

A Lightweight Learning Optimization Method for Precision Diagnosis of Breast Cancer: Cray-ELM

Jie Fang¹, Guangjie Sun¹, Ying Wang², Bingcang Huang^{2*}

¹ School of Gongli Hospital Medical Technology, University of Shanghai for Science and Technology, Shanghai, China

² Department of Radiology, Gongli Hospital of Shanghai Pudong New Area, Shanghai, China

Email Address

hbc9209@sina.com (Bingcang Huang)

*Correspondence: hbc9209@sina.com

Received: 20 December 2025; **Accepted:** 18 January 2026; **Published:** 3 February 2026

Abstract:

To address clinical challenges in breast cancer diagnosis, such as low diagnostic accuracy and detection delays, this study proposes a lightweight learning optimization method named Cray-ELM. The methodology integrates Principal Component Analysis (PCA) for dimensionality reduction—retaining 15 key components—with a Crawfish Optimization Algorithm (COA) and an Extreme Learning Machine (ELM). The COA is specifically employed to optimize ELM parameters, including input weights and hidden biases, thereby enhancing convergence stability and generalization. Experimental validation indicates that the Cray-ELM method achieves an average diagnostic accuracy of 98.95%, with a specificity of 98.98%, sensitivity of 97.67%, and an F1-score of 98.65%. Comparative experiments further indicate that this approach shows notable advantages over mainstream optimization methods across several key performance metrics. Ultimately, the Cray-ELM method offers a high-accuracy, fast-execution, and scalable solution, providing a robust tool for early breast cancer screening and intelligent medical diagnosis.

Keywords:

Breast Cancer Diagnosis, Principal Component Analysis, Crawfish Optimization Algorithm, Extreme Learning Machine

1. Introduction

Breast cancer is one of the most common malignant tumors among women worldwide [1], characterized by insidious onset, rapid progression, and high metastatic potential, posing serious threats to both physical health and psychological well-being. With lifestyle changes and an aging population, the incidence of breast cancer continues to rise and is showing a trend toward younger onset [2,3,4]. Although treatment methods have significantly advanced in recent years—with the application of surgery, radiotherapy, chemotherapy, targeted therapy, and endocrine therapy greatly improving patient survival—early detection remains the key to increasing the cure rate and improving prognosis [5,6]. Therefore, leveraging

advanced technologies to achieve rapid and accurate diagnosis of breast cancer, particularly at the early stages, has become a major focus in clinical medicine and intelligent healthcare research [7]. Traditional breast cancer diagnosis primarily relies on clinical examination, imaging techniques, and needle biopsy [8]. However, commonly used mammography (X-ray imaging of the breast) has low sensitivity in dense breast tissue, making it prone to missing early-stage microlesions [9]. Magnetic Resonance Imaging (MRI) while highly sensitive, is costly, time-consuming, and prone to false positives, which can lead to unnecessary follow-up procedures [10]. Ultrasound can compensate for some of mammography's limitations, but its accuracy heavily depends on the operator's experience and often suffers from poor reproducibility and consistency [11]. Additionally, cytological analysis of biopsy samples still faces challenges in the early detection of breast cancer. Moreover, these traditional diagnostic methods often require multidisciplinary collaboration and specialized equipment, resulting in complex workflows that hinder rapid, accessible, and routine screening. Thus, there is an urgent need to develop a new auxiliary diagnostic technology for breast cancer that is non-invasive, rapid, low-cost, and easy to operate, in order to enhance the efficiency of early screening and promote precise breast cancer prevention and control.

As a key component of artificial intelligence, machine learning has attracted widespread attention in the field of medical diagnosis and has demonstrated significant potential and application value [12]. For example, Mirai, developed by Yala A. and colleagues at MIT [13], is a deep learning model based on mammography that can predict breast cancer risk at multiple time points. It leverages information from potentially missing risk factors and ensures consistent predictions across different mammography devices. Yaqoob et al. [14] integrated the Seagull Optimization Algorithm (SGA) for feature selection with RF for breast cancer classification, improving the efficiency of processing high-dimensional gene expression data. Zeng et al. [15] combined serum Raman spectroscopy with deep learning algorithms to diagnose triple-negative and HER2-positive breast cancers, achieving promising diagnostic results. Abrar Yaqoob et al. [16] proposed a method that integrates Mutual Information (MI) and Particle Swarm Optimization (PSO) for gene selection. MI serves as an initial filter, while PSO refines the selection to identify the optimal gene subset for accurate classification. These advances highlight the broad application prospects of machine learning in breast cancer diagnosis and provide strong support for clinical decision-making. [17]

2. Methods

2.1. Data collection

This paper uses the publicly accessible Wisconsin Breast Cancer Dataset (<https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>), which was collected by Dr. William H. Wolberg at the University of Wisconsin Hospital in the United States [18]. The dataset contains a total of 569 samples, with 357 benign samples and 212 malignant samples. There are 32 attributes in total, including an ID number, a diagnosis result (M for malignant, B for benign), and 30 input features. The specific features are shown in Table 1.

Table 1. Feature names.

Index	Feature	Index	Feature	Index	Feature
1	<i>Mean Radius</i>	11	<i>Radius Standard Error</i>	21	<i>Worst Radius</i>
2	<i>Mean Texture</i>	12	<i>Texture Standard Error</i>	22	<i>Worst Texture</i>
3	<i>Mean Area</i>	13	<i>Perimeter Standard Error</i>	23	<i>Worst Perimeter</i>
4	<i>Mean Perimeter</i>	14	<i>Area Standard Error</i>	24	<i>Worst Area</i>
5	<i>Mean Smoothness</i>	15	<i>Smoothness Standard Error</i>	25	<i>Worst Smoothness</i>
6	<i>Mean Compactness</i>	16	<i>Compactness Standard Error</i>	26	<i>Worst Compactness</i>
7	<i>Mean Concavity</i>	17	<i>Concavity Standard Error</i>	27	<i>Worst Concavity</i>
8	<i>Mean Concave Points</i>	18	<i>Concave Points Std. Error</i>	28	<i>Worst Concave Points</i>
9	<i>Mean Symmetry</i>	19	<i>Symmetry Standard Error</i>	29	<i>Worst Symmetry</i>
10	<i>Mean Fractal Dimension</i>	20	<i>Fractal Dimension Std. Error</i>	30	<i>Worst Fractal Dimension</i>

2.2. PCA-COA-ELM algorithm

The PCA-COA-ELM model achieves efficient breast cancer diagnosis through a well-structured sequence of steps, addressing key challenges in high-dimensional medical data analysis. First, medical datasets often contain a large number of features, some of which may be irrelevant or redundant, potentially degrading model performance. To mitigate this issue, PCA is utilized for dimensionality reduction. PCA effectively retains the most informative features while discarding feature redundancy, thereby enhancing computational efficiency and preventing overfitting. This preprocessing step ensures that the subsequent classification model operates on a more compact and representative dataset.

Next, the COA is integrated to optimize the parameters of the ELM, particularly the input weights and biases. ELM, a single-hidden-layer feedforward neural network, is known for its rapid learning speed and strong generalization capability. However, its performance is highly sensitive to parameter selection. By leveraging COA's metaheuristic optimization, the model achieves improved stability and adaptability across diverse breast cancer datasets. This step enhances classification robustness by ensuring that the ELM model is tuned optimally, reducing training time while improving accuracy. Once optimized, the ELM efficiently classifies tumors as benign or malignant. Compared to traditional models, PCA-COA-ELM reduces reliance on manual feature extraction through PCA, lowers computational complexity with COA's automated parameter tuning, and improves performance on complex, high-dimensional datasets with better generalization. By integrating PCA for dimensionality reduction, COA for optimization, and ELM for classification, this model enhances accuracy and efficiency in breast cancer diagnosis. It also contributes to more automated and reliable medical diagnostic tools, with potential for broader application across imaging modalities and patient demographics. The detailed workflow is illustrated in Figure 1.

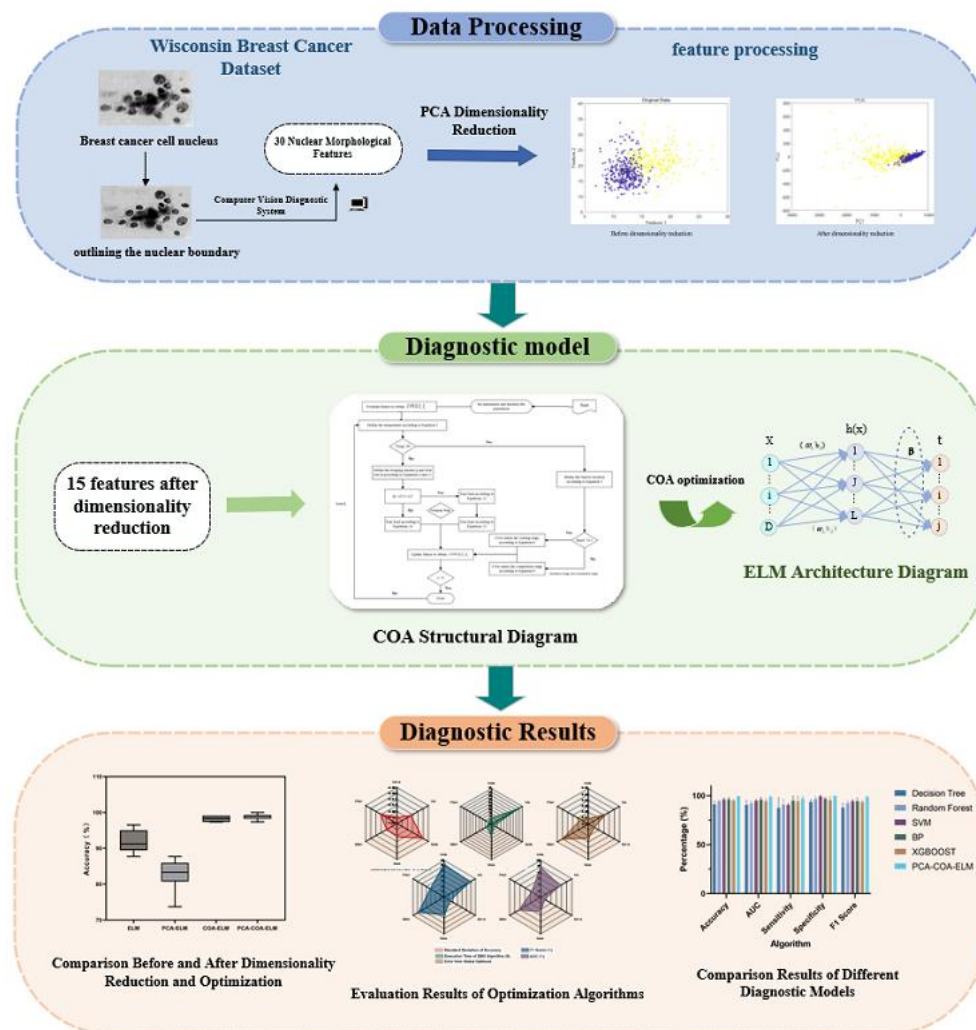


Figure 1. Flowchart of this study.

The research process comprises three key stages: data preprocessing, diagnostic modeling, and result evaluation. Initially, 30 breast cancer cell nucleus features extracted from the Wisconsin Breast Cancer Dataset are subjected to PCA for dimensionality reduction. Subsequently, a diagnostic model is constructed using the PCA-COA-ELM algorithm. Finally, the model's diagnostic performance is evaluated to assess its effectiveness and reliability.

2.2.1. Principal component analysis

Principal Component Analysis is a commonly used dimensionality reduction algorithm, widely applied in data analysis and machine learning. The core idea is to transform high-dimensional data into low-dimensional data through linear transformations while retaining as much of the original data's features and information as possible [19]. PCA is a commonly used dimensionality reduction method that transforms high-dimensional data into a lower-dimensional space while preserving as much important information as possible. Before applying PCA, data is typically standardized to eliminate scale differences between variables. Next, the covariance matrix of the standardized data is computed, which reflects the relationships between different variables. The covariance matrix is then decomposed into its eigenvalues and eigenvectors. The top K eigenvalues and their corresponding eigenvectors are selected as the principal components, as they capture the most

significant variations in the data. By identifying the primary sources of variation, PCA not only aids in data interpretation but also reduces model complexity, improving the efficiency of subsequent analysis and modeling, making data processing more effective and reliable.

2.2.2. Crayfish optimization algorithm

The Crayfish Optimization Algorithm is an emerging population-based intelligent optimization algorithm inspired by the behavior of crayfish when searching for food [20]. This algorithm solves complex optimization problems by simulating the foraging paths and collective cooperation mechanisms of crayfish. The COA algorithm flowchart is shown in Figure 2.

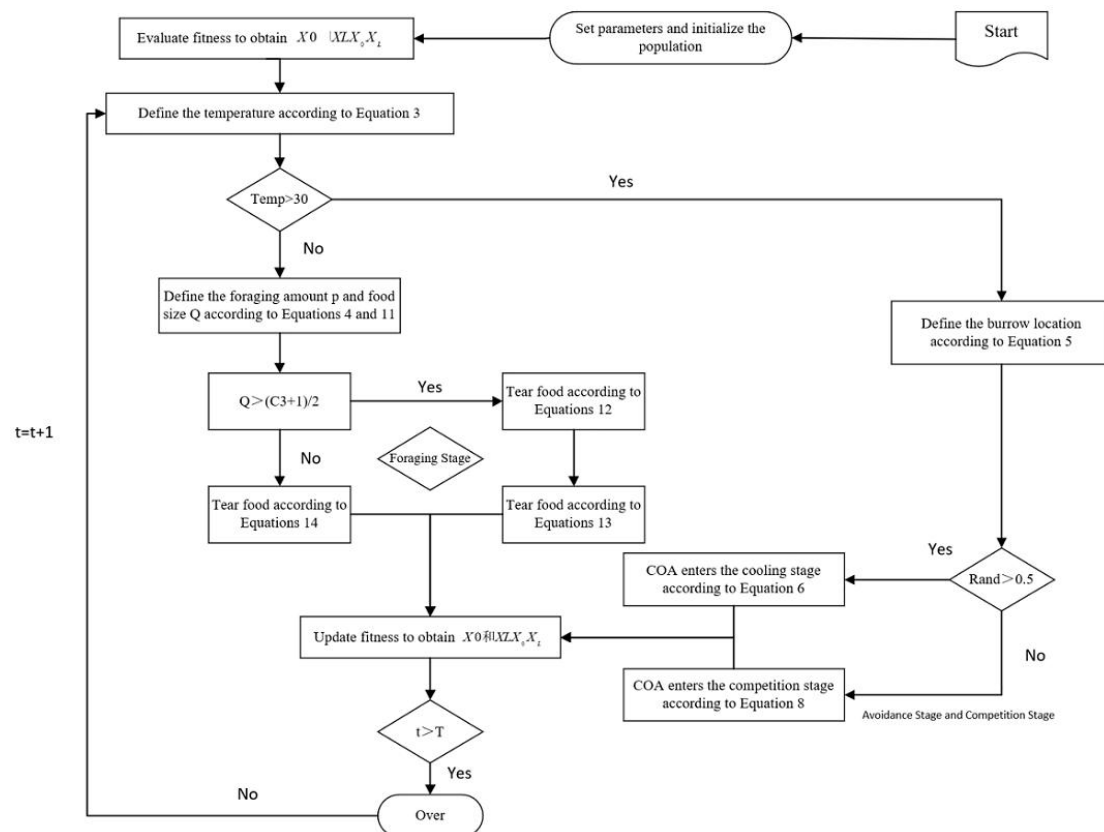


Figure 2. Flowchart of the COA. The execution process is divided into three main stages: avoidance, competition, and foraging.

2.2.3. Extreme learning machine

Extreme Learning Machine is an algorithm proposed by Professor Guang-Bin Huang [21] from Nanyang Technological University for training single-hidden-layer feedforward neural networks (SLFN). ELM randomly generates input layer weights and hidden layer biases, then uses an activation function to compute the hidden layer output. Compared to traditional methods, ELM offers advantages such as fast training speed, good generalization performance, and reduced complexity of traditional neural networks [22].

Assume there are N samples, $\{X_i, t_i | X_i \in R^D, t_i \in R^m, i = 1, 2, \dots, N\}$ where X_i represents the i -th data in the sample, t_i denotes the label of the i -th data, and L is the number of hidden nodes. In the ELM structure, from left to right, there are the input

layer x , the hidden layer $h(x)$, and the output layer. The formula for the hidden layer output $H(x)$ is as follows:

$$H(x) = [h_1(x), \dots, h_L(x)] \quad (1)$$

Where $h_i(x)$ is the output of the i -th hidden layer node, and the formula is as follows:

$$h_i(x) = g(w_i \cdot b_i, x) = g(w_i x + b_i), w_i \in R^D, b_i \in R \quad (2)$$

Where $g(w_i \cdot b_i, x)$ represents the activation function, and (w_i, b_i) denotes the parameters of function.

$$g(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \quad (3)$$

After the hidden layer performs feature mapping through the Sigmoid activation function, the output obtained is as follows:

$$f_{L(x)} = \sum_{i=1}^L \beta_i h_i(x) = H(x) \beta \quad (4)$$

Where β of $\beta_i = [\beta_1, \dots, \beta_L]^T$ represents the output weights between the hidden layer and the output layer.

2.3. Statistical analyses

To evaluate the diagnostic performance of the proposed Principal Component Analysis–Crayfish Optimization Algorithm–Extreme Learning Machine (PCA-COA-ELM) model, the following key metrics were employed:

Accuracy: The proportion of correctly classified benign and malignant cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

Area Under the Curve (AUC): Measures the model's ability to distinguish between malignant and benign cases across different thresholds. A higher AUC indicates better classification performance.

$$\text{AUC} = \frac{1}{P \cdot N} \sum_{i=1}^P \sum_{j=1}^N \mathbb{I}(s_i > s_j) \quad (6)$$

Sensitivity (Recall): The proportion of actual malignant cases correctly identified by the model, which is critical for reducing false negatives in early detection.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

Specificity: The proportion of actual benign cases correctly classified, helping to minimize unnecessary interventions.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

F1 Score: The harmonic mean of precision and recall, providing a balanced measure of model performance.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

Among them, TP (True Positive) refers to cases predicted as malignant and actually malignant; TN (True Negative) refers to cases predicted as benign and actually benign; FP (False Positive) refers to cases predicted as malignant but actually benign; and FN (False Negative) refers to cases predicted as benign but actually malignant. P represents the total number of actual malignant samples, i.e., $TP + FN$, and N represents the total number of actual benign samples, i.e., $TN + FP$. Precision is the proportion of correctly predicted malignant cases among all predicted malignant cases, and its formula is: $\text{Precision} = \frac{TP}{TP + FP}$.

To ensure the robustness of the model, the dataset was split into 80 % for training and 20 % for testing. Five independent trials were conducted using PCA for dimensionality reduction, and random partitioning was applied to assess the model's stability. Performance metrics (accuracy, AUC, sensitivity, and specificity) were recorded for each trial, and the mean $\pm$ standard deviation was reported to evaluate consistency.

To verify the effectiveness of the COA, it was compared with CS, BOA, SSA, DBO and PSO. The comparison included algorithm stability (standard deviation of accuracy), execution time, optimal solution error, F1 score, and AUC. Confusion matrices were also generated to visually present true positives, false positives, true negatives, and false negatives.

The PCA-COA-ELM model was benchmarked against five widely used classifiers: DT, RF, SVM, BP, and XGBoost. The comparison focused on diagnostic accuracy, computational efficiency (training/prediction time), sensitivity-specificity trade-off, AUC, and F1 score.

3. Research

3.1. Feature dimensionality reduction

The dataset used in this study contains 30 features. PCA was applied to reduce the dimensionality of these 30 features, and the visualization of the results is shown in Figure 3.

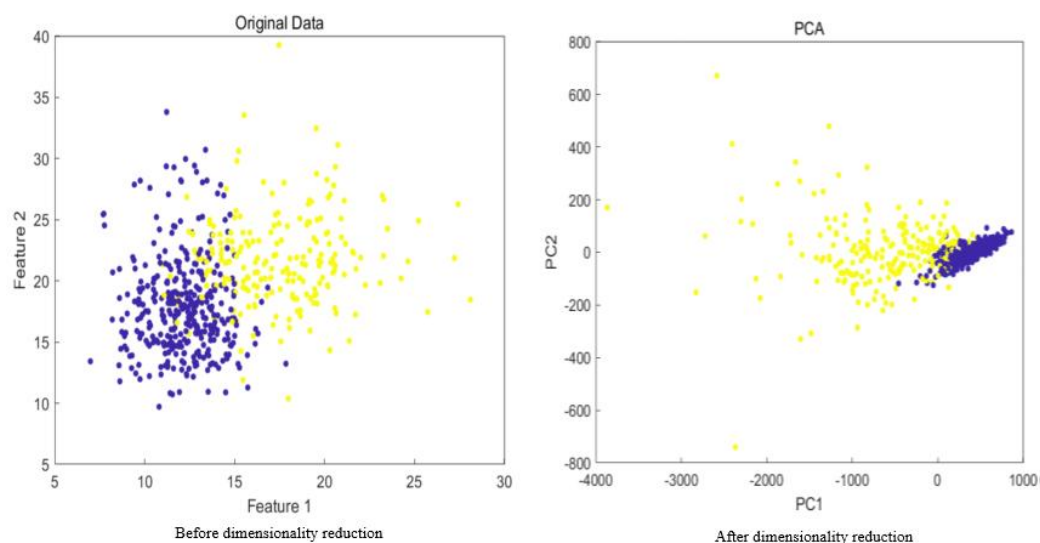


Figure 3. Visualization of data before and after dimensionality reduction.

The left figure illustrates the distribution of the original data, whereas the right figure presents the PCA-transformed results. The left plot represents the original data distribution, where yellow and purple dots indicate benign and malignant cases, respectively. Although there are some differences in their distribution in the original feature space, a significant overlap still exists, suggesting that the distinguishing power of these features is limited. The right plot shows the results after applying PCA, with the horizontal and vertical axes representing the first principal component (PC1) and the second principal component (PC2), respectively. These two principal components retain most of the variance in the data. It can be observed that after PCA processing, the separation between yellow and purple dots has improved, and the major variations in the data are concentrated in PC1 and PC2. This enhances the ability to differentiate between different categories. Overall, PCA removes redundant information and preserves the primary characteristics of the data in a lower-dimensional space, facilitating a better understanding of the intrinsic structure of the data.

3.2. Evaluate the model performance with different numbers of features

The number of features (k) retained after PCA dimensionality reduction can have a certain impact on classification results. If too few features are retained, information loss may occur, reducing the model's accuracy and predictive capability. On the other hand, retaining too many features may introduce redundant information and increase computational complexity. Therefore, in this study, we set the number of retained features to 5, 10, 15, 20, and 25 for analysis and discussion to evaluate the impact of feature count on model performance. This helps us determine the optimal balance between information retention and model complexity. The classification accuracy for different numbers of features is shown in Table 2.

Table 2. Performance Metrics under Different Numbers of Features.

Evaluation index	Number of times	Number of features				
		5	10	15	20	25
Accuracy (%)	1	97.37	99.12	99.12	98.25	97.37
	2	96.49	97.37	98.25	98.25	99.12
	3	97.37	98.25	98.25	97.37	99.12
	4	96.49	98.25	99.12	97.37	97.37
	5	95.61	96.49	100	98.25	96.49
	average	96.67±0.66	97.9±1.00	98.95±0.73	97.90±0.48	97.89±1.18
AUC (%)	1	96.74	98.94	98.91	97.73	96.05
	2	95.56	97.44	97.96	98.61	97.78
	3	96.88	98.15	97.44	96.43	98.68
	4	95.49	98.13	98.84	96.51	97.01
	5	94.64	95.57	100	97.67	95.92
	average	95.87±0.94	97.65±1.28	98.63±0.99	97.59±0.95	97.09±1.17
Sensitivity (%)	1	93.48	97.87	97.83	95.46	92.11
	2	95.56	97.44	97.96	98.61	97.78
	3	95.12	97.73	94.87	92.86	97.37
	4	92.31	97.67	97.67	93.02	95.46
	5	90.70	92.50	100	95.35	91.84
	average	93.43±2.01	96.64±2.32	97.67±1.65	95.06±2.09	94.91±2.52
Specificity (%)	1	100	100	100	100	100
	2	91.11	97.73	95.91	97.22	98.89
	3	98.63	98.57	100	100	100

	4	98.67	98.59	100	100	98.57
	5	98.59	98.65	100	100	100
	average	97.40±3.18	98.71±0.73	98.98±1.77	99.44±1.11	99.49±0.63
<i>F1 score (%)</i>	1	96.63	99.12	99.12	98.25	97.37
	2	95.35	96.63	97.92	97.67	98.88
	3	96.30	97.73	97.37	96.30	98.67
	4	94.74	97.67	98.82	96.39	96.55
	5	93.98	94.87	100	97.62	95.75
	average	95.20±1.00	97.00±1.42	98.65±0.92	97.25±0.77	97.71±1.23

Each algorithm was evaluated five times to assess its performance in terms of accuracy, AUC, sensitivity, specificity, and F1 score under different numbers of features. Based on the data in Table 2, it is evident that the classification accuracy reaches its highest when the number of features is reduced to 15. This indicates that among the tested feature counts of 5, 10, 15, 20, and 25, selecting 15 features achieves the optimal balance between preserving key information and reducing redundancy. Therefore, in subsequent model training, choosing 15 features as the input after dimensionality reduction can effectively enhance classification performance. This result provides an important reference for optimizing breast cancer diagnosis models.

3.3. COA-ELM diagnostic model

When training an ELM, key parameters such as input weights, biases, and the number of hidden layer nodes need to be manually set. If these parameters are not accurately configured, they can significantly impact the results, severely affecting the classification accuracy of the diagnostic model [23]. Therefore, this study proposes an optimization method based on the COA to optimize ELM parameters, thereby improving model performance and generalization ability. By fine-tuning these parameters, the optimal network structure can be identified, enabling the model to achieve higher accuracy on both training and test datasets while enhancing robustness. This approach helps balance model complexity and prevents overfitting. In this study, after applying PCA for dimensionality reduction, the proposed COA-optimized ELM diagnostic model was evaluated on a breast cancer diagnosis dataset, and the accuracy comparison results are shown in Figure 4.

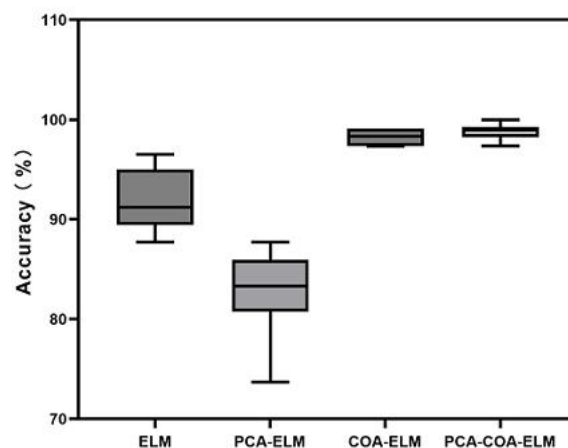


Figure 4. Comparison of model accuracy before and after dimensionality reduction and optimization.

By inputting the 15 features obtained through PCA dimensionality reduction into four models ELM, PCA-ELM, COA-ELM, and PCA-COA-ELM and conducting

seven tests, it was observed that the proposed PCA-COA-ELM model achieved significantly higher accuracy. This indicates that the model excels in both feature extraction and classification capability, effectively improving the accuracy of breast cancer diagnosis. Compared to the other three models, PCA-COA-ELM not only maintained stable performance across all tests but also demonstrated superior generalization ability, confirming its potential and reliability for clinical applications.

3.4. Performance comparison of different optimization algorithms

To evaluate the performance of the COA in the PCA-ELM model, this study employed a variety of optimization algorithms for parameter tuning and conducted comparative analyses. These algorithms include CS, BOA, SSA, DBO, and PSO. Their performances were assessed using multiple evaluation metrics, including: Algorithm stability, measured by the standard deviation of classification accuracy; Execution time of each algorithm; Optimal solution error, calculated by comparing the best solution found by each algorithm to the theoretical optimum; F1 score and AUC, providing a comprehensive evaluation of diagnostic effectiveness in breast cancer classification tasks.

Each optimization algorithm was independently tested five times. The resulting diagnostic performances are illustrated in Figure 5. To further analyze the classification effectiveness of each algorithm in breast cancer diagnosis, Figure 6 presents the corresponding confusion matrices, which intuitively display classification outcomes such as accuracy, misclassification rates, and other key indicators.

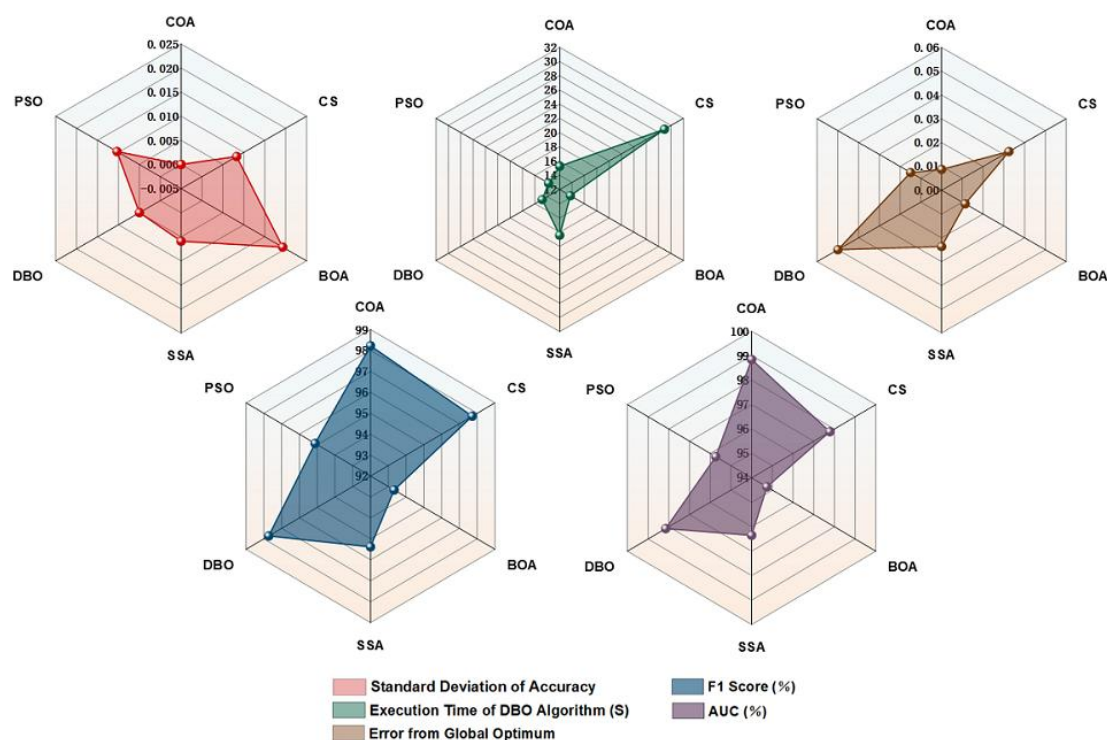


Figure 5. Evaluation Results of Optimization Algorithms Based on Five Metrics.

As shown in Figure 5, the performance of different optimization algorithms within the PCA-ELM model varies significantly across the five evaluation metrics. The COA demonstrates the best performance in key indicators such as the standard deviation of accuracy, F1 score, and AUC, indicating strong advantages in model stability, classification accuracy, and discriminative ability. Additionally, COA shows a

relatively low optimal solution error, reflecting its robust global search capability and effectiveness in approaching the optimal solution. Although its execution time is slightly longer compared to some algorithms like the CS and SSA, the overall efficiency remains within an acceptable range. Therefore, COA is considered the most suitable optimization algorithm for the breast cancer diagnosis task, offering enhanced model performance while maintaining high stability and reasonable computational time.

Based on the confusion matrix results (a–f) presented in Figure 6 for six optimization algorithms applied to the breast cancer diagnosis task, the classification performance for malignant and benign samples was evaluated. Overall, all optimization algorithms demonstrated high classification accuracy, highlighting the potential of optimization-based diagnostic models in intelligent medical image recognition.

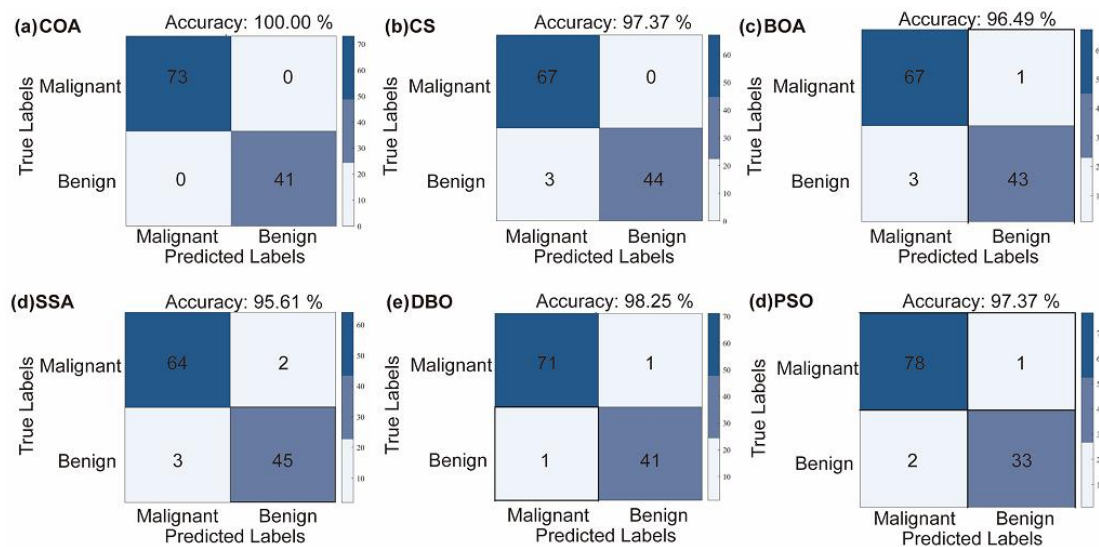


Figure 6. Confusion Matrix Results.

The confusion matrices of breast cancer classification results under different combinations of optimization algorithms and ELM classifiers are presented, including six algorithms: (a) COA, (b) CS, (c) BOA, (d) SSA, (e) DBO, and (f) PSO.

Among them, the COA (Figure 6a) achieved a perfect classification accuracy of 100.00 %, making it the most outstanding performer. All test samples were correctly classified without any misclassifications, demonstrating strong robustness and generalization capability. The CS (Figure 6b) and PSO (Figure 6f) also showed excellent performance, both with an accuracy of 97.37 %. PSO was particularly precise in identifying malignant samples, with only a small number of benign samples misclassified. The DBO (Figure 6e) followed closely with an accuracy of 98.25 %, maintaining both high accuracy and strong stability. In contrast, the BOA (Figure 6c) and the SSA (Figure 6d) achieved accuracies of 96.49 % and 95.61 %, respectively. Although slightly lower than the other algorithms, they still correctly classified the vast majority of samples, making their classification performance practically valuable.

In summary, the COA demonstrated the best performance under the experimental setup of this study, making it the most promising optimization strategy for breast cancer diagnosis. Meanwhile, algorithms such as PSO and DBO also showed strong competitiveness. The confusion matrix not only provides classification accuracy but

also reveals the classification bias of each algorithm for different types of samples, offering valuable insights for future algorithm selection and improvement.

3.5. Comparison of different classifiers

With the rapid development of machine learning and artificial intelligence, we introduced five classical algorithms—Decision Tree (DT), RF, SVM, Backpropagation Neural Network (BP Neural Network), and XGBoost—into our existing model framework (PCA-COA-ELM) for comparative analysis. The accuracy, AUC, sensitivity, specificity, and F1 score of each algorithm are shown in Figure 7.

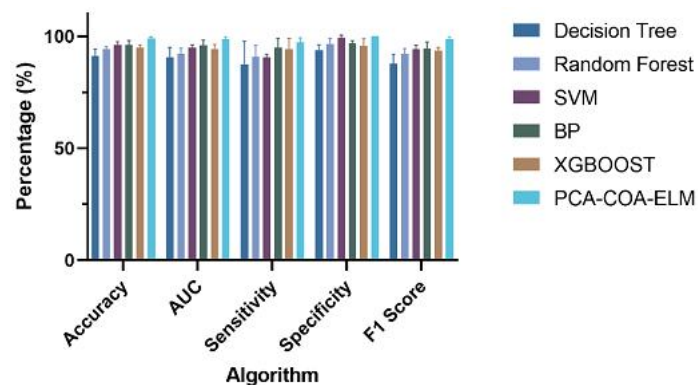


Figure 7. Performance comparison results of different diagnostic models.

From Figure 7, it is evident that the PCA-COA-ELM model performs excellently across all evaluation metrics, showing a clear advantage compared to other classic algorithms. In terms of accuracy, AUC, sensitivity, specificity, and F1 score, PCA-COA-ELM consistently outperforms the other methods. Especially in AUC and F1 score, the model demonstrates strong classification ability and good balance, effectively distinguishing between benign and malignant breast cancer samples. Moreover, PCA-COA-ELM excels in balancing sensitivity and specificity, which helps minimize both false negatives and false positives in breast cancer diagnosis. Although classic algorithms such as Decision Tree (DT), RF, SVM, BP Neural Network, and XGBoost also achieve relatively high scores across various metrics, there remains a noticeable gap compared to PCA-COA-ELM. Therefore, by integrating the strengths of PCA, COA, and ELM, the PCA-COA-ELM algorithm shows greater potential in breast cancer diagnosis, providing a more efficient and accurate solution for early screening and diagnosis in this field.

4. Discussion

This study proposes a breast cancer diagnosis method Cray-ELM: A Lightweight Optimized Learning Framework for Intelligent Breast Cancer Diagnosis. SVM and Logistic Regression (LR) also achieve high accuracy in breast cancer classification but often require complex feature selection and parameter tuning. For example, Laghmati et al. [24] proposed a supervised CAD system combining feature selection, PCA, and ensemble models (such as Majority Voting and Stacking with Logistic Regression), achieving 97.37 % accuracy on the Wisconsin Breast Cancer Diagnostic dataset. Similarly, Ara et al.[25] and Naji et al. [26] employed Random Forest and SVM, reporting accuracies of 96.5 % and 97.2 %, respectively. These methods typically require significant manual intervention for feature selection and hyperparameter optimization, whereas PCA-COA-ELM enhances classification

performance through an automated optimization process, reducing the need for manual adjustments.

Deep learning methods have also made significant advancements in breast cancer diagnosis. For example, Sharmin et al. [27] developed a hybrid deep learning model combining ResNet50V2 and Light Gradient Boosting Classifier, achieving 95% accuracy. Additionally, Islam et al. [28] demonstrated that Artificial Neural Networks (ANNs) achieved the highest accuracy of 98.57 % on the Wisconsin Breast Cancer dataset, highlighting their effectiveness in early diagnosis. However, deep learning methods usually require large amounts of training data and computational resources. In contrast, PCA-COA-ELM performs well on small sample datasets while maintaining high computational efficiency.

The proposed PCA-COA-ELM method demonstrates significant advantages in feature extraction, model optimization, and computational efficiency. By integrating PCA for dimensionality reduction and COA for optimization, this approach efficiently processes high-dimensional data, enhancing model generalization and classification performance. Additionally, PCA-COA-ELM offers superior model interpretability, making it more suitable for resource-constrained environments. However, this study has limitations. The experiments were mainly conducted on the Wisconsin Breast Cancer dataset, which has a limited sample size and diversity, potentially affecting generalization. Future research should explore larger, more diverse datasets and integrate multimodal data to enhance accuracy and clinical applicability.

The proposed PCA-COA-ELM method demonstrates significant advantages in feature extraction, model optimization, and computational efficiency, making it particularly suitable for resource-constrained environments. Despite these achievements, several limitations should be acknowledged: (1) Data Limitations: The current validation is based solely on a single public dataset, with a limited sample size (569 cases) and lack of diversity (only cytological features from fine-needle aspiration), which may affect the model's generalizability. Unlike clinical practice involving multicenter and multimodal data, the model's ability to integrate critical diagnostic information such as ultrasound imaging and molecular pathology remains unverified. (2) Algorithmic Limitations: Although COA exhibits strong optimization performance, it still lags slightly behind newer metaheuristic algorithms, such as the Golden Jackal Optimizer introduced in 2023. Additionally, ELM, as a single-hidden-layer feedforward neural network (SLFN) without iterative training, while offering fast training speed, shows performance instability when dealing with imbalanced datasets—particularly in terms of reduced precision and specificity for minority class recognition. (3) Clinical Application Challenges: Medical AI models are required to meet strict regulatory standards. However, this study is based on single-center retrospective data and lacks prospective multicenter clinical validation. The model's generalizability across different medical devices and operational protocols still needs further investigation.

5. Conclusions

Multiple comparative experiments in this study demonstrate that the proposed model offers three key advantages: efficiency, accuracy, and practicality. PCA reduces computational complexity, while COA enables automated parameter optimization, minimizing manual intervention. The optimized ELM achieves a good

balance between sensitivity and specificity, significantly lowering the risk of misdiagnosis.

This model shows strong real-world applicability. It performs well on small-sample datasets, achieving high accuracy with only 569 cases, making it particularly suitable for use in primary healthcare settings. PCA assists in extracting key features (e.g., radius error, texture standard deviation), enhancing model interpretability. The automated feature selection and parameter tuning process also simplifies operation, making the model easier for clinicians to use.

Additionally, the model exhibits strong potential for extension and transferability. Its general architecture can be adapted to other disease types (such as lung cancer or thyroid cancer) and supports integration with other modalities, such as medical imaging or genomic data, laying the groundwork for multi-source fusion diagnostics.

In the future, integrating breast ultrasound, histopathological slide images, and dynamic follow-up data will help build a cross-modal diagnostic framework, improving the model's adaptability to complex cases and ensuring more precise diagnoses. Efforts will also focus on expanding clinical validation samples to assess the model's robustness across diverse populations and device variations, promoting deployment on mobile platforms to support rapid screening in rural and remote areas. Furthermore, incorporating dynamic learning mechanisms will enable continuous model optimization, with future extensions to include malignancy grading and treatment response prediction capabilities.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This paper was supported by National Natural Science Foundation of China (No. 82372029); Discipline Construction of Pudong New Area Health Commission (No. PWZxk2022-03); Investigator-initiated clinical study of Pudong New District Health Committee (No. 2025-PWDL-24); Shanghai Pudong New District Health Committee Health Industry Special Project (No. PW2024E-02).

References

- [1] Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024, 74, 229-263.
- [2] Tesch, M.E.; Partridge, A.H. Treatment of Breast Cancer in Young Adults. *Am Soc Clin Oncol Educ Book.* 2022, 42, 1-12.
- [3] Rossi, L.; Mazzara, C.; Pagani, O. Diagnosis and Treatment of Breast Cancer in Young Women. *Curr Treat Options Oncol.* 2019, 20, 86.
- [4] Harbeck, N.; Gnant, M. Breast cancer. *Lancet.* 2017, 389, 1134-1150.
- [5] Xiong, X. et al. Breast cancer: pathogenesis and treatments. *Signal Transduct Target Ther.* 2025, 10, 49.

- [6] Barzaman, K. et al. Breast cancer: Biology, biomarkers, and treatments. *Int Immunopharmacol.* 2020, 84, 106535.
- [7] Yan, J. et al. Diagnosis and Treatment of Breast Cancer in the Precision Medicine Era. *Methods Mol Biol.* 2020, 2204, 53-61.
- [8] Jafari, S.H. et al. Breast cancer diagnosis: Imaging techniques and biochemical markers. *J Cell Physiol.* 2018, 233, 5200-5213.
- [9] Bodewes, F.T.H.; Van Asselt, A.A.; Dorrius, M.D.; Greuter, M.J.W.; De Bock, G.H. Mammographic breast density and the risk of breast cancer: A systematic review and meta-analysis. *Breast.* 2022, 66, 62-68.
- [10] Wekking, D. et al. Breast MRI: Clinical Indications, Recommendations, and Future Applications in Breast Cancer Diagnosis. *Curr Oncol Rep.* 2023, 25, 257-267.
- [11] Gong, H.; Qian, M.; Pan, G.; Hu, B. Ultrasound Image Texture Feature Learning-Based Breast Cancer Benign and Malignant Classification. *Comput Math Methods Med.* 2021, 2021, 6261032.
- [12] Sekine, C.; Horiguchi, J. Current status and prospects of breast cancer imaging-based diagnosis using artificial intelligence. *Int J Clin Oncol.* 2024, 29, 1641-1647.
- [13] Yala, A. et al. Toward robust mammography-based models for breast cancer risk. *Sci Transl Med.* 2021, 13, aba4373.
- [14] Yaqoob, A. et al. SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study. *Sci Rep.* 2025, 15, 10944.
- [15] Zeng, Q. et al. Serum Raman spectroscopy combined with convolutional neural network for rapid diagnosis of HER2-positive and triple-negative breast cancer. *Spectrochim Acta A Mol Biomol Spectrosc.* 2023, 286, 122000.
- [16] Yaqoob, A.; Mir, M.A.; Jagannadha Rao, G.V.V.; Tejani, G.G. Transforming Cancer Classification: The Role of Advanced Gene Selection. *Diagnostics (Basel).* 2024, 14, 2632.
- [17] Sharma, N.; Sharma, K.P.; Mangla, M.; Rani, R. Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding. *Multimedia Tools and Applications.* 2023, 82, 4011-4029.
- [18] Street, W.N.; Wolberg, W.H.; Mangasarian, O.L. Electronic imaging.
- [19] Jolliffe, I.T.; Cadima, J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci.* 2016, 374, 20150202.
- [20] Yan, D.; Wang, H.; Gao, Y.; Tian, S.; Zhang, H. Based on improved crayfish optimization algorithm cooperative optimal scheduling of multi-microgrid system. *Sci Rep.* 2024, 14, 24871.
- [21] Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing.* 2006, 70, 489-501.
- [22] Li, H.T.; Chou, C.Y.; Chen, Y.T.; Wang, S.H.; Wu, A.Y. Robust and Lightweight Ensemble Extreme Learning Machine Engine Based on Eigenspace Domain for

- Compressed Learning. *IEEE Transactions on Circuits and Systems I: Regular Papers*. 2019, 66, 4699-4712.
- [23] Bilal, A. et al. BC-QNet: A quantum-infused ELM model for breast cancer diagnosis. *Computers in Biology and Medicine*. 2024, 175, 108483.
- [24] Laghmati, S.; Hamida, S.; Hicham, K.; Cherradi, B.; Tmiri, A. An improved breast cancer disease prediction system using ML and PCA. *Multimedia Tools and Applications*. 2024, 83, 33785-33821.
- [25] Ara, S.; Das, A.; Dey, A. 2021 International Conference on Artificial Intelligence (ICAI). 2021, 97-101.
- [26] Naji, M.A. et al. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Computer Science*. 2021, 191, 487-492.
- [27] Sharmin, S.; Ahammad, T.; Talukder, M.A.; Ghose, P. A Hybrid Dependable Deep Feature Extraction and Ensemble-Based Machine Learning Approach for Breast Cancer Detection. *IEEE Access*. 2023, 11, 87694-87708.
- [28] Islam, M.M. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*. 2020, 1, 290.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)