IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 August 2023, accepted 2 September 2023, date of publication 5 September 2023,
date of current version 11 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312305

RESEARCH ARTICLE

Auxiliary Diagnosis of Breast Cancer Based on Machine Learning and Hybrid Strategy

HUA CHEN[1], KEHUI MEI[1], YUAN ZHOU[1,2], NAN WANG[1], AND GUANGXING CAI[1]
[1]School of Science, Hubei University of Technology, Wuhan 430068, China
[2]School of Computer Science and Technology, Wuhan University of Bioengineering, Wuhan 430415, China

Corresponding author: Kehui Mei (792768156@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502156, in part by the Teaching and Research Project of Hubei Provincial Department of Education under Grant 282, and in part by the Doctoral Startup Fund of the Hubei University of Technology under Grant BSQD13051.

ABSTRACT Breast cancer has replaced lung cancer as the number one cancer among women worldwide. In this paper, we take breast cancer as the research object, and pioneer a hybrid strategy to process the data, and combine the machine learning method to build a more accurate and efficient breast cancer auxiliary diagnosis model. First, the combined sampling method SMOTE-ENN is used to solve the problem of sample imbalance, and the data are standardized to make the data have better separability. Then, the features of the dataset are initially screened using the mutual information method, and further secondary feature selection is performed using the recursive feature elimination method based on the XGBoost algorithm. Thus, the feature dimensionality of the dataset is reduced and the generalization ability of the model is improved. Finally, five different machine learning models are used for classification prediction, the best combination of parameters for each model is found using a grid search method, and the final results of each model are derived using a 10-fold cross-validation method. The experiments are conducted using the Wisconsin Diagnostic Breast Cancer dataset (WDBC), and the results of the study find that after the data are processed by the hybrid strategy, the best prediction results are obtained using the RF model with 99.52% accuracy, which is better than the previous research methods.

INDEX TERMS Breast cancer, machine learning, sample balancing, feature selection, classification forecast.

I. INTRODUCTION

According to the *Cancer Statistics, 2023* statistical estimates, breast cancer, lung cancer, and CRC account for 52% of all new diagnoses, with breast cancer alone accounting for 31% of female cancers [1]. Breast cancer, as one of the common malignant tumors in women, has become a focus of public health attention around the world. Its early diagnosis is important for the success of treatment and patient survival [2]. With the rapid development of machine learning and other technologies, more and more research has been devoted to applying these advanced technologies to the diagnosis and assisted decision making of breast cancer [3].

Machine learning, as an important artificial intelligence technology, has the ability to extract features, discover patterns and build predictive models from a large amount of medical data. It can not only assist doctors in identifying high-risk groups in early screening, but also be used for accurate diagnosis and personalized treatment plan development [4]. For breast cancer diagnosis, the application of machine learning has revolutionized the field and achieved remarkable results. By recording, analyzing and summarizing the data of a large number of breast cancer patients, machine learning algorithms are able to discover the regularity information that exists in them, thus effectively assisting the diagnostic process of breast cancer. This application not only significantly reduces the time and cost required for diagnosis, but also reduces the risk of misdiagnosis and underdiagnosis caused by subjectivity [5].

The problem of breast cancer diagnosis is widely recognized as a challenge in classifying benign and malignant tumors. Typically, machine learning models are used in

The associate editor coordinating the review of this manuscript and approving it for publication was Akin Tascikaraoglu.

combination for breast cancer diagnosis, and the classification performance of the final model is superior or inferior depending on a number of factors, which include the preprocessing of data, the selection of input features, and the setting of model parameters. Therefore, finding an effective strategy to obtain better prediction results remains a difficult and important task. Keyvanpour et al. [6] designed a new method for mass detection and classification based on weighted association rule mining (WARM), which aims to improve the accuracy of detecting and classifying masses in mammographic images and classifying them as benign and malignant. Jha et al. [7] proposed an effective method for conversion of experimental attributes of breast cancer dataset. The method uses latent semantic analysis techniques to convert raw attributes into more informative features and integrate them with classification methods to improve the accuracy of breast cancer identification. Rekha and Amali [8] used k-means clustering algorithm for data preprocessing to divide the samples into different clusters, thus reducing the dimensionality and noise of the data. Then the LFCSO-PSVM algorithm was proposed, which combined local feature selection optimization with parallel support vector machines to further improve the accuracy of breast cancer classification. Sowan et al. [9] proposed an ensemble filter feature selection method that evaluates and ranks features based on their relevance and importance. Then a wraparound feature selection algorithm was proposed in combination with the association classification method, which takes the most relevant features as input to further optimize the classification prediction of breast cancer. Alsubai et al. [10] proposed a genetic-hyperparameter optimization (G-HPO) method for determining the optimal hyperparameter values for the classifier under consideration. This method enabled to find the optimal hyperparameter configuration of the model more precisely, thus improving the accuracy and performance of the classifier. Ramakrishna et al. [4] used Adaboost ensemble technique for classification prediction of breast cancer and used SMOTE oversampling method to deal with the sample imbalance problem. The experimental results showed that the performance of the model was significantly improved when the preprocessed data was used for classification prediction using Adaboost integration technique.

The application of machine learning in the field of breast cancer diagnosis has brought great potential and opportunities to the field. Many scholars have utilized machine learning algorithms to accelerate the diagnostic process of breast cancer, improve the accuracy of models, and reduce the risk of misdiagnosis and underdiagnosis due to subjectivity. Abdar et al. [11] proposed a nested integration approach to detect benign and malignant breast tumors by using stacking and voting as a classifier combination technique. Experimental results on the WDBC dataset showed that the proposed two-layer nested integration model outperformed a single classifier with an accuracy of 98.07%. Vijayalakshmi et al. [12] proposed a multimodal classification model (POS-NDS) for breast cancer prediction, which filters the most significant features associated with breast cancer prediction by using particle swarm optimization and non-dominated ranking of classifier models. Experimental results showed that the proposed model reduced the prediction error with an accuracy of 98.8%. Kumar et al. [13] proposed a new genetic programming fitness function for medical data classification to solve the problem of data imbalance. Experimental results showed that the proposed algorithm can better meet the classification task of medical data with an accuracy of 99.12% on the WDBC dataset. El Rahman et al. [14] used genetic algorithm (GA) for feature selection and compared different classifiers such as decision tree, random forest, logistic regression, K-nearest neighbor and support vector machine. The experimental results showed that the accuracy of C-SVM classification method based on kernel function RBF was 99.04% on the WDBC dataset, which was better than other classification methods. Stephan et al. [15] proposed a hybrid algorithm combining standard ABC and WOA, and adopted HAW algorithm for feature selection and parameter optimization of the neural network model. Experimental results on WDBC dataset showed that the accuracy of HAW-RP was 98.5%. Naseem et al. [16] proposed a classifier integration-based breast cancer diagnosis system and automatic prognostic detection system. Experimental results on the WDBC dataset showed that the integrated method outperformed other single methods, with an accuracy of 98.83%. Badr et al. [17] used the Gray Wolf Optimizer (GWO) to improve the performance of Support Vector Machines (SVMs) and proposed three effective scaling techniques for the classical normalization technique. The experimental results of WDBC showed that the accuracy of the proposed hybrid GWO-SVM model after normalized scaling was 98.60%. In addition, the proposed scaling technique and the proposed GWO-SVM model could converge quickly with an accuracy of 99.30% using the proposed scaling technique. Alshayeji et al. [18] proposed an artificial neural network model (ANN) that can diagnose breast cancer without applying feature optimization or selection algorithms. Experimental results on the WDBC dataset showed that the proposed model can be used to assist in breast cancer diagnosis with an accuracy of 99.47%. Singh et al. [19] proposed a unique feature selection method based on the Eagle Strategy Optimization (ESO), Gravitational Search Optimization (GSO) algorithm and their hybrid algorithms, which could select the least number of features to achieve maximum accuracy. Experimental results showed that the method achieved good results on the WDBC dataset with an accuracy of 98.96%. Mushtaq et al. [20] used extended kernel principal component analysis (K-PCA) to reduce the dimensional space of the data and conducted experiments with five different kernels in K-PCA. The experimental results showed that K-PCA using sigmoid kernels obtained the best results with 99.28% accuracy on the original Wisconsin breast cancer dataset (WBC). Mahesh et al. [21] proposed an improved method for

breast cancer detection. First, to address the imbalance of the data, an oversampling technique (SMOTE) was used, and then, the data were classified using plain Bayes, decision trees, random forests and their integrated models. Experimental analysis on the Kaggle Wisconsin breast cancer dataset showed that the XGBoost-Random Forest integrated model had an accuracy of 98.20% in early detection of breast cancer. Wu and Hicks [22] used support vector machine, KNN, plain Bayesian and decision tree models to distinguish triple-negative breast cancer from non-triple-negative breast cancer. Experimental results on RNA-Seq data found that the support vector machine model was able to more accurately classify breast cancers into triple negative and non-triple negative.

In summary, the development of machine learning has provided great help to the diagnosis of breast cancer. In order to improve the accuracy of existing breast cancer identification methods, this paper takes breast cancer as the research object, creatively designs a hybrid strategy to process the data, and combines machine learning models and their related knowledge to construct a more accurate and efficient breast cancer diagnosis model. The idea of the research in this paper is shown in Figure 1.

The main works of this paper are summarized as follows.

1) We propose a hybrid strategy for processing data that facilitates the training of machine learning models, and thus improves their predictive performance.

2) We use the combined sampling method SMOTE-ENN to solve the problem of sample imbalance, and to standardize the data to improve the separability of the data.

3) We use mutual information method to do the initial screening of the dataset and further use the recursive feature elimination method based on XGBoost algorithm for secondary feature selection, which improves the information quality of the selected features and achieves the maximum accuracy with the least number of features.

4) We use the grid search method to find the best combination of parameters for each model, and use the cross-validation method to derive the predictions for each model.

5) We conduct a comprehensive comparative analysis of the prediction results of various machine learning models, and experimentally prove that our research method is superior to the previous research methods.

The rest of the paper is organized as follows: in Section II, we introduce the theoretical knowledge related to machine learning. In Section III, we present information about the dataset in detail and perform sample equalization and standardization on the data. In Section IV, we use the mutual information method and recursive feature elimination method for secondary feature selection to derive the best feature subset. In Section V, we do three sets of comparative analysis experiments to verify that the research method in this paper is effective and feasible. Finally, we present the research conclusions of this work and discuss future research work in Section VI.

II. RELATED THEORETICAL KNOWLEDGE

A. EXTREME GRADIENT BOOSTING (XGBOOST)

XGBoost is an efficient gradient boosting decision tree algorithm [23]. The core idea of the algorithm is to integrate by iteratively generating many CART classification regression trees, and the new tree generated in each iteration is based on the training and prediction of the tree generated in the previous iteration. That is, the optimization is performed in the direction of the negative gradient of the loss function, and a new tree generated in each iteration corresponds to the model learning a new function to fit the prediction bias of the tree generated in the previous iteration, and iterating until the bias cannot be reduced to improve the performance of the model [24].

The objective function of the sample predicted values consists of two components, the error of the model, i.e., the loss function, and the structural error of the model, i.e., the canonical term, and is optimized using a second-order Taylor expansion with the following objective function:

$$Obj(f_i) = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \tag{1}$$

where n is the total number of samples, $\hat{y}_i^{t-1}$ is the predicted value of sample i by the t-1th learner, $f_i(x_i)$ is the newly added t-th learner, $\Omega(f_t)$ is a regular term that can be pruned to prevent model overfitting, and L is a loss function that is used to calculate the deviation between the true and predicted values of the samples [25].

B. RANDOM FOREST (RF)

RF is an integrated prediction model built on top of decision trees and incorporates random sampling [26]. It generates a series of decision tree models with differences by constructing different training datasets and different feature spaces, and each decision tree model then comes up with a classification result based on its own judgment, after which the final results are then aggregated according to the voting results of each classifier by integrating the classification results of all the trees and conducting voting. Thus, the random forest model does not easily fall into overfitting and has a good resistance to noise [27].

The effectiveness of random forest classification is related to two factors. One is the correlation of any two trees in the forest, the greater the correlation, the greater the error rate, so if the independence between trees is guaranteed the accuracy of the model will be improved. The second is the classification ability of each tree in the forest. If the classification ability of each tree is very strong, the final accuracy of the whole forest will be higher, but if the classification ability of each tree is very weak, even if there are more classifiers, it will not achieve better results. Therefore, in order to improve the final accuracy of the model it is necessary to select the metrics that are as relevant as possible to ensure that the model has better prediction results [28].

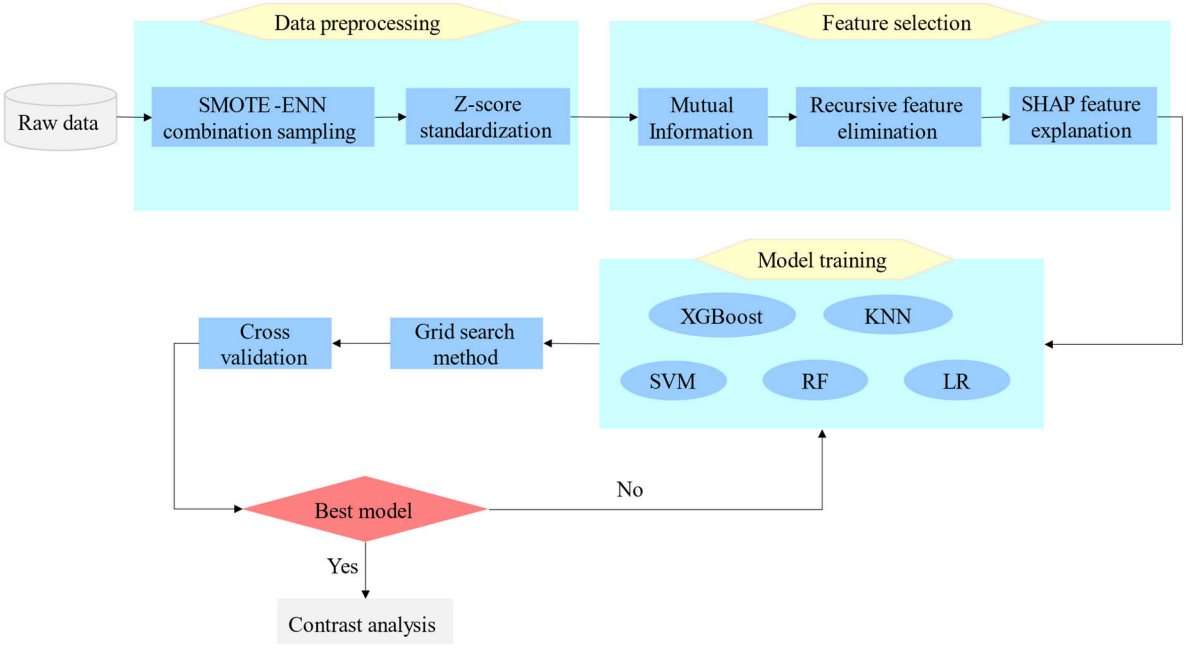


FIGURE 1. Flowchart of breast cancer auxiliary diagnosis based on machine learning and hybrid strategy.

C. SUPPORT VECTOR MACHINE (SVM)

SVM is a binary classification model, which divides samples by finding a hyperplane. The principle of segmentation is to maximize the interval, which is finally transformed into a convex quadratic programming problem for solving [29]. With the continuous development of SVM theory, SVM can transform low-dimensional linear indivisibility into high-dimensional linear divisibility by constructing kernel functions, thus solving the problem of linear indivisibility in low-dimensional space [30].

In practical application, it is very important to select a suitable kernel function for algorithm implementation. Commonly used kernel functions mainly include linear kernel function, polynomial kernel function, Gaussian kernel function and Sigmoid kernel function [31].

1) LINEAR KERNEL FUNCTION

$$K(x_i, x_j) = x_i \cdot x_j \tag{2}$$

Linear kernel function is the most common kernel function, it is directly used as the inner product of the original features, mainly used in the case of linear divisible, with fewer parameters, fast advantages. However, many data in practical applications are not linearly separable, so we need to choose other kernel functions that are more suitable.

2) POLYNOMIAL KERNEL FUNCTION

$$K(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d, \quad \gamma > 0,\ r \geq 0 \tag{3}$$

Polynomial kernel function can use order d to adjust the performance of the algorithm, so that it can be applied to more data, but when d is large, the value of the kernel matrix will tend to infinity or 0, and it is difficult to find the best value, so polynomial kernel function has higher requirements on order d.

TABLE 1. Confusion matrix.

Label		Predicted result	
		Positive	Negative
True situation	Positive	TP	FN
	Negative	FP	TN

3) GAUSSIAN KERNEL FUNCTION

$$K(x_i, x_j) = \exp\left\{-\frac{(x_i - x_j)^2}{2\sigma^2}\right\}, \quad \sigma > 0 \tag{4}$$

Gaussian kernel function has no requirement on sample size and dimension. It is the most widely used kernel function. Since the Gaussian kernel function has only one parameter σ, compared with other kernel functions, it is easier to select the value of the parameter, which greatly reduces the calculation amount of kernel function. However, when the value of parameter σ is too small, it is easy to cause overfitting. Therefore, it is necessary to properly determine the value of parameter σ.

4) SIGMOID KERNEL FUNCTION

$$K(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + r), \quad \gamma > 0,\ r \geq 0 \tag{5}$$

Sigmoid kernel function is a nonlinear function of neurons, which is widely used in deep learning and machine learning. When Sigmoid kernel function is adopted, support vector machine is a kind of multi-layer perceptron neural network.

TABLE 2. Features of the dataset.

No.	Feature	No.	Feature	No.	Feature
1	radius_mean	11	radius_se	21	radius_worst
2	texture_mean	12	texture_se	22	texture_worst
3	perimeter_mean	13	perimeter_se	23	perimeter_worst
4	area_mean	14	area_se	24	area_worst
5	smoothness_mean	15	smoothness_se	25	smoothness_worst
6	compactness_mean	16	compactness_se	26	compactness_worst
7	concavity_mean	17	concavity_se	27	concavity_worst
8	concave points_mean	18	concave points_se	28	concave points_worst
9	symmetry_mean	19	symmetry_se	29	symmetry_worst
10	fractal_dimension_mean	20	fractal_dimension_se	30	fractal_dimension_worst

D. K-NEAREST NEIGHBOR (KNN)

KNN is a supervised learning algorithm that predicts unknown category samples by looking for the nearest K known category samples [32]. The basic idea of KNN is as follows: In order to judge the categories of data to be classified, the distance between the data to be classified and the sample data of known categories is calculated by taking all the sample data of known categories as references. Then select K known samples that are closest to the data to be classified. Finally, according to the voting rule of minority obedience to majority, the data to be classified and the K nearest neighbor samples with the largest proportion of categories are grouped into one category [33].

There are two basic elements in the construction of KNN model, which are the selection of K value and the measurement of distance. Different K values will have a great impact on the accuracy of model prediction. A small K value will easily lead to overfitting of the model, while a large K value will easily lead to underfitting of the model. Choosing an appropriate distance measure is also crucial to the classification accuracy of the model. Commonly used distance measures include Euclidean distance, Manhattan distance, cosine distance, etc. [34].

E. LOGISTIC REGRESSION (LR)

LR is a classification algorithm in machine learning, which is widely used in practice because of its good interpretability and high generalization ability [35]. The idea of logistic regression algorithm is derived from linear regression, and in order to solve the quantitative sensitivity problem of linear regression, logistic regression nests a logistic function on the basis of linear regression [36]. Its core idea is that if the output result of linear regression is a continuous value and the range of values is unbounded, the output result can be mapped to the interval [0, 1] by a sigmoid function. Its function equation is:

$$g(z) = \frac{1}{1 + e^{z}} \tag{6}$$

where $z = w^T \cdot x$, w is the weight to be learned, and x is the sample feature vector. $g(z)$ then denotes the predicted probability value corresponding to the occurrence of the event corresponding to it inferred from the sample.

TABLE 3. Methods of sample balancing.

Sampling methods	Benign	Malignant
Original sample	357	212
Random up-sampling	357	357
Random down-sampling	212	212
SMOTE up-sampling	357	357
SMOTE-ENN combination sampling	311	310

F. EVALUATION INDICATORS OF THE MODEL

In this paper, confusion matrix is used to evaluate the effect of classification model. For each sample, it is divided into positive or negative, and there are four combined results of the real and predicted categories: 1) If a sample is positive and predicted to be positive, it is True Positive (TP). 2) If a sample is positive but predicted negative, it is False Negative (FN). 3) If a sample is negative but predicted to be positive, it is False Positive (FP). 4) If a sample is negative and predicted to be negative, it is True Negative (TN). These four types of results can form the confusion matrix [37].

Based on confusion matrix, evaluation indexes of classification models such as accuracy, precision, recall, and F1-score can be obtained, with specific meanings as follows [38].

(1) Accuracy: represents the percentage of the number of samples predicted correctly in the total. The specific formula is:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{7}$$

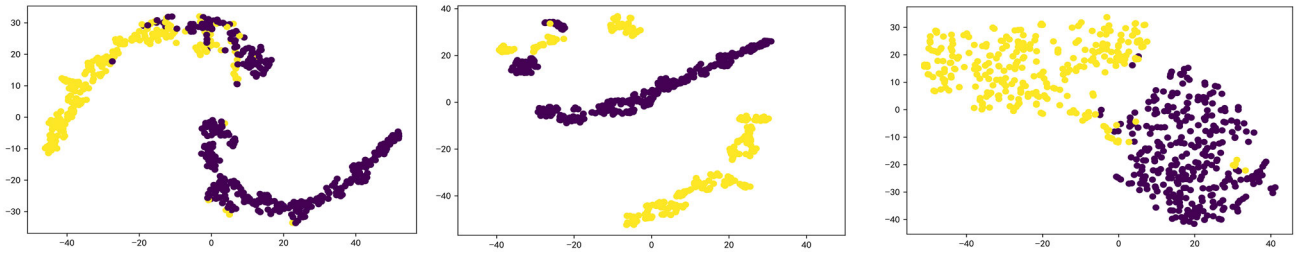
(2) Precision: it indicates the percentage of results that are truly positive samples among those predicted by the model to be positive samples. It is an evaluation indicator for the prediction results. The specific formula is:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

(3) Recall: indicates the percentage of samples that are predicted to be positive out of those that are actually positive. It is an evaluation indicator for the original sample. The specific formula is:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

(4) F1-score: it takes into account both the accuracy and recall of the model, and is a weighted average of the accuracy



(a) Raw data (b) Balanced data (c)Standardized data

FIGURE 2. Data distribution.

TABLE 4. Features obtained by initial screening.

No.	Feature	Score
1	radius_worst	0.6186
2	area_worst	0.6158
3	perimeter_worst	0.6084
4	perimeter_mean	0.5214
5	concave points_worst	0.5188
6	area_mean	0.5043
7	radius_mean	0.5009
8	concave points_mean	0.4936
9	area_se	0.4324
10	concavity_mean	0.4299
11	concavity_worst	0.4027
12	perimeter_se	0.3199
13	radius_se	0.3181
14	compactness_mean	0.2952
15	compactness_worst	0.2930
16	concave points_se	0.1792
17	texture_worst	0.1659
18	concavity_se	0.1561
19	texture_mean	0.1542
20	compactness_se	0.1223

TABLE 5. Features obtained by secondary screening.

No.	Feature	Score
1	perimeter_worst	81.4536
2	area_worst	15.5894
3	concave points_worst	3.3451
4	concave points_mean	2.0234
5	radius_worst	1.8446
6	compactness_worst	1.6175
7	compactness_mean	1.2462
8	radius_mean	0.9411
9	texture_mean	0.8703
10	texture_worst	0.8354
11	concavity_worst	0.5385
12	concave points_se	0.4062
13	concavity_mean	0.1584

and recall of the model. The specific formula is:

$$\mathrm{F1} = \frac{TP}{2TP + FP + FN} \tag{10}$$

The values of the above metrics are all between 0 and 1, and the larger the value, the better the classification effect of the model.

III. DATA SOURCE AND PREPROCESSING

A. DATA SOURCES AND PRESENTATIONS

The data used in this paper are from the Diagnostic Breast Cancer Dataset (WDBC) provided by the Wisconsin Center for Clinical Sciences, which has a total of 569 experimental samples containing 357 benign cases and 212 malignant cases with 32 features. Of these, ID is the patient number, Diagnosis is the sample label, and the remaining 30 features are data calculated from digitized images of fine needle aspiration (FNA) of breast masses, and these feature values describe the morphological characteristics of the cell nuclei in the images, as shown in Table 2. The mean, standard deviation, and maximum value (the average of the three largest feature values in the live sample images) of these features are calculated for each sample image. Features 1-10 are the mean values of the nucleus features in the sample images, which reflect the overall morphological characteristics of the sample nuclei. Features 11-20 are the standard deviations of the nuclei feature values in the sample images, which reflect the fluctuations of the nuclei in each feature value in a sample image. Features 21-30 are the maximum values of the nuclei in the sample images, which is not the maximum value of the whole sample, but the average of the top three values, so as to reduce the impact of errors in the measurement process.

The dataset demonstrates the correlation between the diagnostic results of benign and malignant cases and the size of the cell nuclei feature values. By establishing the model, it can assist physicians in diagnosing the tumor status of patients, improve the efficiency of diagnosis, reduce missed diagnosis and misdiagnosis, and thus improve the survival and cure rate of breast cancer patients.

B. SAMPLE BALANCING

Since there are more benign cases and fewer malignant cases in the data set, there is uneven distribution of categories, and this problem of sample imbalance will affect the training effect of the model. The model may learn such a priori information about the proportion of samples in the training set, resulting in an emphasis on benign cases in the actual prediction [39]. Methods such as random up-sampling, random down-sampling, SMOTE up-sampling, SMOTE-ENN combination sampling, etc. are usually used to solve the problem of sample imbalance, which can reduce the prior information of model study sample proportion, so as to obtain the model that can learn to distinguish the essential characteristics of benign and malignant cases.

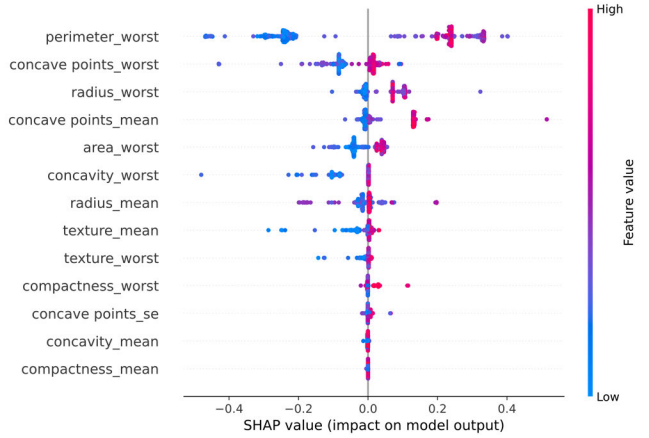


FIGURE 3. SHAP feature interpretation diagram.

As can be seen from Table 3, both benign and malignant cases obtained by the random up-sampling method are 357 cases, which is to achieve sample equalization by synthesizing new malignant cases, but it tends to lead to model overfitting and makes the generalization ability of the model decrease. The number of both benign and malignant cases obtained by the random down-sampling method is 212, which achieves sample equalization by randomly removing some benign cases, but this will lose some important information in the benign cases. The benign and malignant cases obtained by using SMOTE up-sampling method are both 357 cases, and it synthesizes new malignant cases by adding random noise to the malignant cases and according to certain rules, but it may produce the problem of distribution marginalization and increase the difficulty of classification. The combined sampling method SMOTE-ENN is used to obtain 311 benign cases and 310 malignant cases, which is a combination of up-sampling and down-sampling, and generates noisy samples by inserting new points between the marginal anomalies and inner points of malignant cases, and then cleaning the whole sample, which can better solve the problem of data imbalance [40]. Through comparative analysis, this paper finally chose to use the combined sampling method SMOTE-ENN to solve the problem of sample imbalance.

Specific steps of the SMOTE-ENN algorithm [41]: 1) For the unbalanced dataset, it is divided into a minority class $S_{\min}$ and a majority class S_{maj}. 2) For each minority class sample, compute its K nearest neighbors. 3) The number N of new samples that need to be synthesized for each minority class sample is determined based on the imbalance ratio of the dataset. 4) For each minority class sample, N nearest neighbors are randomly selected from its K nearest neighbors. If the nearest neighbors are selected in the process of synthesizing a new sample x_n from sample x, the new sample is constructed according to Equation $x_{new} = x + rand(0, 1) \cdot (x_n - x)$.

C. STANDARDIZED PROCESSING OF DATA

Since each sample data may have different orders of magnitude, if the analysis is performed directly with the raw feature values without any processing, it will highlight the role of features with higher values in the comprehensive analysis. In order to ensure the reliability of the results, the raw data need to be processed. Commonly used processing methods are standardization and normalization. Normalization is to map the data to a specified range to facilitate features of different magnitudes and orders of magnitude to be able to be compared and weighted. Standardization is based on the columns of the feature matrix, and the normalized data retains useful information in the outliers. In contrast, the data used in this paper has a large number of outliers, so it is not suitable for normalization and is suitable for standardization [42].

In this paper, according to the characteristics of the WDBC dataset, the Z-score standardization method is selected to process the dataset, and the processed data conforms to the standard normal distribution with mean 0 and standard deviation 1. The transformation function is:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{11}$$

where μ is the mean of the sample data and σ is the standard deviation of the sample data.

After the data are standardized, the distribution of the data is observed through t-SNE visualization. t-SNE is a commonly used dimensionality reduction algorithm for reducing high-dimensional data to 2 or 3 dimensions, which is mainly used for exploratory data analysis and visualization of high-dimensional data [43]. As can be seen from Figure 2, the data are standardized to have better divisibility.

IV. FEATURE SELECTION

After data preprocessing, feature selection is also needed to eliminate features that are irrelevant or redundant to the problem and reduce the feature dimension of the data set, thus reducing the complexity of the model and improving its generalization ability. In this paper, the best feature subset is screened by a secondary feature selection method. First, the top 20 features with scores are screened by the mutual information method, and then 13 features are screened by the recursive feature elimination method based on the XGBoost algorithm to obtain the final feature subset.

A. INITIAL SCREENING BY MUTUAL INFORMATION METHOD

The mutual information method captures not only the linear relationship between each feature and the label, but also the nonlinear relationship. It takes values between [0, 1], and the larger the value the stronger the correlation, with 0 indicating that the two variables are independent of each other and 1 indicating that the two variables are completely correlated [44]. The mutual information of variable X and variable Y can be defined as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right) \tag{12}$$

where $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively, and $P(x, y)$ is the

TABLE 6. Optimal parameters of the model.

Model	Parameter	Meaning	Set value
XGBoost	n_estimators	Number of decision tree	60
	max_depth	Depth of tree	3
	learning_rate	Learning rate	0.09
RF	n_estimators	Number of decision tree	70
	max_depth	Depth of tree	6
SVM	C	Penalty coefficient	0.9
	kernel	Type of kernel function	rbf
KNN	n_neighbors	k value	1
	weights	Weight of the nearest neighbor sample	uniform
	algorithm	Algorithm	ball_tree
LR	penalty	Penalty item	L2
	solver	Optimization algorithm	liblinear
	C	Inverse of regularized intensity	0.6

TABLE 7. Model prediction results for the control and experimental groups.

	Model	Accuracy	Precision	Recall	F1-score
Control group	XGBoost	**97.89%**	**98.11%**	**96.21%**	**97.12%**
	RF	96.31%	96.24%	92.94%	95.64%
	SVM	91.39%	95.78%	80.78%	87.35%
	KNN	92.98%	92.43%	88.77%	90.37%
	LR	94.38%	93.86%	91.13%	92.28%
Experimental group	XGBoost	99.20%	99.06%	99.35%	99.20%
	RF	**99.52%**	**99.08%**	**99.68%**	**99.21%**
	SVM	98.23%	98.09%	98.39%	98.23%
	KNN	98.07%	97.54%	98.71%	98.09%
	LR	98.71%	98.74%	98.71%	98.71%

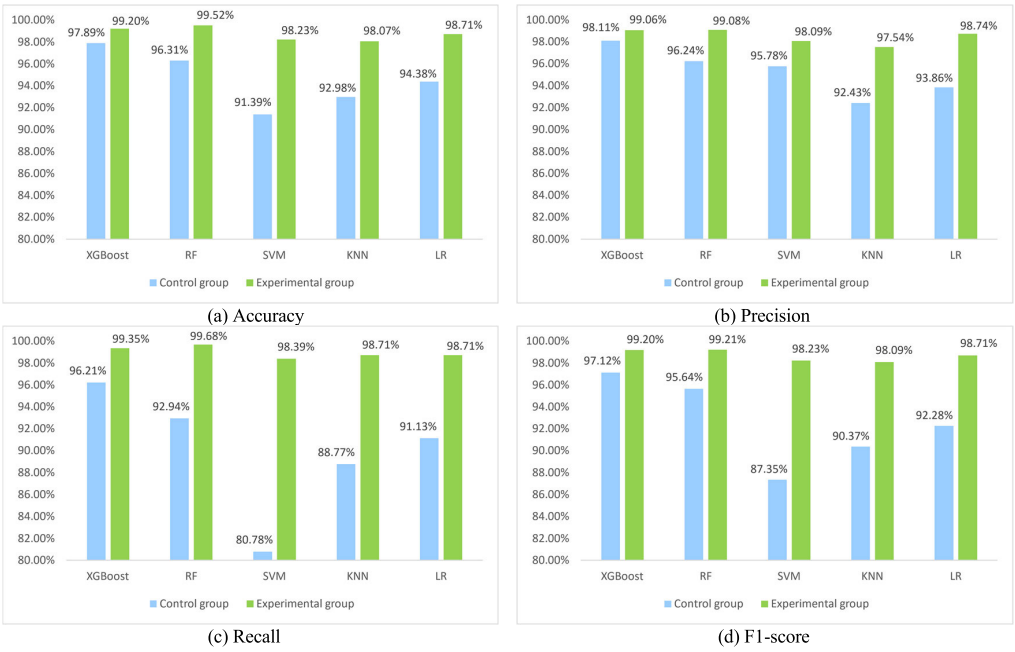


FIGURE 4. Comparison of the results of the control and experimental groups.

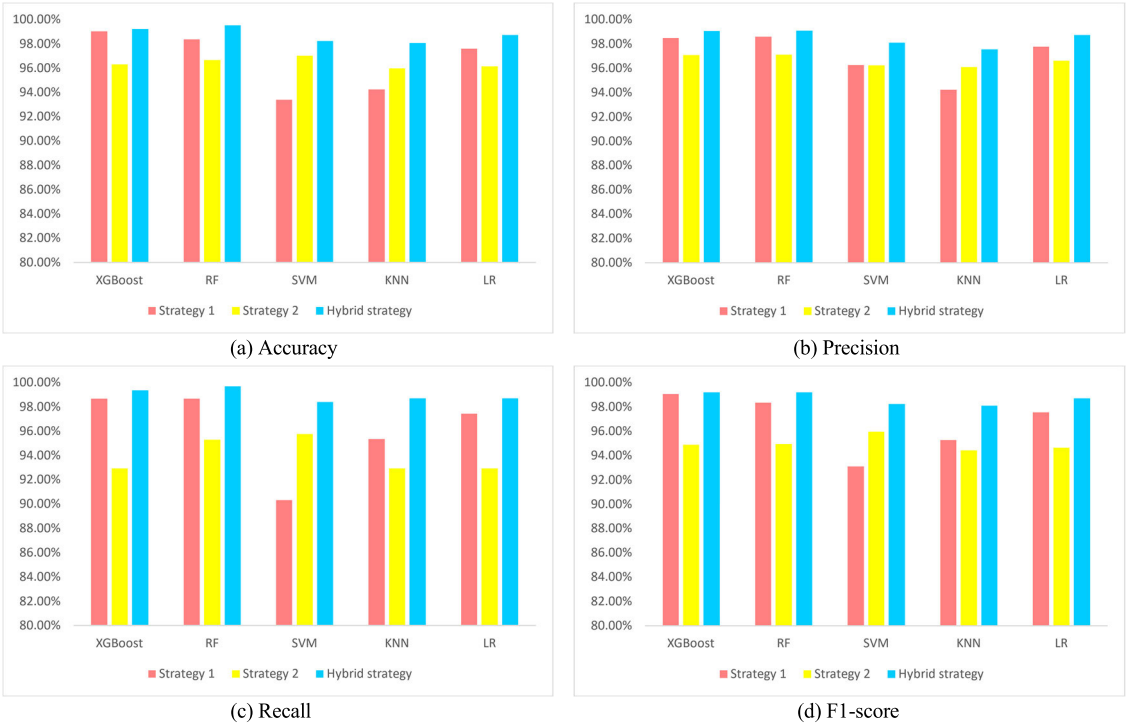
joint probability distribution function of X and Y. The top 20 features of the scores screened by the mutual information method are shown in Table 4.

B. SECOND SCREENING BY RECURSIVE FEATURE ELIMINATION

Recursive feature elimination method can reduce the feature dimension and select the optimal feature subset [45]. The specific steps are as follows: 1) Firstly, 20 features selected by the mutual information method are input into the XGBoost classifier as initial feature subsets. The importance of each feature is measured by the average information gain, and the classification accuracy of the initial feature subsets is obtained by cross validation method. 2) A feature with the lowest feature importance is removed from the current feature subset to obtain a new feature subset, which is input into the

TABLE 8. Comparison of prediction results of different strategies.

Strategy	Model	Accuracy	Precision	Recall	F1-score
Strategy 1	XGBoost	**99.03%**	98.48%	98.67%	**99.05%**
	RF	98.35%	**98.59%**	**98.68%**	98.36%
	SVM	93.40%	96.26%	90.32%	93.11%
	KNN	94.23%	94.23%	95.35%	95.26%
	LR	97.58%	97.76%	97.42%	97.56%
Strategy 2	XGBoost	96.31%	97.09%	92.92%	94.90%
	RF	96.66%	**97.11%**	95.30%	94.93%
	SVM	**97.01%**	96.23%	**95.76%**	**95.95%**
	KNN	95.96%	96.10%	92.92%	94.42%
	LR	96.13%	96.61%	92.94%	94.65%
Hybrid strategy	XGBoost	99.20%	99.06%	99.35%	99.20%
	RF	**99.52%**	**99.08%**	**99.68%**	**99.21%**
	SVM	98.23%	98.09%	98.39%	98.23%
	KNN	98.07%	97.54%	98.71%	98.09%
	LR	98.71%	98.74%	98.71%	98.71%



(a) Accuracy (b) Precision

(c) Recall (d) F1-score

FIGURE 5. Comparative analysis of the results of different strategies.

XGBoost classifier again to calculate the importance of each feature in the new feature subset, and the classification accuracy of the new feature subset is obtained by cross-validation method. 3) Repeat step 2 until no feature is removed. Finally, K different feature subsets are obtained, and the feature subset with the highest classification accuracy is selected as the optimal feature subset. The features screened by recursive feature elimination method are shown in Table 5.

C. SHAP BASED FEATURE INTERPRETATION

SHAP is a framework for interpreting model output. Different from the importance of features trained in machine learning models, SHAP interpretive models make comparison of all features on the basis of consistency, and also have good computational performance. The core idea is to calculate the marginal contribution of features to the model output and then interpret the machine learning model at both global and local levels. SHAP constructs an additive explanatory model that measures the impact of features on the outcome by calculating the contribution of each feature to the prediction outcome, which may be positive or negative, where a positive value improves the prediction outcome and a negative value reduces the prediction outcome [46]. In this paper, we use the SHAP interpretation method to analyze the impact of different features, and the results are shown in Figure 3.

TABLE 9. Training time of each model under different strategies.

Models	Time (seconds)		
	Strategy 1	Strategy 2	Hybrid strategy
XGBoost	0.162	4.248	4.722
RF	0.243	4.354	4.727
SVM	0.112	4.202	4.647
KNN	0.114	4.219	4.665
LR	0.113	4.187	4.651

In the figure, the importance of features decreases from top to bottom, and the color of scattered points from blue to red indicates the value of features from small to large, and each point represents the SHAP value of a sample.

As shown in Figure 3, among the 13 features screened, perimeter_worst is the most important feature that affects the prediction results of the model, and the larger the value of perimeter_worst, the higher the value of SHAP. This indicates that the greater the perimeter of the nucleus, the greater the risk of the patient being diagnosed with a malignant breast tumor. Concave points_worst, radius_worst, concave points_mean, and area_worst are the four next most important features, and all have higher values, the higher the value of SHAP. It indicates that the greater concavity of the nucleus, the larger the radius and the larger the area, all increase the risk of the patient being diagnosed with a malignant breast tumor.

V. RESULTS AND ANALYSIS OF EXPERIMENTS

The operating system used for the experiments is Windows 11, the development environment is Python 3.9.7, the processor is Intel(R) Core (TM) i5-1155G7@2.50GHz2.50GHz, and the memory is 8.00 GB. The experiments use the WDBC dataset, and the results are obtained by applying XGBoost, RF, SVM, KNN, and LR models to classify and predict the benignity and malignancy of breast masses.

A. PARAMETER SETTING OF EXPERIMENTS

During the experiments, 70% of the sample data are used as the training set and 30% of the sample data are used as the test set, and the best combination of parameters for each model is found by the grid search method, as shown in Table 6. The grid search method is used to improve the accuracy of the model by cyclically traversing all possible values of the parameters, comparing and analyzing the effect of each parameter combination on the training of the model, and finally selecting the parameter combination with the best training effect [47]. Also, to make the results of the test more reliable, the analysis is performed using the 10-fold cross-validation method. 10-fold cross-validation means that when training the model, the trained samples are divided into 10 parts, where 1 part of the data is left to validate the model and the remaining 9 samples are trained. The cross validation is repeated 10 times and the average of the 10 test results is used as the final result [48].

B. COMPARATIVE ANALYSIS OF MODELS

In this paper, we propose a method of data processing with a hybrid strategy, and verify the effectiveness of the proposed method by setting up a control group and an experimental group for comparative analysis, as shown in Table 7. The data of the control group are not processed by any method, and the machine learning method is used directly for classification prediction. And the data of the experimental group are processed by one of the hybrid strategy methods proposed in this paper first, and then the classification prediction is performed using the machine learning method.

As can be seen from Table 7, among the five machine learning models, the control group has the best prediction results using the XGBoost model, with accuracy, precision, recall, and F1-score of 97.89%, 98.11%, 96.21%, and 97.12%, respectively. And the experimental group has the best prediction results using RF model, with accuracy, precision, recall, and F1-score of 99.52%, 99.08%, 99.68%, and 99.21%, respectively.

The comparative analysis of Figure 4 reveals that the prediction results of the experimental group are significantly better than those of the control group. For the XGBoost model, the accuracy, precision, recall, and F1-score improve by 1.31%, 0.95%, 3.14%, and 2.08%, respectively. For the RF model, the accuracy, precision, recall, and F1-score improve by 3.21%, 2.84%, 6.74%, and 3.57%, respectively. For the SVM model, the accuracy, precision, recall, and F1-score improve by 6.84%, 2.31%, 17.61%, and 10.88%, respectively. For the KNN model, the accuracy, precision, recall, and F1-score improve by 5.09%, 5.11%, 9.94%, and 7.72%, respectively. For the LR model, the accuracy, precision, recall, and F1-score improve by 4.33%, 4.88%, 7.58%, and 6.43%, respectively. Among all four-evaluation metrics, the percentage of improvement is higher for the recall rate. The recall rate can well reflect the percentage of malignant tumors predicted in breast cancer patients, and the higher recall rate indicates the better prediction result of the model. In summary, a hybrid strategy of data processing proposed in this paper can effectively improve the prediction results of the model and construct a more accurate and efficient breast cancer diagnosis model.

C. COMPARISON ANALYSIS OF SINGLE STRATEGY AND HYBRID STRATEGY

In this paper, three strategies are used for comparative analysis to investigate the impact of two single strategies in a mixed strategy. First, strategy 1 uses a combined sampling method SMOTE-ENN to address the sample imbalance. This approach is able to reduce the a priori information of the model learning sample proportions. Then, strategy 2 uses a mutual information method and a recursive feature elimination method based on the XGBoost algorithm for feature selection. This eliminates features that are irrelevant or redundant to the problem, thus reducing the feature dimensionality of the dataset. Finally, strategy 3 is a hybrid strategy, which

TABLE 10. Accuracy comparison of this study with previous studies.

Author Name	Reference	Year	Model	Accuracy
M. Abdar et al.	[12]	2020	SV-Naïve Bayes-3	98.07%
S. Vijayalakshmi et al.	[13]	2020	POS-NDS	98.80%
A. Kumar et al.	[14]	2020	GP	99.12%
S. A. El Rahman et al.	[15]	2021	RBF-SVM	99.04%
P. Stephan et al.	[16]	2021	HAW-RP	98.50%
U. Naseem et al.	[17]	2022	(SVM+LR+NB+DT) + ANN	98.83%
E. Badr et al.	[18]	2022	GWO-SVM	99.30%
M. H. Alshayeji et al.	[19]	2022	ANN	99.47%
L. K. Singh et al.	[20]	2023	ESGSA	98.96%
this study		2023	Hybrid strategy + RF	**99.52%**

combines strategies 1 and 2 by first using the combined sampling method SMOTE-ENN to solve the sample imbalance problem, followed by feature selection by the mutual information method and the recursive feature elimination method. In order to eliminate the influence of the magnitude and order of magnitude of the sample data in the combined analysis, the data of each strategy are standardized. After the data are processed by each strategy, five different machine learning models are used for classification prediction, and the experimental results are shown in Table 8.

As can be seen from Table 8, for strategy 1, XGBoost and RF model can provide better prediction results. In the prediction results of XGBoost model, the Accuracy and F1-score are higher, which are 99.03% and 99.05%, respectively. The Precision and Recall of RF model were higher, 98.59% and 98.68, respectively. For strategy 2, SVM and RF models were used to predict better results. The Accuracy, Recall and F1-score of the SVM model were relatively high, which were 97.01%, 95.76% and 95.95, respectively. The Precision of RF model is 97.11. For the hybrid strategy, the prediction results of RF model were better, with Accuracy, Precision, Recall and F1-score of 99.52%, 99.08%, 99.68% and 99.21%, respectively.

Through the comparative analysis of Figure 5, it is found that the prediction result of strategy 1 is better than strategy 2 on the whole, but worse than the mixed strategy. However, there are some exceptions, and the prediction results of SVM model are better in Accuracy, Recall and F1-score under strategy 2 than strategy 1. Strategy 2 is better than strategy 1 in Accuracy and Precision of KNN model. The possible reason is that feature selection processing has great influence on SVM and KNN models. Overall, the mixed strategy predicted the best results. It can be inferred that the training effect of the model can be effectively improved by balancing the data. Although the enhancement effect of the model is not very obvious when the data is processed by feature selection, it can further improve the training effect of the model on the basis of balanced.

D. COMPUTATION TIME OF THE TRAINING MODEL

The training time of a computational model is one of the important factors to evaluate the computational cost of a model. In the field of machine learning, training of models is an iterative process that usually requires significant computational resources and time. Knowing the time required for model training can help researchers better plan and manage resources, evaluate the feasibility of a model, and compare it with other testing methods. Therefore, in this paper, the total time for training each model under different strategies is calculated, as shown in Table 9.

Based on the data in Table 9, it can be observed that strategy 1 requires less time, but the classification performance of the model is lower. Strategy 2 requires more time, but the classification performance of the model is higher. The hybrid strategy, on the other hand, slightly exceeds the time consumption of strategy 2 and the model has the best classification performance. This time difference is mainly due to the fact that more time is spent in the process of feature selection on the data. In the hybrid strategy, the XGBoost model and the RF model consume similar amounts of time, however the RF model performs better in terms of classification performance.

E. COMPARATIVE ANALYSIS WITH OTHER STUDIES

The best results of this paper are compared with those of previous studies using the same WDBC dataset, as shown in Table 10. The comparative analysis shows that the research in this paper is better than the results of previous studies with an accuracy rate of 99.52%. The reason for the better results in this paper is the proposed method of data processing with a hybrid strategy, and the processed data facilitate the training of machine learning models, thus improving the accuracy of prediction.

VI. CONCLUSION

The incidence and mortality rate of breast cancer is increasing year by year and has become the number one cancer among women worldwide. In the medical field, the diagnosis and treatment of breast cancer relies heavily on early detection and treatment, and the earlier the treatment, the better the clinical outcome for patients. However, the accurate analysis of breast tumors is a time-consuming and challenging task.

In this study, we propose a method for data processing with a hybrid strategy. Firstly, a combined SMOTE-ENN sampling method is used to solve the problem of sample imbalance.

Then, the features of the dataset are initially screened using the mutual information method, and further the recursive feature elimination method based on the XGBoost algorithm is used for secondary feature selection to derive the best feature subset, and the interpretation method of SHAP is used to analyze the impact of different features. Finally, five different machine learning models, XGBoost, RF, SVM, KNN, and LR, are used for classification and prediction, and the grid search method is used to find the best combination of parameters for each model to construct a more accurate and efficient breast cancer diagnosis model. The experimental results find that the best prediction results are obtained using the RF model, with the accuracy, precision, recall, and F1-score of 99.52%, 99.08%, 99.68%, and 99.21%, respectively. The hybrid strategy proposed in this paper can improve the prediction results of the model, in which the strategy of balanced processing has a greater improvement on the model training, and further using the strategy of feature selection can achieve the maximum accuracy with the least number of features. This indicates that a hybrid strategy of data processing proposed in this study can effectively improve the prediction performance of the model. These techniques, machine learning models and prediction results can help physicians to be able to diagnose patients' tumor conditions accurately and efficiently.

In the future, we will continue to delve deeper into the field of machine learning-based research for breast cancer adjuvant diagnosis in an effort to reduce the research gap that currently exists. We will further develop methods for data imbalance, improve the interpretability of models, and try to investigate generalization methods across datasets and techniques for multimodal data fusion. Through further research and exploration, we will improve the accuracy, reliability, and utility of machine learning-based breast cancer adjuvant diagnosis and promote its application in clinical practice.

REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA, Cancer J. Clinicians*, vol. 73, no. 1, pp. 17–48, Jan. 2023.

[2] M. S. Iqbal, W. Ahmad, R. Alizadehsani, S. Hussain, and R. Rehman, "Breast cancer dataset, classification and detection using deep learning," *Healthcare*, vol. 10, no. 12, p. 2395, Nov. 2022.

[3] Z. Cai, R. C. Poulos, J. Liu, and Q. Zhong, "Machine learning for multi-omics data integration in cancer," *iScience*, vol. 25, no. 2, Feb. 2022, Art. no. 103798.

[4] M. T. Ramakrishna, V. K. Venkatesan, I. Izonin, M. Havryliuk, and C. R. Bhat, "Homogeneous AdaBoost ensemble machine learning algorithms with reduced entropy on balanced data," *Entropy*, vol. 25, no. 2, p. 245, Jan. 2023.

[5] P. E. Jebarani, N. Umadevi, H. Dang, and M. Pomplun, "A novel hybrid K-means and GMM machine learning model for breast cancer detection," *IEEE Access*, vol. 9, pp. 146153–146162, 2021.

[6] M. R. Keyvanpour, M. B. Shirzad, and L. Mahdikhani, "WARM: A new breast masses classification method by weighting association rule mining," *Signal, Image Video Process.*, vol. 16, no. 2, pp. 481–488, Mar. 2022.

[7] S. K. Jha, J. Wang, and R. Shanmugam, "An accurate soft diagnosis method of breast cancer using the operative fusion of derived features and classification approaches," *Expert Syst.*, vol. 39, no. 7, Aug. 2022, Art. no. e12976.

[8] K. S. Rekha and S. A. M. J. Amali, "Efficient feature subset selection and classification using Levy flight-based cuckoo search optimization with parallel support vector machine for the breast cancer data," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 3, pp. 869–881, May 2022.

[9] B. Sowan, M. Eshtay, K. Dahal, H. Qattous, and L. Zhang, "Hybrid PSO feature selection-based association classification approach for breast cancer detection," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5291–5317, Mar. 2023.

[10] S. Alsubai, A. Alqahtani, and M. Sha, "Genetic hyperparameter optimization with modified scalable-neighbourhood component analysis for breast cancer prognostication," *Neural Netw.*, vol. 162, pp. 240–257, May 2023.

[11] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognit. Lett.*, vol. 132, pp. 123–131, Apr. 2020.

[12] S. Vijayalakshmi, A. John, R. Sunder, S. Mohan, S. Bhattacharya, R. Kaluri, G. Feng, and U. Tariq, "Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 11, Nov. 2020, Art. no. 155014772097150.

[13] A. Kumar, N. Sinha, and A. Bhardwaj, "A novel fitness function in genetic programming for medical data classification," *J. Biomed. Informat.*, vol. 112, Dec. 2020, Art. no. 103623.

[14] S. A. El-Rahman, "Predicting breast cancer survivability based on machine learning and features selection algorithms: A comparative study," *J. Ambient Intell. Hum. Comput.*, vol. 12, no. 8, pp. 8585–8623, Aug. 2021.

[15] P. Stephan, T. Stephan, R. Kannan, and A. Abraham, "A hybrid artificial bee colony with whale optimization algorithm for improved breast cancer diagnosis," *Neural Comput. Appl.*, vol. 33, no. 20, pp. 13667–13691, Oct. 2021.

[16] U. Naseem, J. Rashid, L. Ali, J. Kim, Q. E. U. Haq, M. J. Awan, and M. Imran, "An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers," *IEEE Access*, vol. 10, pp. 78242–78252, 2022.

[17] E. Badr, S. Almotairi, M. A. Salam, and H. Ahmed, "New sequential and parallel support vector machine with grey wolf optimizer for breast cancer diagnosis," *Alexandria Eng. J.*, vol. 61, no. 3, pp. 2520–2534, Mar. 2022.

[18] M. H. Alshayeji, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomed. Signal Process. Control*, vol. 71, Jan. 2022, Art. no. 103141.

[19] L. K. Singh, M. Khanna, and R. Singh, "Artificial intelligence based medical decision support system for early and accurate breast cancer prediction," *Adv. Eng. Softw.*, vol. 175, Jan. 2023, Art. no. 103338.

[20] Z. Mushtaq, M. F. Qureshi, M. J. Abbass, and S. M. Q. Al-Fakih, "Effective kernel-principal component analysis based approach for Wisconsin breast cancer diagnosis," *Electron. Lett.*, vol. 59, no. 2, Jan. 2023.

[21] T. R. Mahesh, V. V. Kumar, V. Muthukumaran, H. K. Shashikala, B. Swapna, and S. Guluwadi, "Performance analysis of XGBoost ensemble methods for survivability with the classification of breast cancer," *J. Sensors*, vol. 2022, pp. 1–8, Sep. 2022.

[22] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *J. Pers. Med.*, vol. 11, no. 2, p. 61, Jan. 2021.

[23] P. Liu, B. Fu, S. X. Yang, L. Deng, X. Zhong, and H. Zheng, "Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 1, pp. 148–160, Jan. 2021.

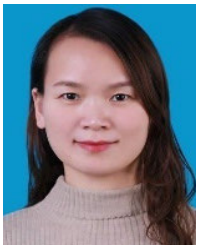
[24] J. Y. Tan, J. Adeoye, P. Thomson, D. Sharma, P. Ramamurthy, and S.-W. Choi, "Predicting overall survival using machine learning algorithms in oral cavity squamous cell carcinoma," *Anticancer Res.*, vol. 42, no. 12, pp. 5859–5866, Dec. 2022.

[25] V. A. Binson, M. Subramoniam, Y. Sunny, and L. Mathew, "Prediction of pulmonary diseases with electronic nose using SVM and XGBoost," *IEEE Sensors J.*, vol. 21, no. 18, pp. 20886–20895, Sep. 2021.

[26] M. U. Rehman, A. Shafique, Y. Y. Ghadi, W. Boulila, S. U. Jan, T. R. Gadekallu, M. Driss, and J. Ahmad, "A novel chaos-based privacy-preserving deep learning model for cancer diagnosis," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 6, pp. 4322–4337, Nov. 2022.

[27] Q. M. Ilyas and M. Ahmad, "An enhanced ensemble diagnosis of cervical cancer: A pursuit of machine intelligence towards sustainable health," *IEEE Access*, vol. 9, pp. 12374–12388, 2021.

[28] Z. Zhao and J. Chen, "A robust discretization method of factor screening for landslide susceptibility mapping using convolution neural network, random forest, and logistic regression models," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 408–429, Dec. 2023.
[29] L. Zou, X. Luo, Y. Zhang, X. Yang, and X. Wang, "HC-DTTSVM: A network intrusion detection method based on decision tree twin support vector machine and hierarchical clustering," *IEEE Access*, vol. 11, pp. 21404–21416, 2023.
[30] M. Aghaabbasi, M. Ali, M. Jasinski, Z. Leonowicz, and T. Novák, "On hyperparameter optimization of machine learning methods using a Bayesian optimization algorithm to predict work travel mode choice," *IEEE Access*, vol. 11, pp. 19762–19774, 2023.
[31] G. Dong and X. Mu, "A novel second-order cone programming support vector machine model for binary data classification," *J. Intell. Fuzzy Syst.*, vol. 39, no. 3, pp. 4505–4513, Oct. 2020.
[32] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020.
[33] P. J. Sudharshan, C. Petitjean, F. Spanhol, L. E. Oliveira, L. Heutte, and P. Honeine, "Multiple instance learning for histopathological breast cancer image classification," *Expert Syst. Appl.*, vol. 117, pp. 103–111, Mar. 2019.
[34] W. Xing and Y. Bei, "Medical health big data classification based on KNN classification algorithm," *IEEE Access*, vol. 8, pp. 28808–28819, 2020.
[35] J. Zhang, C. Zhan, C. Zhang, Y. Song, X. Yan, Y. Guo, T. Ai, and G. Yang, "Fully automatic classification of breast lesions on multi-parameter MRI using a radiomics model with minimal number of stable, interpretable features," *La Radiol. Med.*, vol. 128, no. 2, pp. 160–170, Jan. 2023.
[36] N. Steinauer, K. Zhang, C. Guo, and J. Zhang, "Computational modeling of gene-specific transcriptional repression, activation and chromatin interactions in leukemogenesis by LASSO-regularized logistic regression," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2109–2122, Nov. 2021.
[37] N. Al Mudawi and A. Alazeb, "A model for predicting cervical cancer using machine learning algorithms," *Sensors*, vol. 22, no. 11, p. 4132, May 2022.
[38] M. Khalsan, L. R. Machado, E. S. Al-Shamery, S. Ajit, K. Anthony, M. Mu, and M. O. Agyeman, "A survey of machine learning approaches applied to gene expression analysis for cancer prediction," *IEEE Access*, vol. 10, pp. 27522–27534, 2022.
[39] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective class-imbalance learning based on SMOTE and convolutional neural networks," *Appl. Sci.*, vol. 13, no. 6, p. 4006, Mar. 2023.
[40] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020.
[41] J. Wang, "Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques," *Math. Biosci. Eng.*, vol. 19, no. 10, pp. 10407–10423, 2022.
[42] X. Tang, L. Cai, Y. Meng, C. Gu, J. Yang, and J. Yang, "A novel hybrid feature selection and ensemble learning framework for unbalanced cancer data diagnosis with transcriptome and functional proteomic," *IEEE Access*, vol. 9, pp. 51659–51668, 2021.
[43] K. Yoshida, H. Kawashima, T. Kannon, A. Tajima, N. Ohno, K. Terada, A. Takamatsu, H. Adachi, M. Ohno, T. Miyati, S. Ishikawa, H. Ikeda, and T. Gabata, "Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using radiomics of pretreatment dynamic contrast-enhanced MRI," *Magn. Reson. Imag.*, vol. 92, pp. 19–25, Oct. 2022.
[44] D. K. Rakesh and P. K. Jana, "A general framework for class label specific mutual information feature selection method," *IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 7996–8014, Dec. 2022.
[45] C. Peng, X. Wu, W. Yuan, X. Zhang, Y. Zhang, and Y. Li, "MGRFE: Multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 621–632, Mar. 2021.
[46] D.-C. Feng, W.-J. Wang, S. Mangalathu, and E. Taciroglu, "Interpretable XGBoost-SHAP machine-learning model for shear strength prediction of squat RC walls," *J. Struct. Eng.*, vol. 147, no. 11, Nov. 2021, Art. no. 04021173.
[47] Y. Li, J. Chang, and Y. Tian, "Improved cost-sensitive multikernel learning support vector machine algorithm based on particle swarm optimization in pulmonary nodule recognition," *Soft Comput.*, vol. 26, no. 7, pp. 3369–3383, Apr. 2022.
[48] M. Seifallahi, A. H. Mehraban, J. E. Galvin, and B. Ghoraani, "Alzheimer's disease detection using comprehensive analysis of timed up and go test via Kinect V.2 camera and machine learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 1589–1600, 2022.



HUA CHEN received the Ph.D. degree in applied mathematics from the School of Mathematics and Statistics, Wuhan University, in 2012. She has been an Associate Professor with the School of Science, Hubei University of Technology, since 2015. Her research interests include machine learning, information security, and cryptography.

KEHUI MEI received the B.E. degree in mathematics and applied mathematics from the School of Mathematics and Computer Science, Jianghan University, in 2020. He is currently pursuing the master's degree in applied statistics with the School of Science, Hubei University of Technology. His research interests include data mining and privacy protection.

YUAN ZHOU received the M.S. degree in applied statistics from the College of Science, Hubei University of Technology, in 2023. She is currently with the Wuhan Institute of Biological Engineering. Her main research interests include data mining and privacy protection.

NAN WANG received the B.E. degree from the School of Science, Hubei University of Technology, in 2020, where she is currently pursuing the master's degree. Her research interests include data mining and machine learning.

GUANGXING CAI has been a Professor with the Hubei University of Technology for many years. His research interests include mathematics, information and coding, and cryptography.