Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 20, No. 1, 2026

Research on Breast Cancer Diagnosis and Survival Prediction Based on Ensemble Learning

Jiayao Li, Haojie Huo, Guo Chen, Bo Zhang, Yinuo Tian

North China University of Science and Technology, Tangshan, Hebei, China

Abstract: To address the need for high-precision auxiliary diagnosis and survival prediction in breast cancer, a comprehensive diagnosis and treatment system has been developed. In terms of diagnosis, multivariate linear regression and the Enhanced Whale Optimization Algorithm (EWOA) are employed for feature selection and hyperparameter tuning, combined with stacking ensemble learning to improve the model's accuracy and generalization capability. For survival prediction, the Adaptive Synthetic Sampling (ADASYN) method is used to balance data distribution, and Multi-strategy Improved Nutcracker Optimization Algorithm (MSNOA) is proposed to optimize features and hyperparameters simultaneously. This is integrated with a decision tree-based ensemble learning algorithm (CatBoost) to accurately predict patient survival status. Finally, a multi-platform compatible auxiliary diagnosis and treatment system is developed based on Gradio and Flask frameworks, enabling remote diagnosis and prediction functions, providing intelligent support for clinical personalized treatment, and promoting the development of precision medicine.

Keywords: Breast Cancer; Auxiliary Diagnosis and Treatment System; Enhanced Whale Optimization Algorithm; Improved Starling Optimization Algorithm.

1. Introduction

Breast cancer, as the most common malignant tumor among women, requires early precise diagnosis and survival prediction to improve patient survival rates and optimize treatment plans. Traditional diagnostic methods rely heavily on physician experience, which is subjective and carries a risk of misdiagnosis. Therefore, developing high-precision, highly generalizable, and interpretable intelligent auxiliary diagnosis and survival prediction systems has become an important research direction in the field of precision medicine for breast cancer.

Currently, there have been numerous studies on breast cancer diagnosis both domestically and internationally. Reference [1] utilizes radiomics and Support Vector Machine (SVM) to achieve preoperative diagnosis of triple-negative breast cancer, providing a new approach for early diagnosis; however, a single model lacks sufficient generalization ability when dealing with complex, high-dimensional data. Considering the evident shortcomings of single models in handling complex data, Reference [2] compares algorithms such as Random Forest and XGBoost, confirming that ensemble strategies can effectively improve classification accuracy. Reference [3] enhances the classification performance of malignant tumors by optimizing Random Forest algorithms. Nonetheless, existing ensemble learning methods still face limitations in feature fusion and model combination. In terms of feature fusion, the challenge lies in accurately integrating features from different sources and types to eliminate redundancy and avoid conflicts; in model combination, it is crucial to determine how to allocate model weights and select combination strategies to fully leverage ensemble advantages and maximize performance. Reference [4] employs Whale Optimization Algorithm (WOA) for tuning parameter of Support Vector Machine (SVM). Reference [5] improves the Black Widow Optimization Algorithm for feature selection. Reference [6] integrates DenseNet201 and XGBoost to significantly enhance pathological image accuracy. Reference [7] applies Random Forest to process mammogram data. Reference [8] constructs an SVM model to improve generalization capability. Reference [9] proposes a confidence-weighted voting ensemble framework to boost prediction accuracy. While these studies have laid a certain technical foundation for breast cancer diagnosis, there remains a significant gap between current research and the actual clinical needs for precise diagnosis. There is an urgent need to develop more systematic optimization methods.

Research on survival time prediction has been explored by scholars both domestically and internationally through various technical approaches. Domestic studies mainly focus on risk assessment using classical statistical models and emerging neural networks. References [10] and [11] construct prognostic nomograms based on the SEER database, which can intuitively display key influencing factors and predict the survival time for specific patients. However, such models have limited ability to handle high-dimensional features. Reference [12] attempts to analyze pathological whole-slide images using Graph Neural Networks (GNNs) to assess patient survival risk; nevertheless, this method still faces significant shortcomings in managing class imbalance data and achieving collaborative optimization of feature selection and hyperparameter tuning. International research more often emphasizes algorithm comparison and multimodal data fusion. Reference [13] systematically compares nine machine learning algorithms and finds that Artificial Neural Networks (ANN) perform best in five-year survival prediction tasks. Reference [14] combines genomic data with Long Short-Term Memory (LSTM) models to further improve prediction accuracy, but such complex models often encounter dual constraints of poor interpretability and high computational resource consumption, limiting their clinical application. To address the common class imbalance issue in survival data, Reference [15] employs a combination of SMOTE oversampling techniques and bidirectional LSTM; meanwhile, Reference [16] constructs a model integrating

similarity networks and Convolutional Neural Networks (CNN) to analyze survival features. However, these studies often treat feature engineering and model hyperparameter optimization as separate processes, lacking an integrated end-to-end optimization scheme, which prevents achieving global optimality.

In summary, this study addresses the clinical needs of high-precision diagnosis and survival prediction for breast cancer by constructing a stacking ensemble diagnostic model that integrates multiple linear regression and EWOA, designing a CatBoost survival prediction model based on MSNOA and ADASYN, and ultimately developing a multi-platform-assisted diagnosis and treatment system that combines diagnostic and predictive functions using the Gradio and Flask frameworks. Through deep collaboration between ensemble learning and optimization algorithms, a full-chain closed-loop is achieved, encompassing feature optimization, model construction, and clinical decision support. The effectiveness and practicality of the models and system are validated through clinical data.

2. Research on Breast Cancer Diagnostic Models

This study addresses the issues of limited generalization performance in traditional breast cancer diagnostic models, homogeneity defects among base learners in ensemble learning, and low hyperparameter optimization efficiency. It comprehensively employs sequential feature selection, EWOA, and a double-layer heterogeneous Stacking ensemble technique to construct an intelligent breast cancer diagnostic model, achieving automated classification of benign and malignant breast tumors.

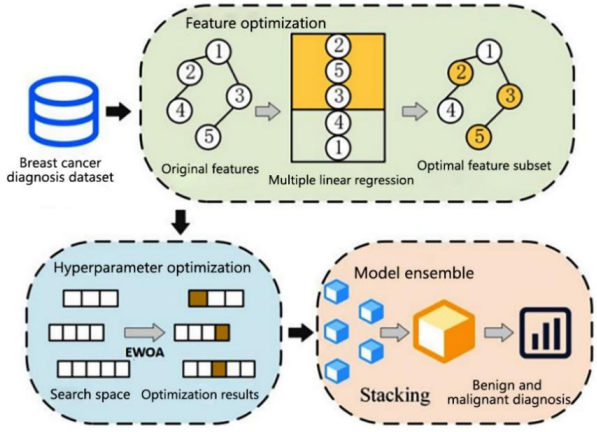

Fig 1. The overall model architecture is depicted

The overall model architecture is depicted in the figure. It takes the original clinical feature data, after standardization and missing value processing, as input, and passes through three core steps in sequence: serialized feature selection, EWOA hyperparameter global optimization, and double-layer heterogeneous Stacking ensemble model training. The process outputs tumor benign/malignant classification results and performance evaluation metrics, while also generating convergence curves for optimization and feature importance analyses, balancing model performance with interpretability.

In the feature engineering stage, a two-phase serialized feature selection strategy combining RFECV and SHAP values is designed: first, recursive feature elimination with cross-validation (RFECV) is used to perform preliminary feature screening, yielding a high-precision, low-redundancy feature ranking from the perspective of base learners; then, SHAP values are employed to quantify the overall importance and local contribution patterns of features, enabling the allocation of feature subsets to different base learners to enhance the diversity of prediction results and provide complementary information for ensemble learning. To address the issues of the basic Whale Optimization Algorithm (WOA) prone to falling into local optima and slow convergence during later iterations, an enhanced EWOA with multiple strategies is proposed, which modifies the convergence factor 'a' to a nonlinear adaptive decay and introduces dynamic inertia weights to balance exploration and exploitation capabilities; additionally, Lévy flight perturbation mechanisms are incorporated to help the algorithm escape local optima. The adaptive fitness function minimizes the classification error rate on the validation set of the ensemble model, enabling automatic global hyperparameter optimization for the meta-learner. A double-layer heterogeneous stacking ensemble model is constructed: the first layer comprises heterogeneous base learners—including support vector machine, random forest, XGBoost, k-nearest neighbors, and multilayer perceptron—each trained on different assigned feature subsets; the second layer uses logistic regression as the meta-learner, which takes the predicted probabilities generated via K-fold cross-validation from the first-layer base learners as meta-features. The hyperparameters are optimized using EWOA, and the final benign/malignant classification decision is made by integrating the predictions of the base learners.

3. Research on survival prediction models for breast cancer

Building upon the research on breast cancer diagnostic models, this section further focuses on precise prognosis assessment by developing a breast cancer survival prediction model. To address issues such as data class imbalance, algorithms prone to local optima, and the separation of feature and hyperparameter optimization in survival prediction tasks, a combined approach integrating ADASYN, MSNOA, and the CatBoost model is employed to construct the MSNOA-CatBoost binary classification survival prediction model. This model achieves high-precision prediction of patient survival status (Alive/Dead), providing scientific support for clinical prognosis evaluation and intervention strategy formulation. The model structure is as shown in the figure.

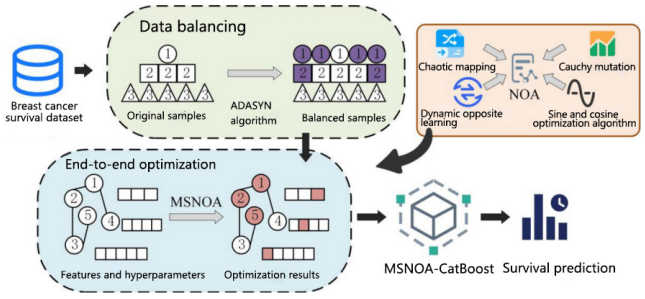

Fig 2. The model structure

To address the issue of class imbalance in survival data, this study employs the ADASYN algorithm to adaptively synthesize new samples based on the density distribution of minority class instances, balancing the number of positive and negative samples while preserving the distribution

characteristics of the minority class, thereby effectively enhancing the model's ability to recognize high-risk death cases. Due to the slow convergence and tendency to fall into local optima inherent in the traditional Sparrow Search Algorithm (SSA), improvements are introduced through Kent chaotic mapping, dynamic reverse learning, Cauchy mutation, and cosine-sine optimization algorithms, leading to the development of MSNOA. This approach leverages Kent chaotic mapping to enhance initial population diversity, incorporates the cosine-sine optimization mechanism during iterations to balance global exploration and local exploitation, and employs Cauchy mutation and dynamic reverse learning to boost the algorithm's ability to escape local optima. Based on MSNOA, an end-to-end collaborative optimization scheme for feature selection and hyperparameter tuning is constructed, breaking the conventional separation between the two. It automatically selects the optimal subset from 14 clinical features and simultaneously optimizes four core hyperparameters of the CatBoost model—learning rate, tree depth, L2 regularization coefficient, and number of iterations. Meanwhile, an adaptive fitness function, Fitness = 1 – (0.7 × Balanced_Acc + 0.3 × Dead_Recall), is designed to balance overall performance and minority class recognition, with the goal of minimizing this function to ensure balanced accuracy while emphasizing recall for deceased cases. This approach achieves deep integration of feature and parameter optimization, conducts global search, reduces feature redundancy, and fine-tunes model parameters for optimal performance.

4. Implementation of the Intelligent Auxiliary Diagnosis and Treatment System

This study designed and developed a B/S architecture-based intelligent auxiliary diagnosis and treatment prototype for breast cancer based on the Gradio and Flask frameworks. The system integrates the constructed EWOA-Stacking diagnostic model with the MSNOA-CatBoost survival prediction model through engineering integration, aiming to provide clinicians with a user-friendly,and accessible across multiple devices one-stop decision support platform. Its core functionalities encompass intelligent diagnosis and survival prognosis prediction for breast cancer.

The system adopts the architecture design of front-end and back-end separation to ensure the modularity, maintainability and scalability. The system's front-end integrates two core functional modules to provide users with an intuitive user experience. In the breast cancer intelligent diagnosis module, the constructed EWOA-Stacking diagnostic model is invoked and integrated. Users can input or upload patient clinical feature data that meets formatting requirements through the interface. The back-end system invokes this diagnostic model in real time for analysis and immediately return the benign/malignant classification results of tumors along with relevant confidence information on the front-end interface. The breast cancer survival prediction module integrates the MSNOA-CatBoost survival prediction model. Users can input prognostic clinical indicators related to patients, and the back-end system invokes this model for risk assessment, outputting the patient's five-year survival prediction results in a clear visual format on the front-end. The back-end service is the core engine of the system. First, the serialized files of the trained EWOA-Stacking and MSNOA-CatBoost models are saved and loaded into memory when the Flask service starts, ensuring the models are always available. Next, corresponding API routes (e.g., /api/diagnosis and /api/prognosis) are constructed in the Flask application to receive JSON-formatted data submitted from the front-end. When requests arrive, the back-end service executes the same data preprocessing workflow as during training, then invokes the corresponding models for inference, ultimately returning structured prediction results to the front-end interface for rendering and display.

5. System Function Verification

Comparative experiments were conducted on the Wisconsin Breast Cancer Original Classification Dataset (WBCD) and the Wisconsin Breast Cancer Diagnostic Breast Cancer Dataset (WDBC), with accuracy, precision, recall, and F1 score as evaluation metrics. The impact of different meta-learners on the diagnostic performance of benign and malignant breast tumors was explored, and the comparison results are shown in the figure.

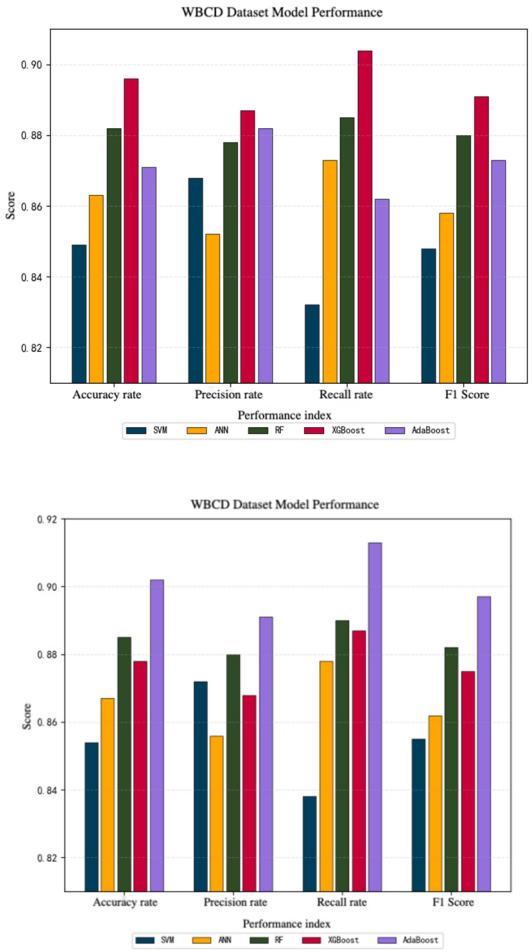

Fig 3. Performance Comparison of the WOA-Stacking Model with Different Meta-Learners

The figure demonstrates that the EWOA-Stacking model significantly improves breast tumor diagnositic performance by integrating outputs from multiple models. On the WBCD dataset, the model achieves optimal performance when using XGBoost as the meta-learner, while on the WDBC dataset, the AdaBoost-based version delivers superior results with all metrics exceeding 85%. Experimental results indicate that the EWOA-Stacking model exhibits high overall diagnostic accuracy and low false-negative rates, effectively identifying

malignant tumors and demonstrating outstanding performance in breast tumor classification.

Table 1 Performance of Different Models on the Breast Cancer Survival Test Set

Models	Accuracy	Precision	Recall	F1
SVM	0.8199	0.9324	0.8490	0.8887
ANN	0.8296	0.9146	0.8798	0.8969
Random Forest	0.8745	0.9278	0.9238	0.9258
XGBoost	0.8795	0.9221	0.9370	0.9295
AdaBoost	0.8534	0.9260	0.8988	0.9122
CatBoost	0.8807	0.9138	0.9487	0.9309
MSNOA-CatBoost	0.8708	0.9313	0.9150	0.9231

For breast cancer survival prediction research, female breast cancer patients from the Surveillance, Epidemiology, and End Results (SEER) database established by the National Cancer Institute (NCI) between 2008 and 2015 were selected as the study subjects. The fitness function was set as 0.7×balanced accuracy + 0.3×mortality recall rate. The co-optimized feature subset and hyperparameters were applied to the CatBoost model, and model training was completed on ADASYN-balanced data. On an independent test set, the model was validated against traditional models such as SVM and XGBoost using metrics including accuracy, balanced accuracy, and mortality recall rate, as shown in the table.

As shown in the table, the MSNOA-CatBoost model constructed in this study demonstrates a slightly lower overall accuracy than the basic CatBoost model, but achieves a precision rate of 0.9313, indicating improved robustness in prediction. It performs well in identifying death cases, a core clinical concern, thereby achieving a clinical balance between overall performance stability and precise identification of high-risk cases. Multi-model comparison experiments confirm that the MSNOA-CatBoost model effectively avoids overfitting risks through collaborative optimization of features and hyperparameters, enhancing the model's adaptability to feature perturbations. Addressing the clinical challenge of "focus on minority class recognition" in survival prediction, it significantly improves the identification capability of high-risk patients, outperforming traditional SVM and ANN models while achieving an optimal balance between model complexity and prediction accuracy. Additionally, the integration of MSNOA does not compromise training efficiency and reduces manual intervention costs through automated optimization, supporting the lightweight deployment of subsequent systems.

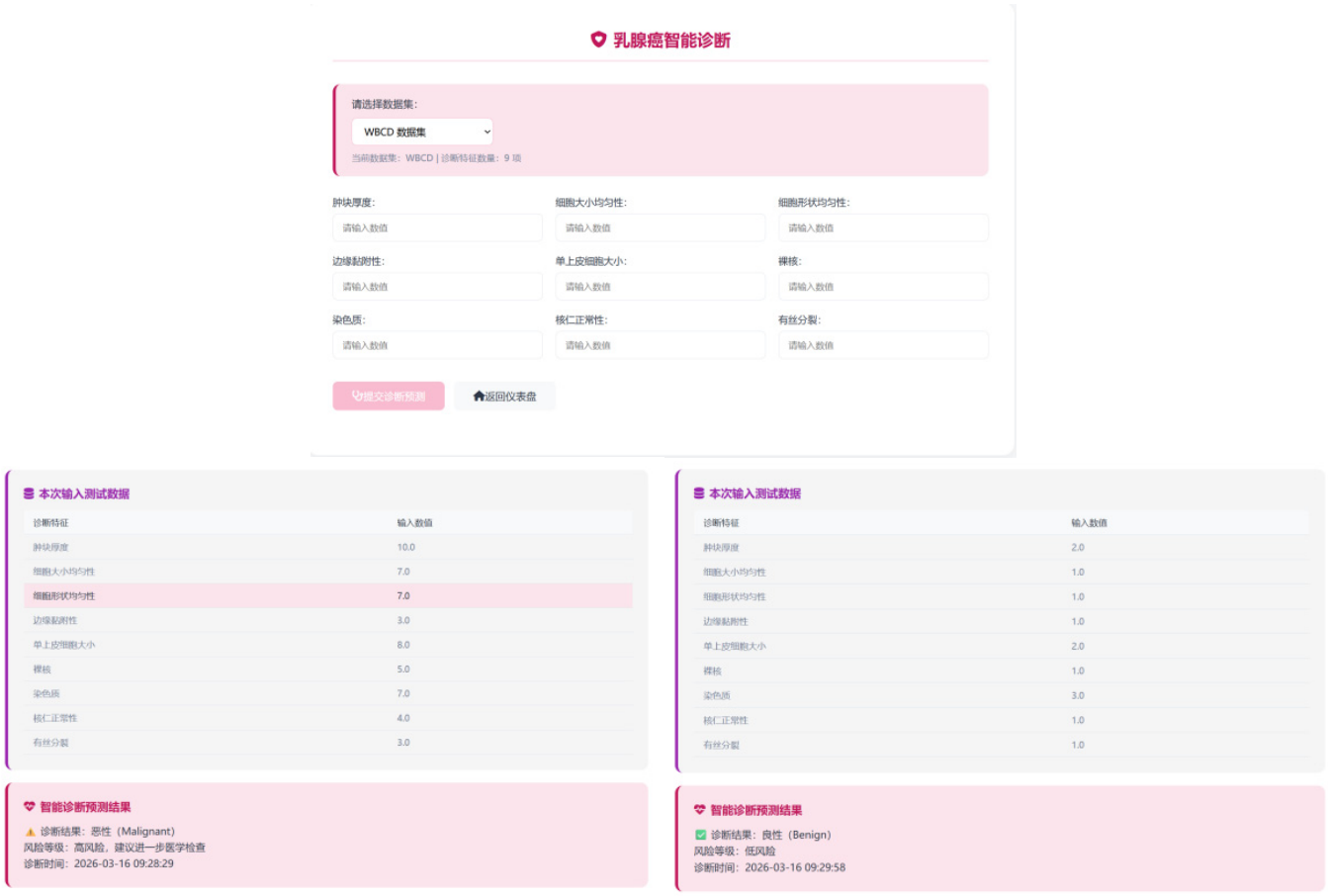

Fig 4. Demonstration and Test Results of the Breast Cancer Diagnosis and Prediction Function

The diagnostic interface of the system features a clear dataset adaptation mechanism and a standardized clinical feature entry process, effectively supporting standardized auxiliary decision-making in clinical diagnosis and treatment. As shown in the figure, a dropdown menu for dataset selection is located at the top of the interface, enabling clinicians to quickly switch to the corresponding dataset system for matched analysis according to the specific tumor subtypes (e.g., different subtypes or pathological grades) of patients. The default WBCD dataset includes 9 core cytological diagnostic features, namely clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, chromatin, normality of nucleoli, and mitoses, with standardized numerical input fields provided to facilitate accurate entry of patients' clinical examination data by clinicians.

After completing dataset selection and feature parameter input, users click the submit button, and the system

automatically invokes the pre-trained EWOA-Stacking intelligent diagnostic model for inference. The result display area adopts a visual card layout: on the one hand, it clearly presents the details of the input test data in a tabular form for clinicians to quickly verify the accuracy of the input information; on the other hand, it intuitively displays the intelligent diagnostic prediction results, including the diagnostic conclusion, risk grade assessment, and specific diagnostic timestamp. For malignant cases identified as high-risk, the interface simultaneously provides clinical prompts for “further medical examination,” realizing a closed loop of data input, inference prediction, and result visualization.

As shown in the figure, the overall layout of the breast cancer survival prediction module covers four core functional areas: model management, data input, parameter configuration, and result visualization. The interface design is closely aligned with the clinical prognosis assessment workflow, enabling automated prediction of 5-year survival status for breast cancer patients. At the top of the module, a model status control area is provided, while the data source section offers dual modes of "using simulated data" and "uploading real data," balancing the needs of clinical testing and practical application. The patient information entry area adopts a standardized form design, collecting key clinical indicators including age, race, marital status, income, residential region, AJCC stage, T/N/M stage, and surgical/radiotherapy/chemotherapy history. This forms a comprehensive feature set covering patient demographics, pathological staging, and treatment history, providing robust data support for survival prediction.The parameter configuration and prediction area allows users to set a follow-up period (default 60 months) and select the MSNOA-CatBoost prediction model. Clicking the "Start Prediction" button triggers the backend inference process. Upon completion, the system intuitively displays detailed prediction results via a pop-up window, including risk grade assessment and core features such as age, T stage, and N stage. The clear and prominent presentation enables clinicians to quickly grasp the patient's prognosis. Additionally, a 5-year survival rate curve visualization module is integrated on the right side of the interface. It dynamically displays the trend of patient survival probability over follow-up time using a line chart, intuitively presenting the model's risk stratification and prognosis assessment logic. Overall, the survival prediction module, through standardized feature entry, automated model inference, and multi-dimensional result visualization, fully covers the clinical decision-making chain of "patient information input – model inference – risk assessment – prognosis visualization."

Fig 5. Demonstration and Test Results of the Breast Cancer Survival Prediction Function

6. Conclusion

This study addresses the clinical needs of adjuvant diagnosis and treatment for breast cancer by constructing an intelligent diagnostic model. The model enhances the diversity of base learners through serialized feature selection, improves hyperparameter optimization efficiency by integrating the EWOA algorithm, and achieves performance enhancement via dual-layer heterogeneous Stacking integration technology. It demonstrates excellent accuracy, robustness, and generalization capability in distinguishing benign from malignant breast cancer. The breast cancer survival prediction model effectively balances the categorical distribution of survival data using the ADASYN algorithm, while achieving end-to-end collaborative optimization of features and hyperparameters based on the MSNOA algorithm. Additionally, it leverages the nonlinear fitting advantage of the CatBoost model to achieve high-precision prediction of patient survival status and enhances the identification of high-risk patients, providing reliable evidence for clinical prognosis assessment. The adjuvant diagnosis and treatment system for breast cancer, developed based on the Gradio and Flask frameworks, successfully achieves engineering encapsulation and integration of the two models, establishing a multi-platform compatible Web application. Functional and compatibility tests validate the system's accurate and stable operation throughout the entire workflow, confirming the feasibility of transforming algorithmic models into clinically applicable tools.

The model and system developed in this study provide technical support for precision medicine in breast cancer, and their engineering implementation lays a solid foundation for subsequent clinical functional validation, system iterative optimization, and practical clinical application.

References

[1] Wu J, Ma Y Y, Huang J, et al. Predictive value of radiomics based on DBT in triple-negative breast cancer[J]. Journal of Clinical Radiology, 2024, 43(8): 1292-1297.

[2] Deng Z, Su B H, Zhang K. Research on breast cancer classification based on ensemble learning[J]. China Medical Equipment, 2020, 35(12): 59-62.

[3] Wang D, Qu Y, Liu Y H, et al. Classification and diagnosis of breast cancer based on optimized random forest algorithm[J]. Computer Engineering and Design, 2022, 43(3): 706-712.

[4] Gao T, Yuan D C. Improved whale algorithm optimized support vector machine for breast cancer prediction[J]. Modern Electronics Technique, 2024, 47(11): 156-160.

[5] Du X X, Wang Z F, Wang B, et al. Black widow optimization algorithm based on chi-square transition strategy and its application[J]. Journal of Shaanxi University of Science & Technology, 2023, 41(6): 162-175.

[6] Maleki A, Raahemi M, Nasiri H. Breast cancer diagnosis from histopathology images using deep neural network and XGBoost[J]. Biomedical Signal Processing and Control, 2023, 86: 105152.

[7] Williamson S, Vijayakumar K, Kadam V J. Predicting breast cancer biopsy outcomes from BI-RADS findings using random forests with chi-square and MI features[J]. Multimedia Tools and Applications, 2022, 81(26): 36869-36889.

[8] Huang M W, Chen C W, Lin W C, et al. SVM and SVM ensembles in breast cancer prediction[J]. PLOS ONE, 2017, 12(1): e0161501.

[9] Abdar M, Makarenkov V. CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer[J]. Measurement, 2019, 146: 557-570.

[10] Wu Z S, Li F, Zhang H Q, et al. Prognostic factors and construction of survival prediction nomogram for breast cancer patients with brain metastasis: a single-center large sample retrospective study[J]. Journal of Medical Colleges of PLA, 2024, 45(5): 493-502.

[11] Wu G L, Liu J, Sun H. Construction of a nomogram for predicting survival probability of early elderly HER2-positive breast cancer based on SEER database[J]. Chinese Journal of Clinical Pharmacology and Therapeutics, 2024, 29(3): 283-295.

[12] Ye S J, Wang Y X. Cancer survival prediction method of WSI based on graph neural network[J]. Electro-Optic Engineering, 2024, 51(4): 66-78.

[13] Nguyen Q T N, Nguyen P A, Wang C J, et al. Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study[J]. Cancer Science, 2023, 114(10): 4063-4072.

[14] Mahmoud A, Alhussein M, Aurangzeb K, et al. Breast cancer survival prediction modelling based on genomic data: An improved prognosis-driven deep learning approach[J]. IEEE Access, 2024.

[15] Yuan H, Xu H. Deep multi-modal fusion network with gated unit for breast cancer survival prediction[J]. Computer Methods in Biomechanics and Biomedical Engineering, 2024, 27(7): 883-896.

[16] Zhang G, Ma C, Yan C, et al. MSFN: a multi-omics stacked fusion network for breast cancer survival prediction[J]. Frontiers in Genetics, 2024, 15: 1378809.