An Explainable PSO-XGBoost Algorithm Framework for Concentration Prediction with High-Dimensional Data

Ya'nan Gao

School of International Business, Tianjin Foreign Studies University, Tianjin, China

gaoyanan339@163.com

Keywords: NIPT; Spearman's rank correlation; PSO-XGBoost; SHAP; Prediction of fetal chromosomal concentration

Abstract: Noninvasive Prenatal Testing (NIPT) serves as a crucial tool for prenatal screening. Given the high-dimensional, heterogeneous, and nonlinear characteristics of NIPT data, an analytical framework that balances accuracy with interpretability is essential for effective prenatal screening. We propose an interpretable PSO-XGBoost framework that integrates Spearman correlation for feature screening, PSO-based hyperparameter optimization, and SHAP analysis to predict fetal Y-chromosome concentration. Experimental results demonstrate a significant positive correlation between fetal Y chromosome concentration and gestational age. The PSO-XGBoost model achieved an R^2 value of 0.958, indicating that the model exhibits high accuracy and robust stability. SHAP analysis further reveals that model predictions are primarily driven by core features such as X chromosome concentration, Y chromosome Z-score, and gestational age, with significant nonlinear interactions and individual variation present. Future integration of multimodal data could further improve the precision of prenatal diagnosis and clinical decision-making.

1. Introduction

With the rapid advancement of gene sequencing and biotechnology, non-invasive prenatal testing (NIPT) has become a significant tool in prenatal screening. Accurate prediction of fetal chromosomal concentration is central to NIPT and directly affects diagnostic sensitivity and specificity. However, NIPT testing data typically exhibits characteristics such as high dimensionality, strong sample heterogeneity, and complex nonlinear relationships between features. Traditional linear or parametric methods often underperform on such data because of rigid assumptions, leading to biased predictions. Therefore, a framework that couples high predictive accuracy with strong interpretability can enhance prenatal screening and advance data-driven precision medicine.

In exploring fundamental biological correlations, non-parametric statistical methods such as Spearman's rank correlation provide preliminary evidence for understanding the influence of key variables. Yang et al.[1] employed Spearman analysis to investigate the relationship between fetal free DNA (FF) concentration and Z-score in NIPT results, revealing a significant positive correlation between the two in positive samples. Kim et al.[2] employed Spearman test to assess the relationship between FF and indicators of placental function, finding that low FF was significantly negatively correlated with multiple adverse pregnancy outcomes associated with placental dysfunction. Hanxiao et al.[3] employed Spearman's rank correlation analysis model to compute rank correlations for pedigree SNP (single nucleotide polymorphism) data in order to infer fetal haplotypes. The results demonstrated that the rank correlation model achieved a high positive predictive rate in paternal haplotype prediction. Duarte-Delgado et al.[4] employed Spearman test to assess the relationship between cytokine levels and clinical or haematological indicators, revealing that multiple cytokines demonstrated significant correlations with disease activity measures. Wu et al.[5] employed Spearman to investigate the effects of periodontitis on gut microbiota and faecal metabolites. Findings revealed that periodontitis significantly altered both the composition of gut microbiota and the faecal metabolite profile, with a marked correlation observed between gut microbiota and metabolites.

To enhance direct predictive capabilities for high-dimensional non-linear data, researchers have

  DOI: 10.25236/iwmecs.2025.005

adopted a strategy combining ensemble optimisation algorithms with advanced machine learning models. Cao et al.[6] employed multi-feature selection combined with PSO to optimise XGBoost for constructing a cardiovascular disease prediction model. Results demonstrated significant improvements in metrics such as accuracy and AUC compared to the unoptimised baseline. Dias Júnior et al.[7] employed a hybrid model combining deep convolutional features with PSO-XGBoost for the automated classification of COVID-19 patients based on chest X-ray images. Results demonstrated that PSO-XGBoost outperformed the reference classifier across multiple metrics. Tseng et al.[8] integrated image segmentation with the PSO-XGBoost algorithm to construct an MRI brain tumour detection model, achieving an accuracy rate of 99.42%, significantly outperforming the comparison model. Radhakrishnan et al.[9] employed noise processing and stacked machine learning models to achieve automatic sleep staging in wearable devices. The results demonstrated an accuracy of 98.42% on public datasets, representing an outstanding performance. Zhou et al.[10] employed PSO to refine the hyperparameters of XGBoost. The results demonstrated an R^2 exceeding 0.98, with predictive accuracy and stability significantly surpassing conventional methodologies.

To ensure transparency and trustworthiness in clinical decision-making for such complex models, interpretability frameworks such as SHAP have proven effective in quantifying feature contributions. Allgaier et al.[11] evaluated the interpretability of SHAP in medical machine learning models through a systematic review, demonstrating that SHAP effectively quantifies feature contributions and enhances the credibility and transparency of clinical decision-making. Vimbi et al.[12] conducted a systematic review comparing the interpretability of SHAP and LIME in the early detection of Alzheimer's disease. They confirmed that SHAP's global explanations outperform LIME's local analysis, thereby effectively enhancing the clinical credibility of the model. Luo et al.[13] employed the SHAP framework for feature selection and model interpretation to construct a predictive model for one-year readmission risk in elderly heart failure patients. The resulting model achieved an AUC of 0.87, significantly enhancing both interpretability and practical utility. Xu et al.[14] employed machine learning and SHAP to construct a predictive model for feeding intolerance in preterm infants. By quantifying feature contributions, they accurately identified core risk factors such as birth weight and feeding method, significantly enhancing the interpretability of clinical decision-making. Fan et al.[15] employed XGBoost and SHAP to construct an interpretable diagnostic model for knee osteoarthritis, achieving an AUC of 0.94. Through SHAP visualisation, they elucidated the specific contributions of key risk factors, which including BMI and knee joint injury—to diagnostic decision-making. Lugner et al.[16] employed machine learning and the SHAP to analyse UK Biobank data, precisely identifying the ten key predictors of type 2 diabetes mellitus through SHAP.

The organisational logic of the remainder of this paper is as follows: In Chapter Two, we shall systematically elaborate upon the constructed interpretable PSO-XGBoost algorithmic framework. This encompasses a feature selection method based on Spearman's rank correlation, the optimisation of XGBoost hyperparameters via PSO, the principles and workflow for constructing the PSO-XGBoost model, and the establishment of a feature importance analysis framework utilising SHAP. In Chapter Three, we shall conduct experimental validation of the analytical framework constructed. Firstly, the data sources are specified, pre-processing is conducted, and the distribution characteristics of variables are described. Secondly, Spearman's correlation analysis is performed to assess the relationship between each indicator and the Y chromosome. Subsequently, the predictive performance of the PSO-XGBoost model is evaluated on the test set, with comparative analysis against the unoptimised XGBoost model. Finally, by integrating the SHAP method to analyse feature contribution in model predictions, the paper examines the influence mechanisms of each feature on Y chromosome concentration prediction from both individual sample and overall distribution perspectives. This enables a comprehensive evaluation of the proposed framework's efficacy and interpretability.

2. Method

2.1 Spearman's Correlation Analysis of Characteristics

Spearman's correlation is a non-parametric measure of a monotonic association between two variables based on the Pearson correlation of their ranks. It is based on the rank correlation coefficient and possesses a significant advantage in that it imposes no requirements on the characteristic distribution of variables, is insensitive to outliers, and can handle non-linear relationships. Accordingly, the Spearman coefficient is a rank-based measure that is robust to outliers and non-normality, suitable for continuous variables and data that do not necessarily follow a normal distribution. The Spearman correlation coefficient r_s is given by formula (1).

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{1}$$

Where n represents the number of observations, and d_i represents the rank difference between $R(x_i)$ and $R(y_i)$. The rank difference of a number is the position it occupies after the numbers in its column have been sorted in ascending order.

The Spearman correlation coefficient ranges from -1 to 1. The closer the Spearman correlation coefficient is to 0, the weaker the relationship between the two variables; the closer its absolute value is to 1, the stronger this relationship. A negative values indicate negative correlation, while positive values indicate positive correlation.

2.2 Chromosome Concentration Prediction Based on PSO-XGBoost

2.2.1 Principles of XGBoost

XGBoost is an ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDT). It iteratively trains decision trees on the residuals of prior models and combines them into a weighted ensemble. Key advantages include L1/L2 regularisation to mitigate overfitting and support for parallel computation. This enables effective handling of complex nonlinear relationships and demonstrates robust fitting capabilities. It demonstrates particularly outstanding performance in scenarios involving high-dimensional data and numerous features. During XGBoost operation, the construction of each decision tree is associated with the stepwise minimisation of the loss function. Assuming the current model is f, the next step aims to minimise the following objective function:

$$\min_{\theta} \sum_{i=1}^{n} L\left(y_i, F_m(x_i) + f(x_i;\theta)\right) \tag{2}$$

Here, L denotes the loss function; y_i represents the actual value; $F_m(x_i)$ signifies the current model's predicted value; and $f(x_i;\theta)$ indicates the newly incorporated base learner. XGBoost approximates the loss function using a second-order Taylor expansion to optimise the objective function:

$$L(y_i, F_m(x_i) + f(x_i;\theta)) \approx L(y_i, F_m(x_i)) + g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2 \tag{3}$$

Where g_i denotes the first derivative; h_i represents the second derivative.

2.2.2 PSO(Particle Swarm Optimisation)

Particle Swarm Optimisation (PSO) is an optimisation algorithm inspired by the foraging behaviour of flocks of birds, which employs collective search to find optimal solutions. Each particle updates its position using its personal best and the global best, gradually converging to an optimal parameter set. In PSO, each solution is regarded as a 'particle', which explores the optimal solution by continuously adjusting its position within the search space. The movement of particles is influenced by their own historical optimal positions and the optimal positions of other particles within the swarm. Based on this information, they adjust their velocity and position to progressively

approach the global optimum solution. Each particle possesses two crucial attributes: position and velocity. The position update and velocity update formulas in PSO are shown in equations (4) and (5) respectively:

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \tag{4}$$

$$v_i^{(t+1)} = wv_i^{(t)} + c_1 r_1 (p_i^{(t)} - x_i^{(t)}) + c_2 r_2 (g^{(t)} - x_i^{(t)}) \tag{5}$$

Where $x_i^{(t)}$ denotes the position of particle i at time t; $v_i^{(t+1)}$ indicates the velocity of particle i at time $t+1$; $v_i^{(t)}$ represents the velocity of particle i at time t; $p_i^{(t)}$ represents the historical optimal position of particle i at time t; $g^{(t)}$ signifies the global optimal position within the swarm; w implies the inertia weight; c_1 and c_2 denote acceleration constants; r_1 and r_2 denote random numbers generated within the interval [0,1].

In summary, the main steps of the traditional PSO algorithm are as follows:

Step 1: Initialisation. Set the population size, initial particle positions and velocities, and parameters w、c_1、c_2, and t;

Step 2: Evaluate individuals. Calculate the fitness of each particle in the population;

Step 3: Update the particle's individual best position $p_i^{(t)}$ and the global best position $g^{(t)}$;

Step 4: Update the particle's velocity and position respectively using the given velocity formula and position formula;

Step 5: Determine whether the termination condition is satisfied. If satisfied, output the global optimum solution; otherwise, proceed to Step 2 and repeat Steps 2 to 5.

2.2.3 PSO-XGBoost model

1) Initialise particles

First, initialise the position and velocity of the particles, where each particle is represented as a set of parameters comprising the $subsample$ and min_hild_eight values from the XGBoost model. The position and velocity of particles in PSO are as follows:

$$x_i = (subsample_i, min\_child\_weight_i) \tag{6}$$

$$v_i = (v_{subsample}, v_{min\_child\_weight}) \tag{7}$$

2) Evaluation of Fitness Functions

The position of each particle (i.e., the hyperparameter combination) is used to train the XGBoost, with its fitness assessed by calculating the model's R^2 value. The objective is to maximise R^2, hence the fitness function $f(x_i)$ is defined as:

$$f(x_i) = R^2 = 1 - \frac{\sum_{i=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{i=1}^{n}(y_j - \bar{y})^2} \tag{8}$$

Here, y_i denotes the true value of Y chromosome concentration in i sample; $\hat{y}_j$ represents the predicted value based on the XGBoost model; $\bar{y}$ signifies the mean value of Y chromosome concentration; and n indicates the number of samples.

3) Update the position and velocity of particles

Update the position and velocity of each particle according to the above formulae (4) and (5).

4) Update the global optimum solution

For each particle, if its fitness value R^2 is superior to that of the current global optimum solution, the global optimum solution is updated. That is, the global optimum solution is the particle position with the highest R^2 value.

5) Stop condition

When the positional changes of all particles within the particle swarm become negligible and the fitness R^2 approaches stability, the algorithm may be deemed to have converged. The parameters at this juncture constitute the optimal parameters $(subsample_{best}, min\_child\_weight_{best})$.

After selecting the optimal parameters, serialise the model (e.g., with pickle) for subsequent

evaluation and use.

2.3 Feature Importance Analysis Based on SHAP

SHAP is a post-hoc interpretability method derived from Shapley values in cooperative game theory. Its core methodology borrows from game theory used to calculate the relative contributions of different participants within a coalition. It can be used to quantify the marginal contribution of each input feature in predicting a single sample, combining both global and local interpretability. SHAP can decompose a model's prediction output into the cumulative effects of individual features, thereby assigning importance scores to each feature within the model. By evaluating the SHAP values of each input variable, we measure their contribution to the predicted value, thereby enabling a quantitative attribution analysis of the model's prediction results. The greater the SHAP value, the greater the contribution of the input feature to the predicted value. The SHAP interpretability method is used to provide explanations for trained models. The formula for calculating the SHAP value corresponding to the prediction factor is:

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{M!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{9}$$

Where M denotes the set of all predictive factors; S denotes any subset of the set of predictive factors that excludes factor i; $S \cup \{i\}$ denotes any subset of the predictive factors that includes factor i; $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ denotes the result fitted from the set of predictive factors $S \cup \{i\}$; $f_S(x_S)$ denotes the result fitted from the set of predictive factors S.

Unlike traditional feature importance assessment methods (such as tree-based Gain or Split), SHAP not only ranks the overall contribution of features but also reveals, at the level of each individual sample, the specific direction and magnitude of how the value of a feature influences the model's output. It is currently one of the mainstream tools for explaining black-box models.

3. Experiment

3.1 Data description

This paper employs the NIPT-related data provided in https://www.mcm.edu.cn/upload_cn/node/759/SvpohSGacdffe718bcaa3b6e835c03ae3461cab1.zip to validate the performance of the proposed method. This dataset comprises two sub-datasets: 'male foetus detection data' and 'female foetus detection data'. The present study utilises the 'male foetus detection dataset', which contains 1,082 samples and 31 variables.

Our analysis focuses on associations between Y-chromosome concentration and predictors such as gestational age and maternal BMI. Given that male foetuses possess an XY sex chromosome configuration, the X and Y chromosomes exhibit a strong association, and GC content serves as a crucial indicator in assessing sequencing data quality. Therefore, the data in this paper retains only variables relevant to the research objectives, including the pregnant woman's code, age, height, weight, detection of gestational age, pregnant woman’s BMI, GC content, X chromosome Z-value, Y chromosome Z-value, X chromosome concentration, and Y chromosome concentration. All other irrelevant characteristic variables have been removed. The 11 variables are described in detail as shown in Table 1.

In accordance with the reliability standards for NIPT testing, this paper combines data from pregnant women with identical maternal codes based on the Y chromosome threshold concentration, retaining only data meeting the threshold. For outliers in GC content, data exceeding the normal range were excluded. However, considering that minor fluctuations may occur in the data and values close to the normal range still hold reference value, GC data approaching the standard range were retained to ensure data integrity and representativeness. Finally, we converted gestational age to a numeric format and imputed missing values using the mean. Identify outliers by plotting box plots and removing data points that are clearly anomalous. To illustrate the distribution of each indicator more

clearly, box plots have been constructed for each metric, as shown in Figure 1.

Table 1 Specific Descriptions of Variables

Sequence	Variable name	Variable Specification	Type
01	Pregnant Woman Code	The same pregnancy code refers to the same pregnant woman.	Text
02	Age	Age of pregnant women	Integer
03	Height	Height of pregnant women	Integer
04	Weight	Weight of pregnant women	floating-point number
05	Detection of gestational age	Gestational age at the time of this examination (weeks + days)	Text
06	pregnant woman's BMI	BMI Index of pregnant women	floating-point number
07	GC content	The proportion of the bases G (guanine) and C (cytosine) within a sequence constitutes a crucial metric for assessing sequencing data quality. The normal GC content range is 40% to 60%. An excessively high or low GC content, or an abnormal distribution, may indicate issues with sequencing quality.	floating-point number
08	X chromosome Z-value	Z-value of the X chromosome	floating-point number
09	Y chromosome Z-value	Z-value of the Y chromosome	floating-point number
10	X chromosome concentration	X chromosome concentration (whose value is estimated through bioinformatics analysis of data under certain assumptions and may yield negative values)	floating-point number
11	Y chromosome concentration	The proportion of free Y-chromosome DNA fragments	floating-point number

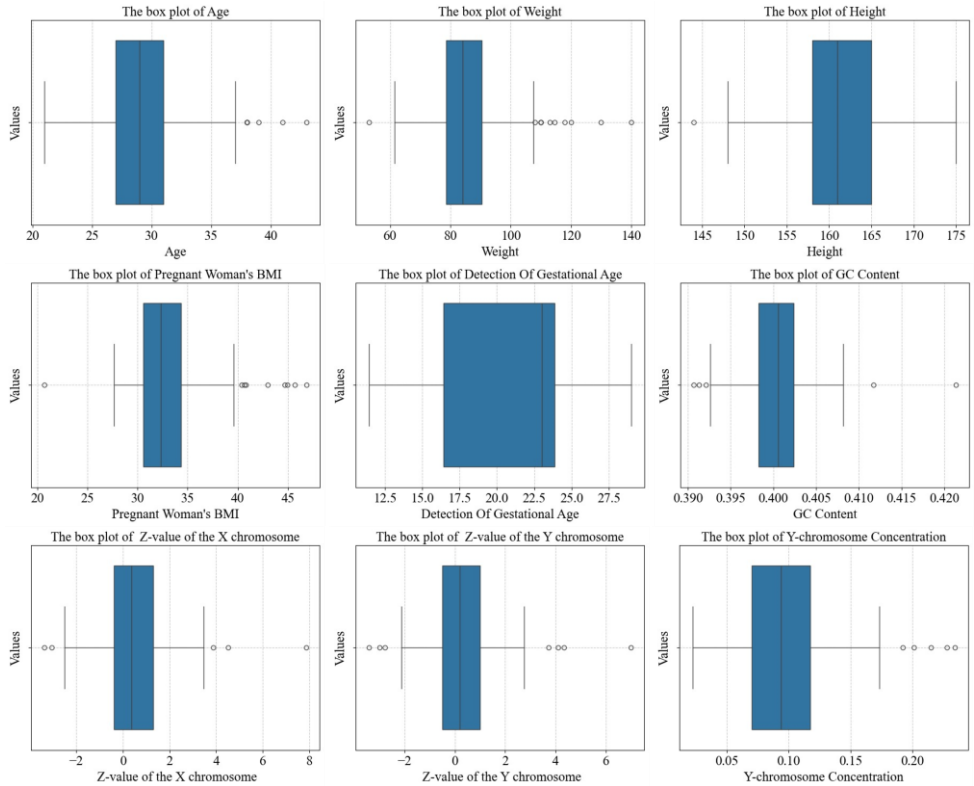

Figure 1 Box Plots for Various Indicators

Figure 1 results indicate that Age, Weight, BMI, GC content, Z-value of the X chromosome, Z-value of the Y chromosome, and Y-chromosome Concentration exhibit relatively symmetrical distributions with a small number of outliers, reflecting moderate variability. Among these, BMI, GC content, Z-value of the X chromosome, Z-value of the Y chromosome, and Y-chromosome Concentration show more concentrated distributions. Height exhibits a slight right skew with one outlier on the left, indicating a more dispersed distribution. Detection of Gestational Age shows a pronounced right skew and the widest distribution range, though no outliers are present. The outliers identified in the figure were subsequently removed.

3.2 Results of Spearman's correlation analysis

To analyse the correlation between fetal Y chromosome concentration and maternal gestational age, BMI, and other indicators, a Spearman correlation matrix was constructed, and a correlation diagram was plotted as shown in Figure 2.

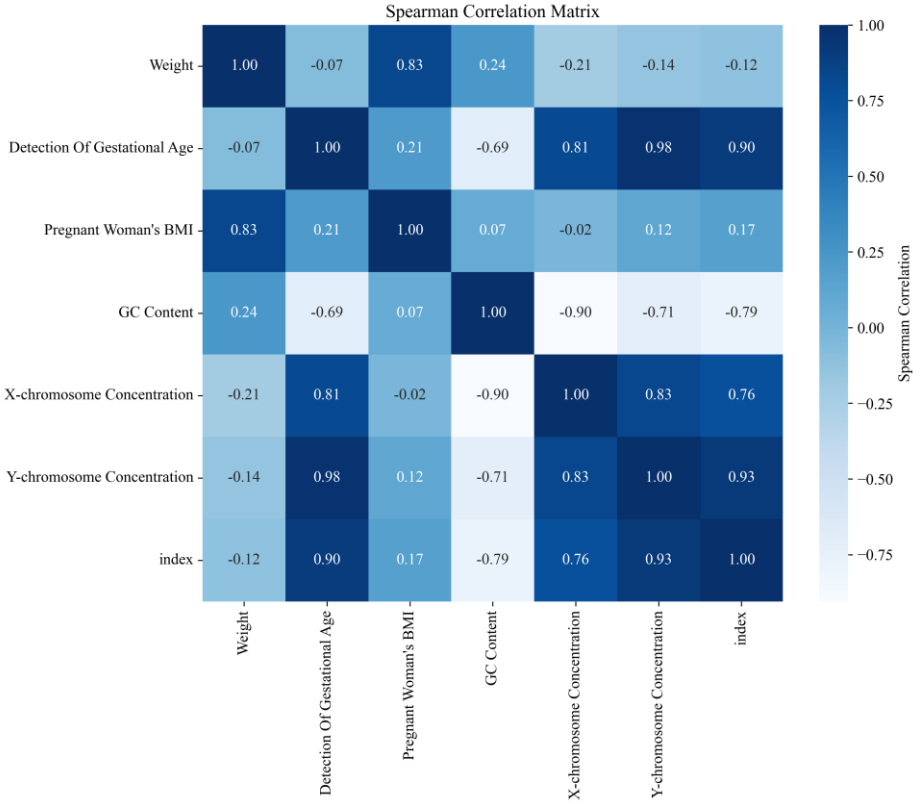

Figure 2 Spearman Correlation Matrix

Figure 2 shows a significant positive correlation between Y-chromosome concentration and gestational age, and no significant association with BMI. The former finding aligns with established medical knowledge. During the early stages of pregnancy, the concentration of fetal cell-free DNA is relatively low, particularly with regard to Y chromosome DNA, which may not reach detectable levels. However, as the pregnancy progresses, the concentration of fetal DNA gradually increases; consequently, an extended gestation period is typically associated with a rise in Y chromosome DNA concentration. Although no significant direct correlation was observed between BMI and Y chromosome concentration, BMI holds considerable importance in subsequent predictive analyses. This suggests that BMI may indirectly influence fetal Y chromosome concentration by affecting factors such as the overall health and metabolic rate of the pregnant woman.

3.3 PSO-XGBoost Prediction Results

Using Python to define a function for finding the maximum R^2 and optimal parameter combination, the solution results are as shown in Table 2:

Table 2 Changes in R ²Fitness

Number of iterations	1	2	3	4	5	6	7	8	9	10
R^2 Fitness	0.9556	0.9564	0.9564	0.9573	0.9573	0.9573	0.9573	0.9582	0.9582	0.9582

As shown in Table 2, after 10 iterations, the optimal solution yields an R ² value of 0.9582, approaching 1 and demonstrating high significance. The optimal parameters are subsample = 0.8201726 and $\min_{c}\text{hild}_{w}\text{eight}$ = 1.

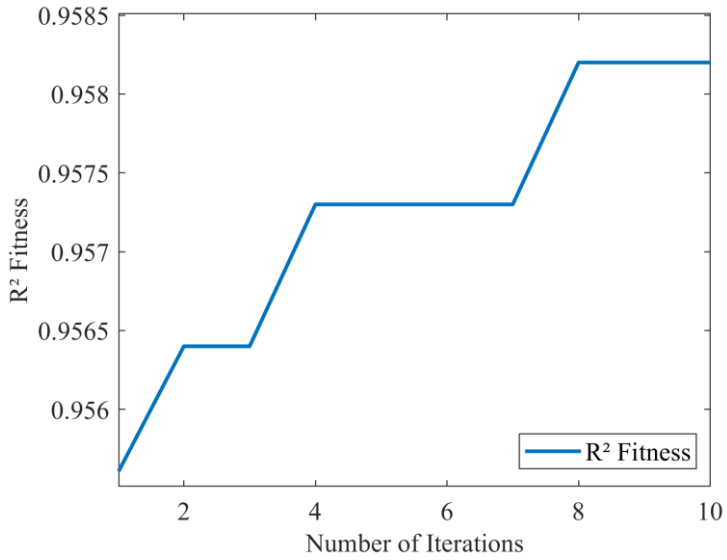

Figure 3 Change in R ²Fitness

Figure 3 illustrates the trend of the R ²fitness function during the iteration process of the PSO algorithm. From the graph, it is clearly observable that during the early stages of iteration, the R ² fitness value exhibits a pronounced upward trend, reflecting the PSO algorithm's ongoing exploration towards more optimal solutions. Upon reaching the third iteration, the model commenced its convergence phase. The R ²fitness value gradually stabilised during subsequent iterations, ultimately settling at approximately 0.9579. As the number of iterations increased further, by the tenth generation, the model successfully identified the global optimum solution. This behaviour, wherein the fitness value rapidly increases during the iterative process before stabilising and converging to ultimately locate the global optimum, fully demonstrates that the PSO-XGBoost model possesses excellent optimisation performance and stability. It is capable of efficiently and reliably searching for optimal solutions in optimisation tasks.

Call the optimal model via Python's scikit-learn library for metric evaluation. The processed data is divided into training and test sets, which are then input into the optimal model to compute the model error metric. Subsequently, the error metrics of the optimised model were compared with XGBoost, with the results presented in Table 3:

Table 3 Model Error Metrics

Indicator	PSO-XGBoost Performance	XGBoost Performance
MAPE	6.56831	28.69994
RMSE	0.00757	0.025026
MAE	0.00546	0.01781
R^2	0.9579	0.5392

As shown in Table 3, the MAPE of the PSO-XGBoost model stands at 6.57%, significantly lower than the 28.70% recorded for the unoptimised XGBoost model. The optimised model achieved an R ² value of 0.958, indicating an excellent fit. The root mean square error (RMSE) obtained from the optimal model optimised using PSO is marginally higher than the mean absolute error (MAE). This indicates that although a small number of samples with significant prediction deviations remain

within the optimised model, both the overall absolute error and root mean square error are maintained at low levels. These results indicate well-controlled predictive error. Both RMSE and MAE are substantially lower than the baseline. Moreover, the optimised model exhibits a goodness-of-fit approaching 1, markedly exceeding the pre-optimisation value, demonstrating a substantial improvement in the PSO-XGBoost model's fitting performance following optimisation. This indicates that the optimal XGBoost model obtained through PSO-optimised hyperparameters demonstrates significantly enhanced performance, exhibiting greater stability and higher predictive accuracy.

3.4 SHAP analysis results

To more intuitively reveal the mechanisms by which various features in the pregnant women dataset influence Y chromosome concentration, SHAP plots were generated to visualise the distribution of SHAP values for each feature within the samples. This quantifies the contribution of each feature to the model's prediction of Y chromosome concentration outcomes. The results are shown in Figure 4.

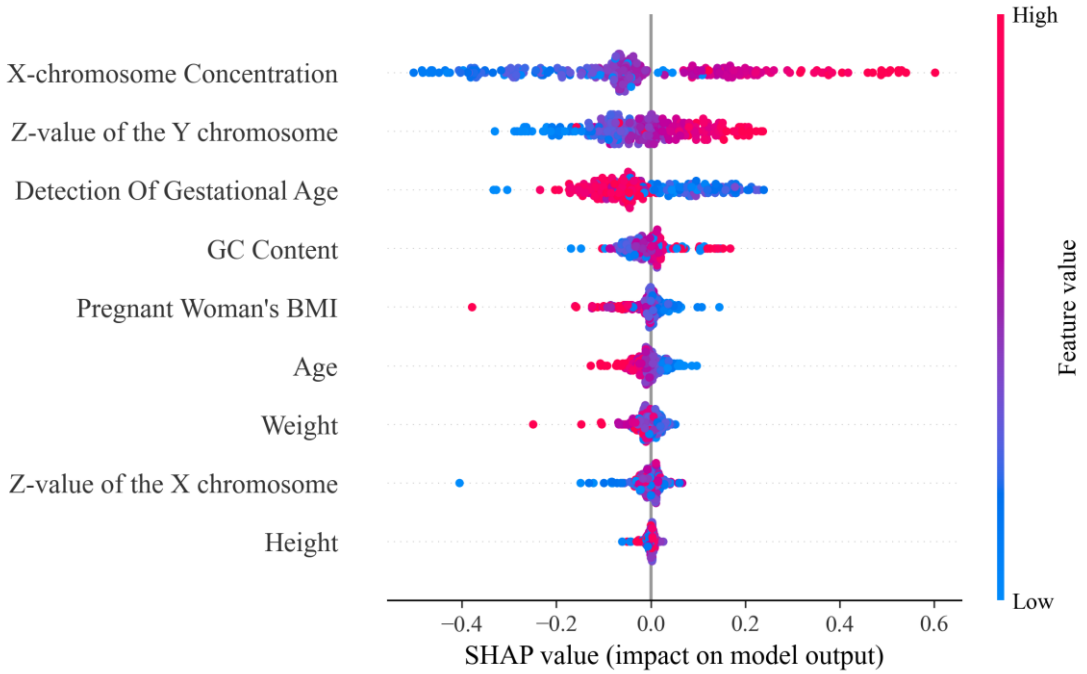

Figure 4 SHAP distribution plot

As shown in Figure 4, the horizontal axis represents SHAP values, indicating the degree of influence each feature exerts on Y chromosome concentration, while the vertical axis denotes the respective features. In SHAP summary plots, colours encode feature values (blue = low, red = high), while the horizontal axis shows SHAP values. A positive value indicates a positive contribution to the forecast result, while a negative value indicates a negative contribution.

From the distribution of SHAP values for the features, the SHAP values for X-chromosome Concentration exhibit a broad range, spanning from negative to positive values. This indicates that this feature significantly influences the prediction of Y-chromosome concentration, with its contribution varying markedly across different samples. The SHAP values for the Z-value of the Y chromosome predominantly cluster in the positive region, indicating that this feature makes a positive contribution to predicting Y chromosome concentration in the majority of samples. Gestational age shows a moderately concentrated SHAP distribution with some spread, indicating directionally mixed contributions across samples. The SHAP value distribution for GC content similarly exhibits distinct positive and negative intervals, indicating that GC content contributes ambivalently to the prediction results. The SHAP value distributions for the features Pregnant Woman's BMI, Age, Weight, Z-value of the X chromosome, and Height are relatively narrow. Notably, Pregnant Woman's BMI exhibits only a sparse distribution in certain local regions. This indicates that these physical characteristics

exert a minor influence on predicting Y chromosome concentration, primarily serving as secondary corrective factors.

By combining colour with SHAP values, it is evident that the majority of feature samples fluctuate within the SHAP value range [-0.02, 0.02]. This indicates that individual features exert limited influence on Y chromosome concentration predictions. However, the cumulative SHAP values of multiple features collectively exert a significant effect, jointly determining the predicted Y chromosome concentration. Concurrently, the alternating red-blue patterns and dispersed distributions observed in the feature distributions also reveal a pronounced non-linearity in the influence of characteristics such as X chromosome concentration, gestational age, and BMI upon predicting Y chromosome concentration. Furthermore, different samples exhibit significant individual variation under the influence of these features, potentially suggesting the presence of other latent feature interactions within the samples that may subsequently impact the prediction of Y chromosome concentration.

To further elucidate the operational mechanism of features at the individual sample level, a SHAP plot was generated for the third sample, as depicted in Figure 5. This provides a visual representation of each feature's specific contribution to predicting the Y chromosome concentration for this particular sample.

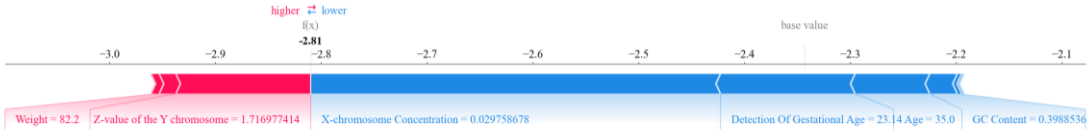

Figure 5 SHAP plot for the third sample

Figure 5 results indicate that the model baseline value is approximately −2.4, with the final predicted value being −2.81, suggesting a strong overall negative feature contribution. Among these, X-chromosome concentration made the most significant negative contribution, being the primary factor pulling down the predicted results. The Z-value of the Y chromosome exhibits a positive contribution, partially offsetting the negative effects. Detection of gestational age and GC content both exerted a moderate negative influence, whereas weight demonstrated only a slight positive effect. This aligns with the overall pattern observed in Figure 4: X chromosome concentration exerts the most significant and directionally inconsistent influence on predicting Y chromosome concentration, while gestational age and GC content exhibit bidirectional effects, and morphological variables exert a weaker influence. Overall, the decline in predictions for this sample was primarily driven by negative characteristics such as X chromosome concentration and gestational age, reflecting the model's high sensitivity to these features and potential non-linear interactions.

4. Conclusion

When confronted with high-dimensional, heterogeneous NIPT data exhibiting complex nonlinearity, traditional linear or parametric methods are prone to systematic bias due to overly stringent assumptions. In summary, a high-precision, interpretable framework enables more accurate prediction of fetal chromosomal concentration, improving screening performance and supporting data-driven precision medicine.

This paper constructs an interpretable PSO-XGBoost algorithm framework. First, variables were filtered through data preprocessing, employing Spearman's correlation analysis to assess the relationship between indicators and Y-chromosome concentration. Subsequently, a PSO-XGBoost model was utilised to predict Y-chromosome concentration, with results compared against a baseline model. Finally, the SHAP method was applied to analyse feature contributions at both the individual and overall levels.

The results indicate that the Spearman correlation matrix demonstrates a significant positive correlation between fetal Y chromosome concentration and gestational age, whilst showing no significant correlation with BMI. The PSO-XGBoost model achieved an R^2 value of 0.958, demonstrating excellent fitting performance. The MAPE, RMSE, and MAE were all significantly

lower than pre-optimisation values, indicating enhanced stability and higher predictive accuracy for this model. SHAP analysis indicates that model predictions are driven by a small number of core features, including X chromosome concentration, Y chromosome Z-value, and gestational age, with significant nonlinear interactions and individual variation observed among these features.

In the future, as clinical information continues to accumulate, this research framework may further incorporate multimodal features such as genomic sequences and epigenetic markers to construct integrated predictive models that fuse multidimensional information. This approach will comprehensively enhance the precision of prenatal screening and diagnostic treatment.

References

[1] J. Yang et al., "Combined fetal fraction to analyze the Z-score accuracy of noninvasive prenatal testing for fetal trisomies 13, 18, and 21," J Assist Reprod Genet, vol. 40, no. 4, pp. 803–810, Apr. 2023, doi: 10.1007/s10815-022-02694-8.

[2] S.-H. Kim et al., "The Association between Low Fetal Fraction of Non-Invasive Prenatal Testing and Adverse Pregnancy Outcomes for Placental Compromise," Diagnostics, vol. 14, no. 10, p. 1020, Jan. 2024, doi: 10.3390/diagnostics14101020.

[3] D. Hanxiao et al., "Noninvasive prenatal prediction of fetal haplotype with Spearman rank correlation analysis model," Molecular Genetics & Genomic Medicine, vol. 10, no. 8, p. e1988, Aug. 2022, doi: 10.1002/mgg3.1988.

[4] N. P. Duarte-Delgado et al., "Cytokine profiles and their correlation with clinical and blood parameters in rheumatoid arthritis and systemic lupus erythematosus," Sci Rep, vol. 14, no. 1, p. 23475, Oct. 2024, doi: 10.1038/s41598-024-72564-z.

[5] L. Wu et al., "Alterations and Correlations of Gut Microbiota and Fecal Metabolome Characteristics in Experimental Periodontitis Rats," Front. Microbiol., vol. 13, Apr. 2022, doi: 10.3389/fmicb.2022.865191.

[6] K. Cao et al., "Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model," Sci Rep, vol. 15, no. 1, p. 12406, Apr. 2025, doi: 10.1038/s41598-025-96520-7.

[7] D. A. Dias Júnior et al., "Automatic method for classifying COVID-19 patients based on chest X-ray images, using deep features and PSO-optimized XGBoost," Expert Systems with Applications, vol. 183, p. 115452, Nov. 2021, doi: 10.1016/j.eswa.2021.115452.

[8] C.-J. Tseng and C. Tang, "An optimized XGBoost technique for accurate brain tumor detection using feature selection and image segmentation," Healthcare Analytics, vol. 4, p. 100217, Dec. 2023, doi: 10.1016/j.health.2023.100217.

[9] B. L. Radhakrishnan, K. Ezra, I. J. Jebadurai, I. Selvakumar, and P. Karthikeyan, "An Autonomous Sleep-Stage Detection Technique in Disruptive Technology Environment," Sensors, vol. 24, no. 4, p. 1197, Jan. 2024, doi: 10.3390/s24041197.

[10] H. Zhuo, T. Li, W. Lu, Q. Zhang, L. Ji, and J. Li, "Prediction model for spontaneous combustion temperature of coal based on PSO-XGBoost algorithm," Sci Rep, vol. 15, no. 1, p. 2752, Jan. 2025, doi: 10.1038/s41598-025-87035-2.

[11] J. Allgaier, L. Mulansky, R. L. Draelos, and R. Pryss, "How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare," Artificial Intelligence in Medicine, vol. 143, p. 102616, Sept. 2023, doi: 10.1016/j.artmed.2023.102616.

[12] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Brain Inf., vol. 11, no. 1, p. 10, Apr. 2024, doi: 10.1186/s40708-024-00222-1.

[13] H. Luo et al., “SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation,” Sci Rep, vol. 14, no. 1, p. 17728, July 2024, doi: 10.1038/s41598-024-67844-7.

[14] H. Xu, X. Peng, Z. Peng, R. Wang, R. Zhou, and L. Fu, “Construction and SHAP interpretability analysis of a risk prediction model for feeding intolerance in preterm newborns based on machine learning,” BMC Med Inform Decis Mak, vol. 24, no. 1, p. 342, Nov. 2024, doi: 10.1186/s12911-024-02751-5.

[15] Z. Fan et al., “XGBoost-SHAP-based interpretable diagnostic framework for knee osteoarthritis: a population-based retrospective cohort study,” Arthritis Res Ther, vol. 26, no. 1, p. 213, Dec. 2024, doi: 10.1186/s13075-024-03450-2.

[16] M. Lugner, A. Rawshani, E. Helleryd, and B. Eliasson, “Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data,” Sci Rep, vol. 14, no. 1, p. 2102, Jan. 2024, doi: 10.1038/s41598-024-52023-5.