Received: 21 April 2025 | Revised: 20 June 2025 | Accepted: 23 June 2025 | Published online: 18 August 2025

RESEARCH ARTICLE

Smart Wearable Technology
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewSWT52025979

BON VIEW PUBLISHING

Study on Diabetes Detection Based on DBO-RF Pulse Wave Analysis

Quanyu Wu[1,*], Shilong Gong[1], Mingying Hu[1], Lingjiao Pan[1], Xiaojie Liu[1] and Weige Tao[1]

[1]*School of Electrical and Information Engineering, Jiangsu University of Technology, China*

Abstract: As society progresses, diabetes has emerged as one of the major global health concerns. Traditional diagnostic methods for diabetes rely heavily on blood tests, which are often invasive, costly, and require specialized equipment. Consequently, this study adopts a non-invasive pulse wave analysis technique, utilizing key feature extraction from pulse wave signals combined with machine learning techniques to predict the presence of diabetes risk. Data for this study were sourced from the Guilin People's Hospital, with a collection of 657 pulse wave samples from 219 participants. Through signal preprocessing and feature extraction, a classification model centered on Dung Beetle Optimization (DBO) and Random Forest (RF) was developed to detect diabetes. The preprocessing included the use of Complete Ensemble Empirical Mode Decomposition with Adaptive Noise combined with Permutation Entropy and wavelet thresholding methods to enhance signal quality. Subsequently, 29 key features, including time domain, entropy domain, and statistical features, were extracted from the processed signals to create a comprehensive feature set. Here optimize the model's classification performance and address issues of missing values in feature vectors and sample imbalance, further evaluate various machine learning algorithms and select the most effective one. Ultimately, the DBO algorithm was applied to optimize the number of decision trees and the selection of feature numbers in the RF classifier. Experimental results demonstrated that our method achieved a 100% accuracy rate on the training set and 92.9% on the test set, significantly outperforming traditional machine learning approaches. These findings confirm the potential of non-invasive pulse wave analysis in the early detection of diabetes, offering possibilities for future clinical applications.

Keywords: signal denoising, diabetes detection, pulse wave analysis, DBO algorithm, RF

1. Introduction

As society progresses, diabetes has emerged as one of the major global health concerns [1]. In response, the Chinese government has implemented a series of measures to address the challenges posed by the high prevalence and increasing trend of diabetes, which has garnered significant national attention and has become a critical policy issue. Recent years have seen innovative technological advances in diabetes detection research based on pulse wave analysis, encompassing areas such as signal processing, feature engineering, model selection, and optimization to enhance data accuracy and usability. In signal processing, Liang [2] employed techniques including wavelet denoising, rational cycle segmentation, and outlier elimination. Using the ShuffleNet neural network, he successfully analyzed and processed fingertip pulse signals to detect hyperglycemia with an accuracy of up to 86%. Liu [3] adopted improved signal preprocessing techniques for analyzing photoplethysmographic (PPG) and electrocardiogram (ECG) signals and integrating Extreme Learning Machine, convolutional neural network, and fractional order system methods achieved effective blood sugar detection with an accuracy exceeding 85%. Ramu Reddy et al. [4] explored the role of heart rate variability and PPG signal waveform characteristics in the classification of Type 2 diabetes mellitus (DM). Temporal (F1), frequency (F2), nonlinear (F3), and waveform (F4) features were utilized to develop a support vector machine (SVM) model, achieving a performance of 82%. The study further validated the effectiveness of the DM classification system by combining different feature sets and feature percentages.

In the realm of feature engineering, Shi [5] focused on processing electromyographic noise and motion artifacts, employing techniques for extracting temporal, frequency, and nonlinear features. A total of 52 features were initially derived, from which 29 key features were retained after optimization for sleep classification. The integration of Genetic Algorithms (GA), Dung Beetle Optimization (DBO), and Radial Basis Function (RBF) neural networks demonstrated a comprehensive classification performance of 74%. Similarly, Chen et al. [6] enhanced the accuracy of non-invasive blood glucose monitoring by combining temporal and frequency domain analyses. Temporal analysis was used to capture waveform characteristics of PPG signals, while frequency domain analysis employed fast Fourier transform to extract spectral information. A BP-based glucose detection model was established and optimized using GA. This integrated approach exhibited excellent predictive accuracy in oral glucose tolerance tests, with both MAE and root mean square error (RMSE) remaining within reasonable ranges. Jiang [7] proposed a feature extraction method for single-cycle pulse wave signals, capturing physiological information such as main waves, tidal waves, and reperfusion waves, based on Gabor time-frequency atoms and sparse representation. The extracted Gabor feature vectors achieved an accuracy of 93.54%

*Corresponding author: Quanyu Wu, School of Electrical and Information Engineering, Jiangsu University of Technology, China. Email: wuquanyu@jsut.edu.cn

in SVM-based pulse wave diabetes classification. Xiao et al. [8] developed a blood glucose control detection method based on collision entropy, utilizing features extracted from pulse and ECG signals to accurately differentiate between healthy individuals, well-controlled, and poorly controlled diabetic patients. Hettiarachchi and Chitraranjan [9] delved into the morphological features related to PPG waveforms and their derivatives, successfully identifying characteristics closely associated with Type 2 DM and validating the feasibility of predicting this condition using short PPG signals. Among various classification models based on the selected feature sets, linear discriminant analysis achieved the highest area under the ROC curve, reaching 79%. Saha et al. [10] implemented the XGBoost algorithm to create 35 feature subsets, evaluating their individual and collective impacts on predictive modeling. The evaluation highlighted that specific feature combinations significantly influenced model accuracy, particularly those including the mean and standard deviation of instantaneous frequency, with the highest accuracy reaching 96%, underscoring its critical role in predicting Type 2 DM.

In the realm of model selection and optimization, Li [11] utilized processed PPG signals to extract glucose-related features and employed a RBF neural network optimized with Particle Swarm Optimization to enhance the accuracy and practicality of non-invasive glucose monitoring. Zhang et al. [12] analyzed diabetes patients' pulse signals through a 9-level wavelet decomposition, employing a Stacking ensemble learning algorithm to underscore the significance of arterial damage in early diabetes detection, achieving high accuracy where the weak learner, Random Forest (RF), reached an accuracy of 91.1%. Bavkar and Shinde [13] applied various machine learning methods, extracting various frequency and time-domain features from single pulse wave PPG signals, and trained neural networks for glucose detection. Their results indicated that the decision tree algorithm performed best in diabetes prediction, achieving an accuracy of 89.97%. Zhang et al. [14] proposed a blood glucose monitoring system based on pulse wave analysis, achieving an accuracy rate of 81.49%. Shi et al. [15] input selected features into seven widely used machine learning algorithms, evaluated their performance using stratified 10-fold cross-validation, and applied multiple regularization techniques to prevent overfitting. Experimental results showed that the SVM based on the RBF model performed best, with an average accuracy of 84.7%, a G-mean of 84.54%, and an *F*-score of 84.03%.

Overall, physiological signals, due to their cost-effectiveness and ease of use, have been widely employed for diabetes estimation [16–18], diabetes diagnosis [19–21], and the detection of diabetic complications [22–24]. On the other hand, they also have been used in the field of hypertension research.

2. Methods

2.1. Database used

This study's data originated from the People's Hospital of Guilin, China, and involved pulse wave sensor (PPG) data from 219 participants, totaling 657 entries. The data collection captured physiological information and cardiovascular history of the participants using portable hardware devices, synchronously collecting PPG waveforms from the left fingertip and blood pressure data from the right forearm. The entire data recording process was completed within three minutes by professional medical staff, with a sampling rate of 1 kHz and a precision of 12-bit AD conversion. Each participant's data were divided into three segments, each containing 2100 sample points, corresponding to a duration of 2.1 s. All data were saved in text files (*.txt format). Moreover, data quality was assessed using the Skewness SQI value to ensure the accuracy of the records obtained.

2.2. Data preprocessing

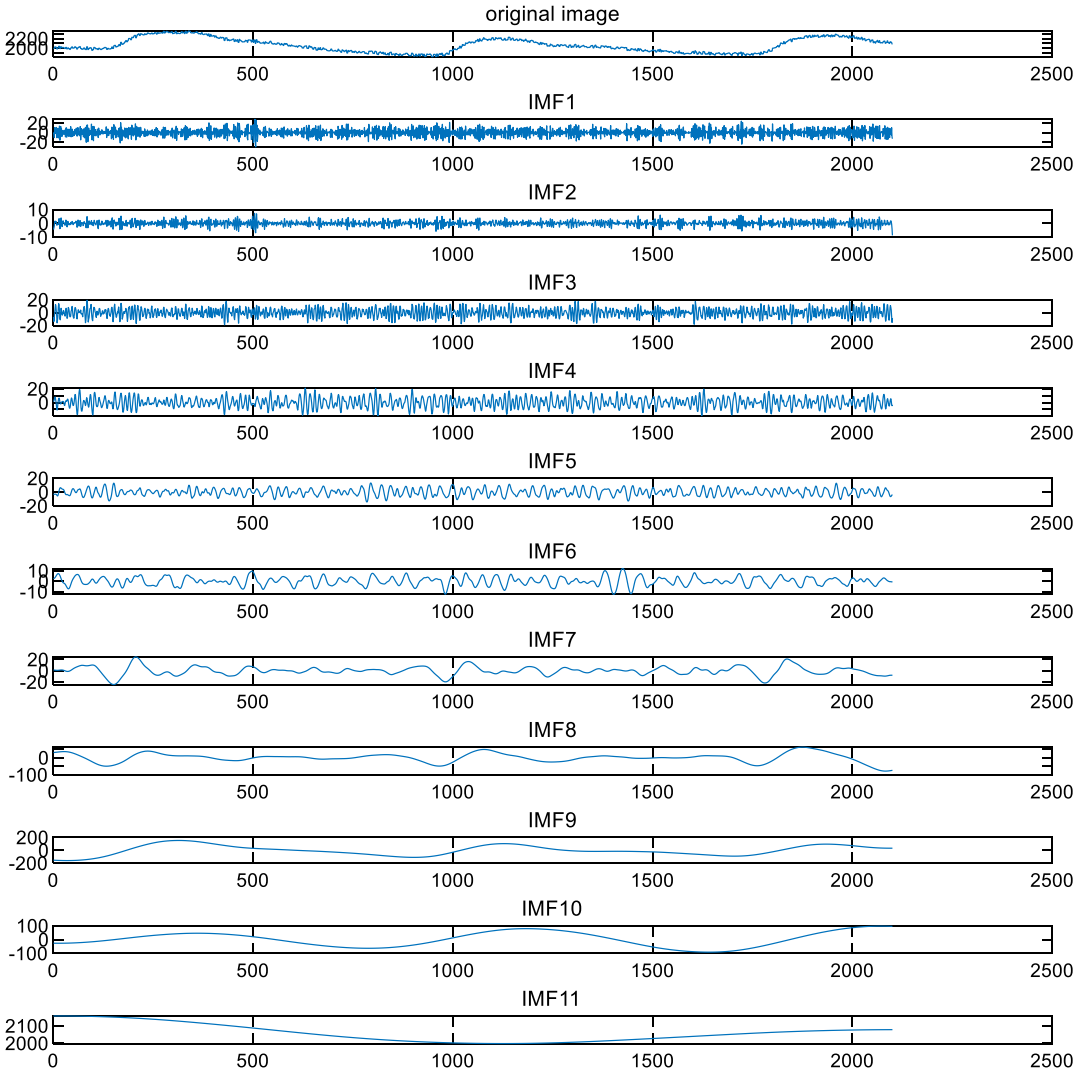
Human fingertip pulse signals are non-stationary, low-frequency, weak signals with a frequency range of 0–40 Hz, where the main energy is concentrated in the 0–1.0 Hz range. The low-frequency components reflect the characteristics of the signal, while the high-frequency components capture subtle differences. However, the signal may contain noise due to the movement of the subject's body, particularly high-frequency noise, which can impact the performance of the model in recognizing signal patterns.

Zhang et al. [25] proposed a pulse signal denoising method that integrates Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Permutation Entropy (PE). The process begins with the decomposition of the signal using CEEMDAN to obtain intrinsic mode functions (IMFs). PE is then utilized to determine the boundary between noise and signal components, followed by adaptive threshold processing of the noisy IMFs, and finally reconstructing the denoised signal. Results indicate that this method improves the signal-to-noise ratio (SNR) and reduces the (RMSE) across various SNRs, outperforming traditional methods. Wu et al. [26] developed a pulse wave denoising algorithm based on the wavelet thresholding method. By selecting an appropriate wavelet base and decomposition levels, the algorithm effectively suppressed motion artifacts. The best results were achieved using the db9 wavelet base with six levels of wavelet decomposition, significantly enhancing the SNR. Chen et al. [27] introduced a photoplethysmogram signal denoising method that combines Ensemble Empirical Mode Decomposition (EEMD) with wavelet thresholding. The signal is decomposed through EEMD, the coherence of modal components is calculated, and noise components are processed using wavelet thresholding, effectively removing high-frequency noise and baseline drift. This approach improves SNR and reduces RMSE, providing new insights for accurate measurements. Li et al. [28], Liu et al. [29], and Lou et al. [30] also conducted significant research on denoising methods for vibration signals, heart sounds, and rain noise using techniques like CEEMDAN and wavelet thresholding. These studies demonstrate the effectiveness and applicability of combining CEEMDAN with wavelet-related technologies in various signal processing contexts, offering efficient solutions for noise suppression and signal preservation.

Based on the aforementioned research, the CEEMDAN and PE and wavelet threshold denoising techniques were applied to the dataset. CEEMDAN addresses the mode mixing issue present in traditional EMD methods by decomposing the signal multiple times and incorporating white noise, thereby yielding a series of IMF components. Through repeated experiments, it was found that the best decomposition results occurred when the ratio of the standard deviation of the added white noise to the standard deviation of the vibratory radial signal (amplitude) was 0.2, with an average of 500 iterations and a maximum of 2100 iterations.

This approach effectively extracts the true characteristics of the signal, enhancing the precision and reliability of subsequent analysis and processing. The 11 IMF components (IMF1 to IMF11) are derived from the original signal through CEEMDAN decomposition, with the results displayed in Figure 1. After obtaining all IMF components through CEEMDAN, we here introduce PE as a nonlinear dynamical metric to assess the complexity and regularity of each component. PE quantifies the dynamic information of a time series, allowing us to determine whether each component primarily

Figure 1
CEEMDAN treatment effect



contains useful signal or noise based on its numerical value. The specific PE values are illustrated in Figure 2.

For IMF components identified as noise-dominated, we next apply wavelet threshold denoising techniques for processing. Wavelet thresholding is capable of selectively eliminating noise based on the frequency characteristics and energy distribution of the signal, while maximally preserving the effective information of the signal. Here, the soft thresholding method is utilized, processing IMF components with thresholds exceeding 0.3. After processing all IMF components using wavelet threshold denoising, they are recombined to achieve denoising reconstruction of the original signal. The soft wavelet reconstruction of image is displayed in Figure 3.

Finally, the preprocessed signal undergoes smoothing, as illustrated in Figure 4. The original signal is represented in blue in the upper part of the figure, displaying significant peaks and fluctuations, likely caused by measurement noise or other interfering factors. The red represents the smoothed signal, where it is evident that the fluctuations are substantially reduced, and the peaks and valleys are smoother and easier to discern. From a signal processing perspective, smoothing helps to reduce random

Figure 2
The PE value of each IMF component

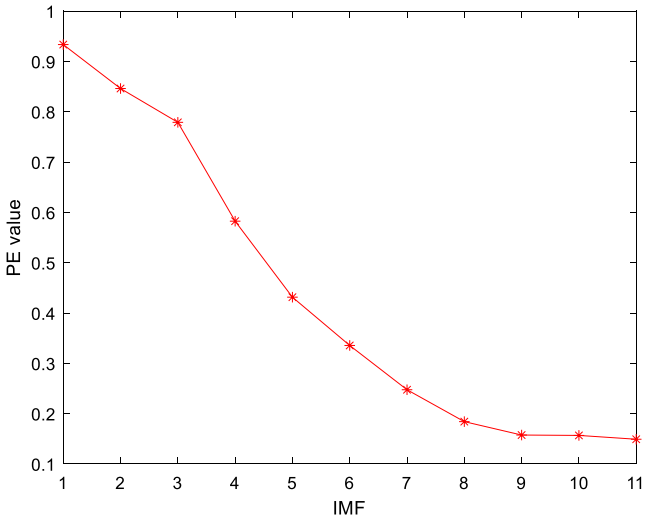


Figure 3
Soft wavelet reconstruction effect image

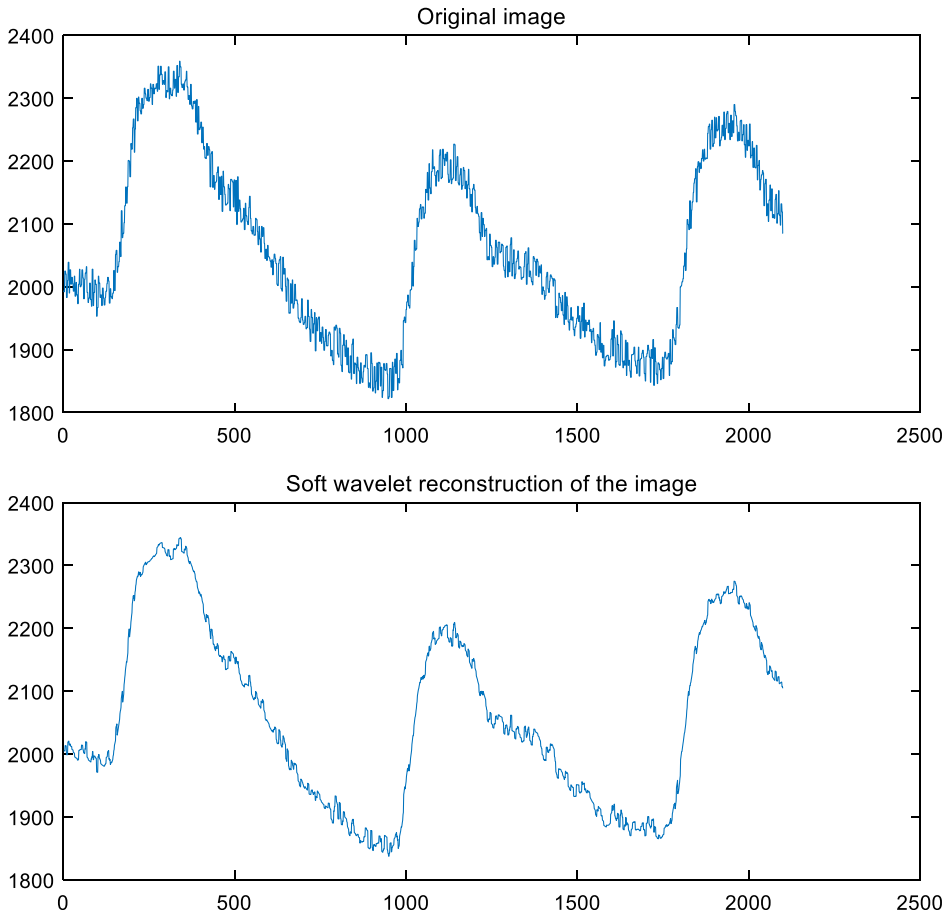
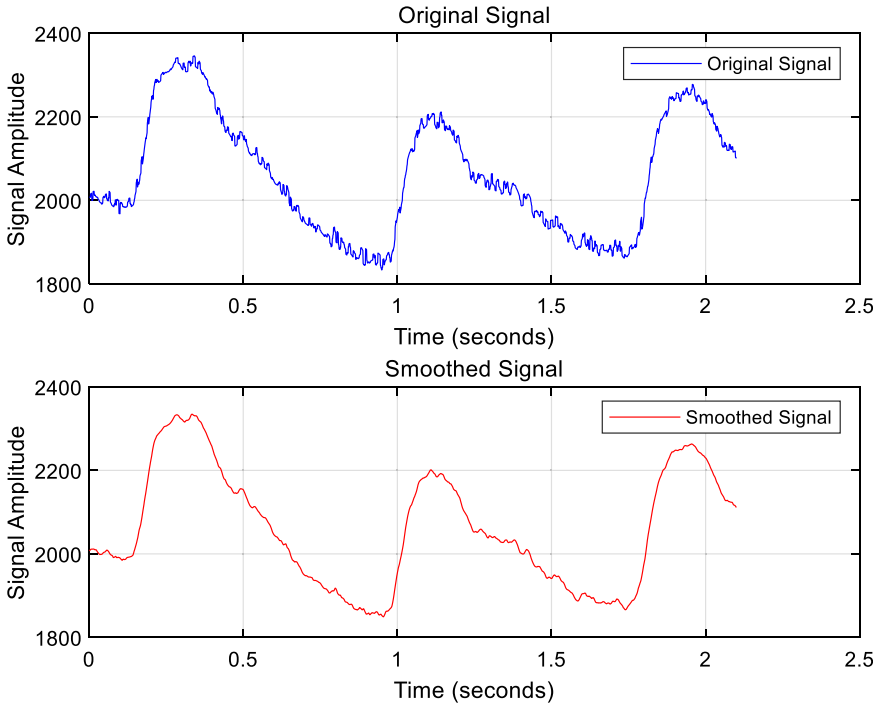


Figure 4
Smoothing effect images



variability in the data, making the main trends and periodic characteristics of the signal more apparent. This treatment is commonly used to enhance data readability, reduce noise impact, and provide a clearer data foundation for subsequent analysis.

2.3. Feature extraction

In the medical field, pulse wave analysis is an expanding area of research, particularly in its application of non-invasive methods for health monitoring and disease prevention. As a crucial biosignal reflecting the cardiovascular status of individuals, pulse wave analysis offers substantial potential for the early detection of cardiovascular diseases. This study successfully extracted 29 features from the pulse wave signal, covering time domain, entropy domain, and statistical characteristics, thereby constructing a comprehensive feature set. These features were used to perform a binary classification of health states, differentiating between healthy and unhealthy individuals.

2.4. Missing value handling

During the extraction of time-domain features in biomedical signal processing, missing value processing is particularly important. The occurrence of missing values can be attributed to various factors, including limitations of the signal capture equipment, poor sensor contacts, or physiological changes in the subjects. Addressing missing values is crucial to ensure the quality of data and the accuracy of research outcomes. Incorrect handling of missing data could lead to biased analysis results, which in turn could impact the final diagnosis of diseases.

Therefore, we chose to delete data records containing missing values as a resolution strategy. Additionally, we excluded signals that did not meet the analytical requirements. To evaluate the effectiveness of this data handling strategy, we designed a series of comparative experiments to assess the performance of the processed data against the original data using the same classification model. We employed a RF model, setting the training and testing set ratio at 8:2. The specific outcomes of this comparison are presented in Tables 1 and 2.

The experimental results indicated that the classification accuracy improved by 2% after processing the data. This not only validates the rationale behind our strategy of deleting records with missing values but also confirms the effectiveness of this approach. By employing this method, we ensure a higher quality of datasets used for training and testing, thereby enhancing the reliability and accuracy of the entire study.

Table 1
Diagnosis rate of unprocessed missing values

	Diagnose	Misdiagnosis	Diagnosis rate
Diabetes	104	0	100%
No diabetes	6	22	21.4286%
Accuracy		83.3333%	

Table 2
Diagnosis rate after processing missing values

	Diagnose	Misdiagnosis	Diagnosis rate
Diabetes	98	6	94.23%
No diabetes	5	12	29.41%
Accuracy		85.124%	

3. Model Selection

3.1. Model comparison

In this study, we employed a variety of traditional machine learning algorithms to address the binary classification problem of pulse wave signals. These algorithms include Naive Bayes, SVM, AdaBoost, RF, K-Nearest Neighbors, and Decision Trees. By analyzing the 29 extracted features, our objective was to determine which algorithm is most suitable for the classification of this type of biomedical data. After a series of detailed experimental designs and data analyses, the results are presented in Table 3.

Table 3
Classification accuracy of traditional machine learning

Algorithm	Accuracy
NB	66.6667%
SVM	81.6092%
AdaBoost	82.5758%
RF	85.124%
KNN	72.7273%
DT	75.7576%

In this study, we compared the accuracy of several algorithms to determine the most suitable method for pulse signal classification. The RF algorithm performed the best among all tested algorithms, achieving an accuracy rate of 92.9%. This high level of accuracy is attributed to the algorithm's ability to handle complex feature interactions and reduce the risk of overfitting through ensemble learning. Furthermore, the majority voting mechanism of the RF algorithm further enhances the model's stability and accuracy, making it an ideal choice for pulse signal classification.

3.2. Sample equilibrium

In our study, while RF demonstrated excellent overall classification accuracy, a deeper analysis of the experimental results revealed some issues, particularly in the performance on the test set: the classification accuracy for patients without disease reached as high as 94%, but it was only 29% for patients with disease, as shown in Table 2. This huge difference caught our attention, which was carefully analyzed, and believe it was primarily due to sample imbalance. Sample imbalance is a common problem in machine learning, especially in the medical field, where data for healthy individuals are often more abundant than data for individuals with diseases. In our dataset, the number of samples for patients without disease far exceeded those for patients with disease. This caused the model, during its learning process, to overemphasize the features of the majority class (patients without disease) and fail to adequately learn the critical features necessary to identify the minority class (patients with disease). To address the issue of sample imbalance, we utilized the synthetic minority oversampling technique (SMOTE), which generates synthetic samples by interpolating between existing minority class samples. SMOTE not only increases the number of the minority class but also enhances the diversity of the data. This diversity is crucial for the model to learn more generalized features, which can improve its ability to correctly classify minority class instances. The effectiveness of this approach is demonstrated in Table 4, which shows the improved performance metrics after applying SMOTE to our dataset.

Table 4
Sample imbalance

	Label	Number	Percentage
Before equilibrium	Healthy	495	81.95%
	Unhealthy	109	18.05%
After equilibration	Healthy	495	50.00%
	Unhealthy	495	50.00%

Table 5
Effect of the sample before and after equilibration

Algorithm	Before equilibrium	After equilibration
NB	66.6667%	68.1818%
SVM	81.6092%	84.8126%
AdaBoost	82.5758%	84.3434%
RF	85.124%	86.8687%
KNN	72.7273%	73.2323%
DT	75.7576%	78.7879%

After adjusting for sample balance, we retrained the RF model as well as other models and compared the classification performance before and after the adjustments. The models processed for sample balance showed significant improvement in identifying patients with the disease, with a substantial increase in accuracy, while still maintaining high accuracy for identifying patients without the disease, as shown in Table 5. These results underscore the importance of sample balance in enhancing the overall performance of models.

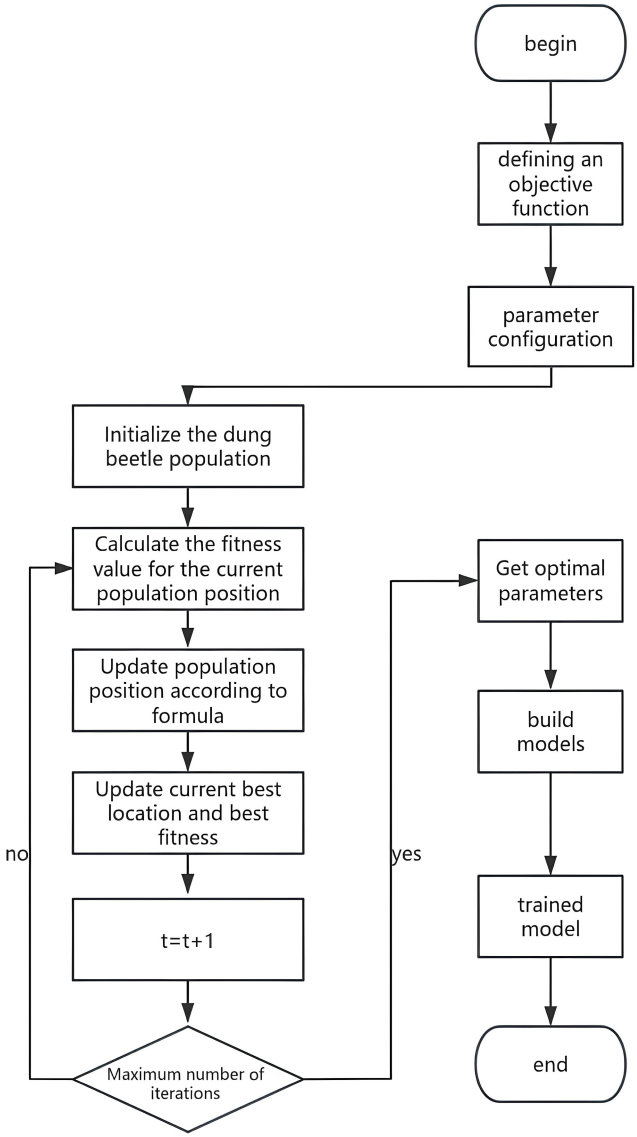
4. Classification Model of Diabetes Based on DBO-RF

4.1. Classification and evaluation

The RF algorithm is widely favored for its robust classification capabilities and excellent resistance to overfitting. However, its performance heavily depends on the setting of several key parameters, including the number of trees, tree depth, and other variables. Inappropriate parameter settings can lead to poor model performance on specific datasets. Here apply the DBO algorithm, inspired by dung beetle behavior, to optimize RF parameters. This optimization approach is used to fine-tune the number of trees and the maximum number of features considered during the splitting process of each decision tree in the RF, aiming to achieve higher prediction accuracy. Flowchart of the DBO-RF Algorithm as shown in Figure 5.

The DBO algorithm classifies the population into four roles based on the social divisions of dung beetles: ball-rolling beetles, brooding ball beetles, juvenile beetles, and thief beetles. In a population of 30, these roles are allocated to 6, 6, 7, and 11 individuals, respectively. Each type of beetle is represented by different symbols—R, B, L, and T—which denote their positions in the solution space. When the population count N equals 30, the number of beetles in these four roles is set to 6, 6, 7, and 11, respectively. If the dimension of the optimization problem is D, with a corresponding objective function f, then each beetle's position, representing a solution, is expressed as $\{x_i1, x_i2\cdots, x_iD\}$, and individual fitness values, indicating optimization quality, are denoted by $f(x_i)$. Fitness values are represented similarly, with smaller values indicating better optimization outcomes. Thus, a smaller fitness value signifies a better position for survival. The mathematical representation of the best and worst positions based on these strategies is as follows:

Figure 5
Flowchart of the DBO-RF algorithm



$$\begin{cases} X^b = \{X_i \in X, i = 1, 2\cdots, N | \forall X_j,\ f(X_i) \le f(X_j)\} \\ X^w = \{X_i \in X, i = 1, 2\cdots, N | \forall X_j,\ f(X_j) \le f(X_i)\} \end{cases} \quad (1)$$

The standard DBO algorithm steps are as follows.

Step 1: Set the maximum number of iterations as T, population size as N, randomly initialize the population, and calculate the fitness value of each individual.

Step 2: Update the positions of the ball-rolling dung beetles. If $\lambda < \gamma$, update the position under the unobstructed state using Equation (2); otherwise, update it under the obstructed state using Equation (4), where λ is a random number within [0,1] and $\gamma = 0.9$.

$$R_{new,e}^{t+1} = R_1^t + \alpha \times k \times R_e^{t-1} + u \times \Delta x \quad (2)$$

$$\Delta x = |R_e^t - X^w| \quad (3)$$

$$R_{new,e}^{t+1} = R_e^t + \tan(\theta)|R_e^t - R_e^{t-1}| \quad (4)$$

where t represents the current iteration count; R_e^t denotes the position of the e-th rolling beetle after the t-th iteration; $k = 0.2$ represents the position deflection coefficient; α represents the natural

environmental interference coefficient; u is a random factor; θ represents the deflection angle in radians, and θ is a random number with $\theta \in [0, \pi]$. If θ is 0, π/2, or π, the position is not updated.

Step 3: Update the position of the brood ball according to Equation (6) and use the upper and lower bounds in Equation (5) to constrain the new position.

$$Lb^* = \max\left(X^{b*} \times (1 - Q), Lb\right)$$

$$Ub^* = \min\left(X^{b*} \times (1 + Q), Ub\right) \quad (5)$$

$$B_{new,m}^{t+1} = X^{b*} + a_1 \times \left(B_m^t - Lb^*\right) + a_2 \times \left(B_m^t - Ub^*\right) \quad (6)$$

In Equations (5) and (6), Ub^* and Lb^* represent the upper and lower bounds of the spawning area, respectively; X^{b*} denotes the current optimal position of each dung beetle; Q = 1–t/T, where T represents the maximum number of iterations; Ub and Lb are the global upper and lower bounds of the problem; B_m^t represents the coordinates of the m-th brood ball after the t-th iteration; a_1 and a_2 are both 1×D-dimensional random vectors; D is the dimension of the optimization problem solution.

Step 4: Update the position of the little beetle from Equation (7).

$$L_{new,h}^{t+1} = L_h^t + C_1 \times \left(L_h^t - Lb'\right) + C_2 \times \left(L_h^t - Ub'\right) \quad (7)$$

In Equation (7), L_h^t represents the coordinate of the h-th dung beetle after the t-th iteration; C_1 follows a normal distribution and $C_1 \in [0,1]$, while C_2 is a random vector with all components falling within the range of [0,1];

Step 5: Update the location of the thief dung beetle by Equation (8).

$$T_{new,z}^{t+1} = X^b + S \times g \times \left(\left|T_z^t - X^{b*}\right| + \left|T_z^t - X^b\right|\right) \quad (8)$$

In Equation (8), X^b denotes the globally optimal position; X^{b*} is the best solution obtained through the integration of various roles in the current iteration. T_z^t signifies the coordinate of the z-th thieving dung beetle after the t-th iteration; S = 0.5; g is a random vector with all components lying between [0,1].

Step 6: Update the global optimal and worst positions.

Step 7: Determine if the algorithm has reached the number of iterations. If so, terminate the operation and return to the optimal position, which is the optimal solution to the problem. Otherwise, proceed to step 2. We initiated training using the established model, setting the number of decision trees between 100 and 200, and the optimal number of features for each split between 1 and 5. Additionally, the model was trained with a population of 30, and the total number of iterations was set at 50. The experimental results show that the model achieved an accuracy rate of 92.9% on the test set. The results are shown in Figure 6. The fitness change curve is depicted in Figure 7.

4.2. Model evaluation index

The confusion matrix is an important tool for evaluating the performance of classification models, especially in handling binary classification issues. It provides a detailed view of performance by comparing the model's predicted categories with the actual categories, as shown in Figure 8.

Finally, to comprehensively evaluate the classifier's performance, precision, recall, and F1-score metrics were derived from the confusion matrix. The results were subsequently compared with the performance of the RF algorithm and other algorithms, as shown in Table 6.

Figure 6
The comparison of prediction results between test sets

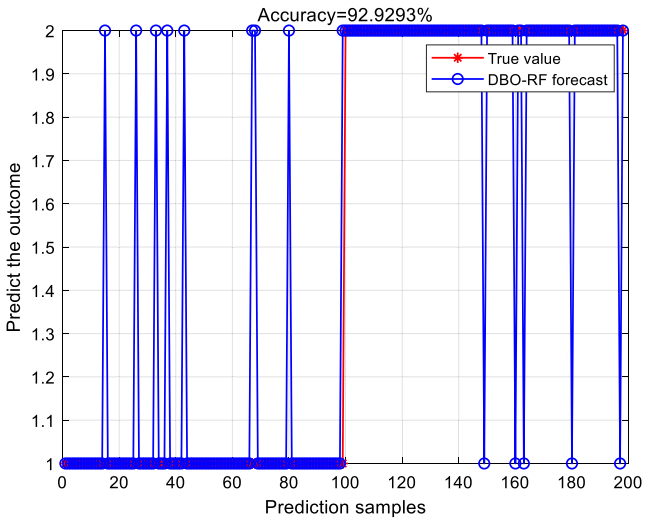
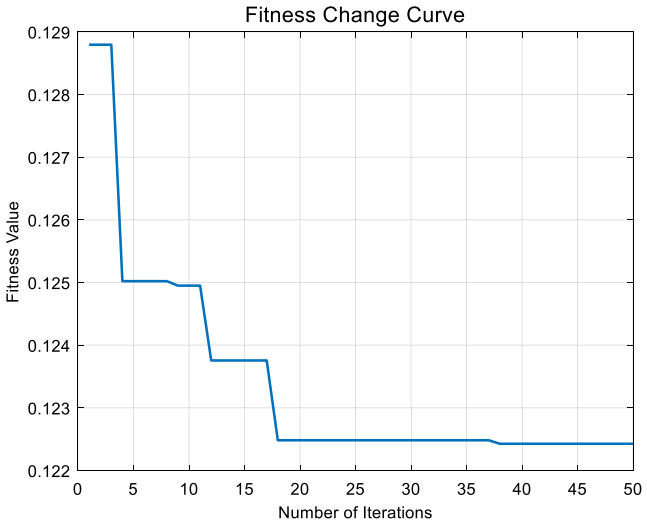


Figure 7
Fitness change curve of DBO algorithm



According to the data presented in Table 6, it is observed that the DBO-RF model outperforms the RF model across all evaluation metrics, demonstrating a significant enhancement in the overall performance of the DBO-RF model. The substantial improvements in accuracy and recall specifically highlight the DBO-RF model's advantages in effectively identifying healthy class samples. These results effectively validate the efficacy of the methodology proposed in this paper.

5. Some Common Mistakes

In our research, we here obtained pulse wave data from the Guilin People's Hospital and subjected it to a comprehensive series of preprocessing, feature extraction, and classification experiments. Our study distinctly emphasized the importance of removing missing values in experiments, where adopting

Figure 8
Training set confusion matrix

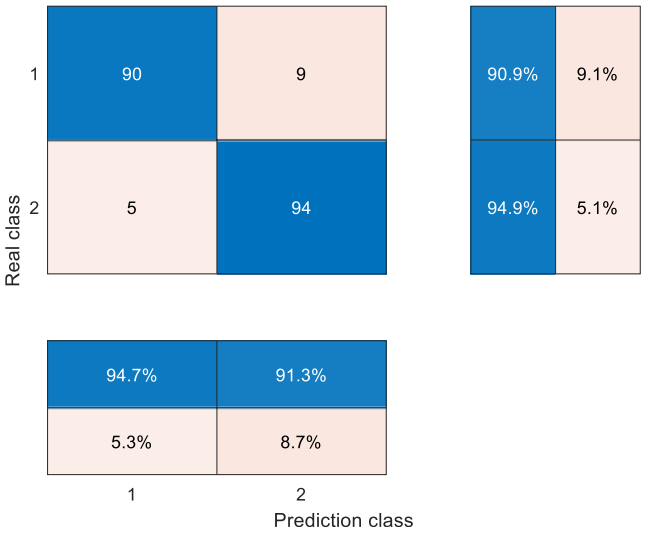


Table 6
Evaluation effect of the model

	Accuracy	Precision	Recall	*F*1 score
RF	89.4%	90.0%	89.1%	0.895
DBO-RF	92.7%	90.91%	94.7%	0.927
SSA-RF	90.9%	82.83%	98.78%	0.901
NGO-RF	89.4%	87.88%	90.63%	0.892

appropriate strategies ensured the high quality of training and testing datasets, thereby enhancing the reliability and accuracy of the research. Furthermore, sample imbalance is a major issue when handling medical data. Through techniques such as oversampling the minority class and undersampling the majority class, we successfully balanced the sample distribution, significantly improving the model's ability to recognize diabetes risk (the minority class), while maintaining high accuracy for identifying healthy individuals (the majority class). This achievement highlights the importance of balancing samples to enhance the overall performance of classification algorithms. Additionally, our experiments tested a variety of machine learning algorithms, including RF, SVM, and logistic regression. Notably, after data balancing, the RF algorithm performed the best, underscoring the importance of choosing the right algorithm based on specific data characteristics. Finally, we utilized the DBO algorithm to adjust model parameters, improving accuracy from 82% to 92.9%. These findings not only demonstrate the potential of our methods in handling real medical data but also highlight the necessity of finely tuning model parameters in complex medical data analysis.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created.

Author Contribution Statement

Quanyu Wu: Conceptualization, Methodology, Validation, Formal analysis, Resources, Writing – review & editing, Visualization. **Shilong Gong:** Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Mingying Hu:** Software, Investigation, Writing – original draft, Writing – review & editing. **Lingjiao Pan:** Writing – review & editing. **Xiaojie Liu:** Supervision. **Weige Tao:** Project administration.

References

[1] Guo, L. (2020). 2021 Nián tángniàobìng lǐngyù niándù zhòngdà jìnzhǎn huígù [2021 annual review of major progress in diabetes]. *Chinese Journal of Diabetes Mellitus*, *14*(1), 1–8. http://doi.org/10.3760/cma.j.cn115791-20220114-00034

[2] Liang, W. (2024). *Jīyú zhǐ jiān màibó xìnhào de gāo xiětáng rénqún de wúchuàng jiǎncè fāngfǎ yánjiū [Research on non-invasive detection method of hyperglycemia based on fingertip pulse signal]*. Master's Thesis, Beijing University of Posts and Telecommunications. http://doi.org/10.26969/d.cnki.gbydu.2023.000704

[3] Liu, Y. (2018). *Jīyú PPG hé ECG xìnhào rónghé de wúchuàng xiětáng jiǎncè fāngfǎ yánjiū [Research on non-invasive blood glucose detection method based on the fusion of PPG and ECG signals]*. Master's Thesis, Guangdong University of Technology. http://doi.org/10.7666/d.D01523882

[4] Ramu Reddy, V., Dutta Choudhury, A., Jayaraman,S., Kumar Thokala, N., Deshpande, P., & Kaliaperumal, V. (2017). PerDMCS: Weighted fusion of PPG signal features for robust and efficient diabetes mellitus classification. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies, 5*, 553–560. https://doi.org/10.5220/0006297205530560

[5] Shi, S. Y., He, K., & Luo, T. (2023). Research on signal noise processing method based on PPG sleep staging. *Software Guide*, *23*(4), 67–73. https://doi.org/10.11907/rjdk.231300

[6] Chen, J., Ren, J., Yang, J., Guo, Y., & Qiao, W. (2024). Jīyú shí pín yù zònghé fēnxī de wúchuàng xiětáng jiǎncè jìshù yánjiū [Study on non-invasive blood glucose detection technology based on time frequency domain analysis]. *Spectroscopy and Spectral Analysis*, *44*(2), 318–324.

[7] Jiang, Z. (2020). *Màibó bō cǎijí xìtǒng shèjì yǔ shí pín tèzhēng fēnxī fāngfǎ yánjiū [Research on pulse wave signal acquisition system design and time-frequency feature analysis method]*. PhD Thesis, Harbin Institute of Technology. http://doi.org/10.27061/d.cnki.ghgdu.2020.004985

[8] Xiao, M., Lu, C., Na, T., & Wang, T. (2023). Jīyú màibó xīn diàn xìnhào pèngzhuàng shāng de xiětáng kòngzhì zhuàngtài jiǎncè yánjiū [Research of blood glucose control state detection based on percussion entropy]. *Acta Metrologica Sinica*, *44*(04), 657–663. https://doi.org/10.3969/j.issn.1000-1158.2023.04.25

[9] Hettiarachchi, C., & Chitraranjan, C. (2019). A machine learning approach to predict diabetes using short recorded photoplethysmography and physiological characteristics. In *Artificial Intelligence in Medicine: 17th Conference on*

Artificial Intelligence in Medicine, 322–327. https://doi.org/10.1007/978-3-030-21642-9_41

[10] Saha, S., Saha, U., Saha, S., & Fattah, S. A. (2024). PPG-based feature extraction for type II diabetes prediction: A machine learning approach. In *2nd International Conference on Computer, Communication and Control*, 1–6. https://doi.org/10.1109/IC457434.2024.10486372

[11] Li, S., & Chen, X. (2022). Continuous non-invasive blood glucose detection method based on PSO-GRU. In *4th International Conference on Information Science, Electrical, and Automation Engineering, 12257*, 376–381. https://doi.org/10.1117/12.2640183

[12] Zhang, M., Zhang, T., Zhong, M., & Cheng, Y. (2022). Stacking jíchéng xuéxí suànfǎ yànzhèng dòngmài sǔnshāng duì tángniàobìng zǎoqí jiǎncè de yìyì [Verifying the significance of arterial injury for early detection of diabetes by Stacking ensemble learning algorithm]. *Chinese Journal of Medical Physics*, *39*(8), 1003–1009. https://doi.org/10.3969/j.issn.1005-202X.2022.08.015

[13] Bavkar, V. C., & Shinde, A. A. (2021). Machine learning algorithms for diabetes prediction and neural network method for blood glucose measurement. *Indian Journal of Science and Technology*, *14*(10), 869–880. https://doi.org/10.17485/IJST/v14i10.2187

[14] Zhang, G., Mei, Z., Zhang, Y., Ma, X., Lo, B., Chen, D., & Zhang, Y. (2020). A noninvasive blood glucose monitoring system based on smartphone PPG signal processing and machine learning. *IEEE Transactions on Industrial Informatics*, *16*(11), 7209–7218. https://doi.org/10.1109/TII.2020.2975222

[15] Shi, B., Dhaliwal, S. S., Soo, M., Chan, C., Wong, J., Lam, N. W. C., . . . , & Ang, S. B. (2023). Assessing elevated blood glucose levels through blood glucose evaluation and monitoring using machine learning and wearable photoplethysmography sensors: Algorithm development and validation. *JMIR AI*, *2*, e48340. https://doi.org/10.2196/48340

[16] Lu, W. R., Yang, W. T., Chu, J., Hsieh, T. H., & Yang, F. L. (2022). Deduction learning for precise noninvasive measurements of blood glucose with a dozen rounds of data for model training. *Scientific Reports*, *12*(1), 6506. https://doi.org/10.1038/s41598-022-10360-3

[17] Porumb, M., Stranges, S., Pescapè, A., & Pecchia, L. (2020). Precision medicine and artificial intelligence: A pilot study on deep learning for hypoglycemic events detection based on ECG. *Scientific Reports*, *10*(1), 170. https://doi.org/10.1038/s41598-019-56927-5

[18] Sen Gupta, S., Kwon, T. H., Hossain, S., & Kim, K. D. (2021). Towards non-invasive blood glucose measurement using machine learning: An all-purpose PPG system design. *Biomedical Signal Processing and Control*, *68*, 102706. https://doi.org/10.1016/j.bspc.2021.102706

[19] Goutham, S., Ravi, V., & Kutti Padannayil, S. (2018). Diabetes detection using deep learning algorithms. *ICT Express*, *4*(4), 243–246. https://doi.org/10.1016/j.icte.2018.10.005

[20] Yildirim, O., Talo, M., Ay, B., Baloglu, U. B., Aydin, G., & Acharya, U. R. (2019). Automated detection of diabetic subject using pre-trained 2D-CNN models with frequency spectrum images extracted from heart rate signals. *Computers in Biology and Medicine*, *113*, 103387. https://doi.org/10.1016/j.compbiomed.2019.103387

[21] Wang, L., Mu, Y., Zhao, J., Wang, X., & Che, H. (2020). IGRNet: A deep learning model for non-invasive, real-time diagnosis of prediabetes through electrocardiograms. *Sensors*, *20*(9), 2556. https://doi.org/10.3390/s20092556

[22] Imam, M. H., Karmakar, C. K., Khandoker, A. H., Jelinek, H. F., & Palaniswami, M. (2016). Heart rate independent QT variability component can detect subclinical cardiac autonomic neuropathy in diabetes. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 928–931. https://doi.org/10.1109/EMBC.2016.7590853

[23] Jelinek, H. F., Cornforth, D. J., & Kelarev, A. V. (2016). Machine learning methods for automated detection of severe diabetic neuropathy. *Journal of Diabetic Complications & Medicine*, *1*(2), 1000108. https://doi.org/10.4172/2475-3211.1000108

[24] Rashid, M., Alkhodari, M., Mukit, A., Ahmed, K. I. U., Mostafa, R., Parveen, S., & Khandoker, A. H. (2022). Machine learning for screening microvascular complications in type 2 diabetic patients using demographic, clinical, and laboratory profiles. *Journal of Clinical Medicine*, *11*(4), 903. https://doi.org/10.3390/jcm11040903

[25] Zhang, H., Wang, L., & Li, X. (2019). Yī zhǒng yòng CEEMDAN hé páiliè shāng qùchú màibó xìnhào zàoshēng de fāngfǎ [A method to remove pulse signal noise using CEEMDAN and permutation entropy]. *China Sciencepaper*, *14*(3), 250–254. https://doi.org/10.3969/j.issn.2095-2783.2019.03.003

[26] Wu, X., Lin, L., Chen, H., & Xu, Z. (2021). Jīyú xiǎobō yùzhí fǎ de màibó bō qù zào suànfǎ [Pulse wave denoising algorithm based on the wavelet Threshold method]. *Beijing Biomedical Engineering*, *40*(1), 38–45. https://doi.org/10.3969/j.issn.1002-3208.2021.01.005

[27] Chen, Z., Wu, X., & Zhao, F. (2019). EEMD jiéhé xiǎobō yùzhí de guāngdiàn róngjī màibó bō xìnhào jiàng zào [Denoising and implementation of a photoplethysmography signal based on EEMD and wavelet threshold]. *Optics and Precision Engineering*, *27*(6), 1327–1334. https://doi.org/10.3788/ope.20192706.1327

[28] Li, C., Kong, L., Dong, Y., Yang, S., & Huang, T. (2024). Jīyú CEEMDAN hé xiǎobō bāo fēnjiě de zhámén zhèndòng xìnhào jiàng zào yánjiū [Research on noise reduction in gate vibration signals based on CEEMDAN and wavelet packet decomposition]. *Mechanical & Electrical Technique of Hydropower Station*, *47*(1), 16–18. https://doi.org/10.13599/j.cnki.11-5130.2024.01.005

[29] Liu, Q., Xu, Y., Liang, C., & Yuan, Y. (2023). Jīyú CEEMDAN hé xiǎobō shāng de xīnyīn xìnhào qù zào suànfǎ yánjiū [Research on heart sound signal denoising algorithm based on CEEMDAN and wavelet entropy]. *Computer Simulation*, *40*(2), 321–325. https://doi.org/10.3969/j.issn.1006-9348.2023.02.059

[30] Lou, H., Xing, H., Li, J., & Shi, C. (2023). Jīyú gǎijìn CEEMDAN hé xiǎobō yùzhí de yǔ shēng xìnhào qù zào suànfǎ yánjiū [Research on denoising algorithm of rain signal based on improved CEEMDAN and wavelet threshold]. *Electronic Measurement Technology*, *46*(7), 103–109. https://doi.org/10.19651/j.cnki.emt.2211278

How to Cite: Wu, Q., Gong, S., Hu, M., Pan, L., Liu, X., & Tao, W. (2025). Study on Diabetes Detection Based on DBO-RF Pulse Wave Analysis. *Smart Wearable Technology*. https://doi.org/10.47852/bonviewSWT52025979