testing time to ensure sufficient Y chromosome concentration. This validates the algorithm's advantage in balancing risk and testing success rate. In practical application, this means that stratified BMI management and differentiated testing strategies for pregnant women is an effective way to reduce overall potential risk.

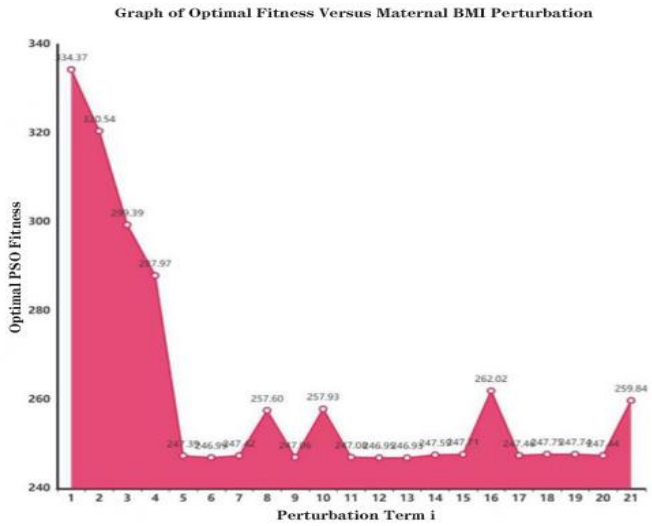

Figure 3. Sensitivity analysis results chart.

To evaluate the model's robustness to data errors, we conducted a sensitivity analysis[13]. As shown in Figure 3, after applying different degrees of perturbation to the input variable BMI, the model's output of the minimum risk value (optimal fitness) exhibited significant fluctuations. This indicates that the model's final decision results are relatively sensitive to measurement errors. In clinical applications, this means it is crucial to ensure the accuracy of input data and also suggests that there is room to further improve the model's robustness when facing data uncertainty.

3.3. Analysis of the Multi-Factor Comprehensive Decision Model

In order to further refine the accuracy of decision-making, we constructed a comprehensive decision model that takes into account more factors, such as height, weight, and age, and analyzed its results.

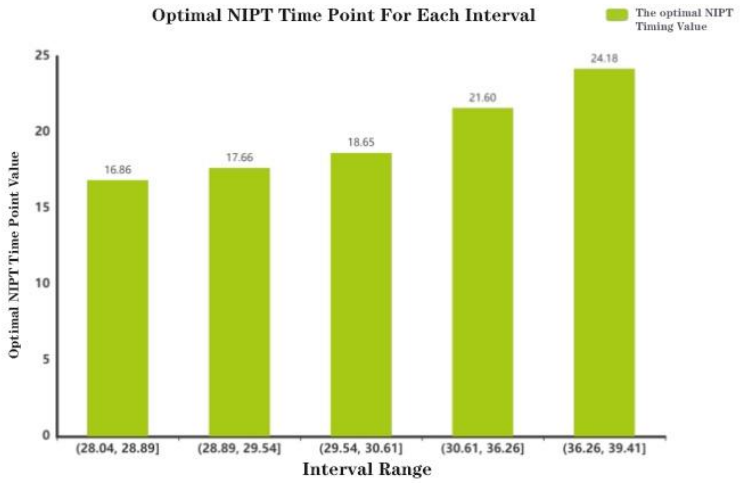

Figure 4. Best NIPT by interval for the multi-factor model.

As shown in Figure 4, after incorporating more dimensions of individual physiological characteristics, the model divides the population of pregnant women into five groups. Compared to the previous model, this model features more detailed BMI grouping intervals and a wider distribution range for the optimal timing of NIPT, spanning from a minimum of 16.86 weeks to a maximum of 24.18 weeks. This change reveals the complex interaction between factors such as age, height, and BMI. The comprehensive model can capture this individual variability with greater precision. For example, at similar BMI levels, pregnant women of different ages may be assigned to different optimal testing windows. This validates the comprehensive model's superiority in achieving highly personalized decision-making: it no longer relies solely on a single indicator, but instead constructs a multi-dimensional risk assessment profile, thereby providing each pregnant woman with a more tailored testing plan.

3.4. Performance Evaluation of the SVM Classification Model for Female Fetal Chromosomal Abnormality

For the intelligent classification task of chromosomal abnormalities in female fetuses, we conducted a comprehensive evaluation of the performance of the trained SVM model on an independent test set.

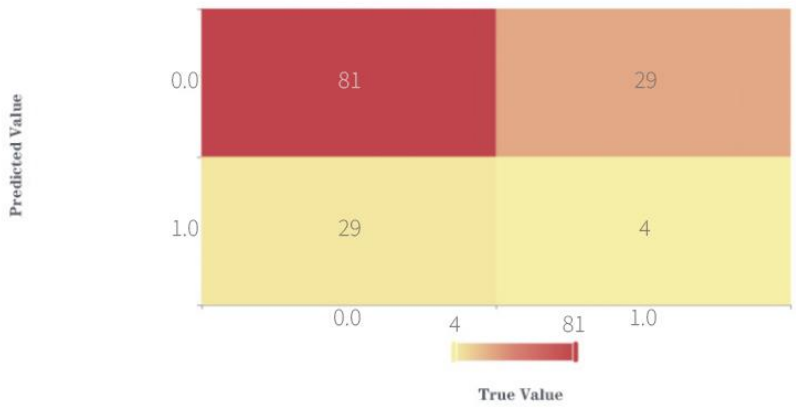

Figure 5. Heatmap of the confusion matrix for the test set

As shown in the confusion matrix in Figure 5, the model demonstrates robust performance in distinguishing normal samples (negative class), achieving a specificity of 73.64%. However, its ability to identify abnormal samples (positive class) is relatively limited, with a recall rate of only 36.36%. This indicates a considerable number of false negatives. This result vividly highlights the challenge the model faces when handling severely imbalanced datasets (where normal samples greatly outnumber abnormal ones). In clinical applications, high specificity means that the model can effectively avoid false positives in normal pregnant women, but the low recall rate indicates a risk of missed diagnoses. This is a key issue that needs to be addressed in future model optimization.

Figure 6. ROC curve for the training set.

To evaluate the model's overall discrimination capability, we plotted its ROC curve as shown in Figure 6. The curve rises steeply towards the top-left corner, and the area under the curve (AUC) reached 0.924 on the test set. Such a high AUC indicates that, although the recall rate at a specific threshold is not high, the model overall possesses excellent potential for distinguishing between positive and negative samples. It can achieve a good balance at different decision thresholds, effectively differentiating between healthy fetuses and those at risk of chromosomal abnormalities. This validates the powerful ability of the SVM algorithm to learn effective discriminative patterns from high-dimensional complex features and demonstrates the great application value of this model as an efficient screening tool.

4. Conclusions

This paper successfully designs and implements an intelligent decision support framework that integrates prediction, optimization, and classification to address key clinical issues in non-invasive prenatal testing (NIPT). The core contribution of the research lies in elevating data analysis from single-dimension "prediction" to comprehensive "decision-making" by organically combining algorithms such as multiple nonlinear regression, particle swarm optimization, and support vector machines. This provides scientifically based, quantitative, and personalized optimal NIPT testing timing solutions for pregnant women with different physiological characteristics and achieves highly accurate intelligent identification of fetal chromosomal abnormalities.

Future research may further refine and improve the model in several aspects. At the algorithmic level, more advanced models such as gradient boosted decision trees (GBDT) or deep learning could be explored to replace the current regression and classification algorithms, aiming to capture deeper data patterns and enhance handling of class imbalance issues. In terms of optimization, multi-objective optimization algorithms could be introduced for greater flexibility in balancing multiple potentially conflicting goals in clinical practice (e.g., risk, cost, time, etc.). Finally, future studies should aim to integrate larger-scale and more multi-center clinical datasets, and incorporate multi-omics information such as genomics, to build a more comprehensive, dynamic, and robust intelligent risk management system for pregnancy.

References

[1] Jayashankar S S, Nasaruddin M L, Hassan M F, et al. Non-invasive prenatal testing (NIPT): reliability, challenges, and future directions[J]. Diagnostics, 2023, 13(15): 2570.

[2] van Eekhout J C A, Bekker M N, Bax C J, et al. Non-invasive prenatal testing (NIPT) in twin pregnancies affected by early single fetal demise: A systematic review of NIPT and vanishing twins[J]. Prenatal Diagnosis, 2023, 43(7): 829-837.

[3] Liehr T. False-positives and false-negatives in non-invasive prenatal testing (NIPT): what can we learn from a meta-analyses on> 750,000 tests?[J]. Molecular Cytogenetics, 2022, 15(1): 36.

[4] Bedei I, Wolter A, Weber A, et al. Chances and challenges of new genetic screening technologies (NIPT) in prenatal medicine from a clinical perspective: a narrative review[J]. Genes, 2021, 12(4): 501.

[5] Mazarei A, Sousa R, Mendes-Moreira J, et al. Online boxplot derived outlier detection[J]. International journal of data science and analytics, 2025, 19(1): 83-97.

[6] Hu Y, Zhang L. Achieving long-term fairness in sequential decision making[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(9): 9549-9557.

[7] Abualigah L. Particle Swarm Optimization: Advances, Applications, and Experimental Insights[J]. Computers, Materials & Continua, 2025, 82(2).

[8] Salman H A, Kalakech A, Steiti A. Random forest algorithm overview[J]. Babylonian Journal of Machine Learning, 2024, 2024: 69-79.

[9] Kumari R, Nigam A, Pushkar S. An efficient combination of quadruple biomarkers in binary classification using ensemble machine learning technique for early onset of Alzheimer disease[J]. Neural Computing and Applications, 2022, 34(14): 11865-11884.

[10] Roy A, Chakraborty S. Support vector machine in structural reliability analysis: A review[J]. Reliability Engineering & System Safety, 2023, 233: 109126.

[11] Alsaqr A M. Remarks on the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data[J]. African Vision and Eye Health, 2021, 80(1): 10.

[12] Muff S, Nilsen E B, O'Hara R B, et al. Rewriting results sections in the language of evidence[J]. Trends in ecology & evolution, 2022, 37(3): 203-210.

[13] Pirsaheb M, Hadei M, Sharafi K. Human health risk assessment by Monte Carlo simulation method for heavy metals of commonly consumed cereals in Iran-Uncertainty and sensitivity analysis[J]. Journal of Food Composition and Analysis, 2021, 96: 103697.