scientific reports

Check for updates

OPEN

Application of an improved LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization for breast cancer diagnosis

Xiaoyan Sun

Cancer, as a shocking disease, is one of the most common malignant tumors among women, posing a huge threat to the physical health and safety of women worldwide. With the continuous development of science and technology, more and more high and new technologies are involved in the diagnosis and prediction of breast cancer. In recent years, intelligent medical assistants supported by data mining and machine learning algorithms have provided necessary support for doctors' diagnosis. This study proposes an improved LightGBM hybrid integration model. Introducing gradient harmonic loss and cross entropy loss to enhance the model's attention to minority classes in the dataset and alleviate the impact of data imbalance on diagnostic results. Designing whale optimization algorithm to improve LightGBM to achieve iterative optimization of hyperparameters, and enhance the overall performance of the model. Proposing Jacobian regularization method to denoise LightGBM to solve the problem of model sensitivity to noise. Developing the LightGBM hybrid integration model to ensure the accuracy and stability of model diagnosis on diverse and imbalanced datasets. The effectiveness of the proposed method has been comprehensively compared and verified through the dataset in the UCI machine learning repository, and the results show that the proposed method has achieved good diagnostic performance in all indicators. The hybrid integration model proposed in this paper can provide effective auxiliary support for doctors to diagnose breast cancer.

Keywords Breast cancer diagnosis, Whale optimization algorithm (WOA), Light gradient boosting machine (LightGBM), Hybrid integration, Data mining

With the gradual development of the social economy and the continuous improvement of material living standards, people's emphasis on physical health is increasing day by day. As a malignant tumor with a high incidence rate in the world, breast cancer has a great impact on women's health[1–3]. In the clinical diagnosis of cancer, doctors often combine various physiological indicators and imaging data of patients with experience to make a diagnosis. However, with the updating and iteration of science and technology, as well as the continuous development of statistical disciplines, the diagnosis of cancer is no longer limited to traditional data surfaces. It is also necessary to explore the hidden information behind the data, discover more statistical patterns from the data, and provide valuable reference for doctors' diagnosis[4,5].

Research shows that the 5-year average survival rate of early detection and treatment of breast cancer can reach more than 90%. Due to the lack of awareness of regular screening, many patients go to the hospital only after they have symptoms. During this period, the examination result is usually breast cancer, and metastasis has occurred, such as brain metastasis, lung metastasis, etc. The 5-year survival rate of these patients is only about 20%. For breast cancer, early detection, early diagnosis and early treatment can effectively cut off the progress of cancer, achieve clinical cure, and enable patients to better survive[6,7].

At present, the focus and difficulty of clinical research is how to carry out early diagnosis and treatment for breast cancer, so as to improve the survival period and quality of life of patients[8,9]. With the continuous development of medicine, the current clinical auxiliary diagnosis technology is constantly developing. The diagnosis of breast cancer mainly includes three types of technology: medical imaging diagnosis, molecular

Obstetrics and Gynecology, Jinan Maternity and Child Care Hospital, Jinan 250000, Shandong, China. email: drsxy0302@163.com

biology diagnosis and pathological biopsy. Generally speaking, machine learning algorithms such as data mining and neural networks can provide necessary assistance for doctors to quickly diagnose.

In recent years, machine learning has made unprecedented progress in data processing and classification, providing new ideas for solving key problems in the diagnosis and treatment of breast cancer. By integrating breast cancer data and using advanced machine learning and artificial intelligence technology to mine in-depth information, the speed and accuracy of disease diagnosis have been improved. The purpose of studying the diagnosis and prediction method of breast cancer based on neural network models is to combine clinical practice with machine learning[10–14], which will help medical workers to more quickly and accurately determine whether the disease is present, while solving the problems of overfitting, high rates of missed diagnosis and misdiagnosis in existing models, and improving the accuracy of prediction models. The LightGBM model occupies less memory, trains faster, and is more efficient, so it is widely used in disease diagnosis[15]. However, most models overlook that disease datasets are often imbalanced and have not been processed beforehand. Moreover, the LightGBM model has numerous parameters and is sensitive to noise. Therefore, it is necessary to establish a reasonable model to automatically optimize and denoise its parameters. Therefore, in order to further improve the accuracy of breast cancer diagnosis, this paper proposes an improved LightGBM hybrid integration model combining gradient harmonization regularization and Jacobian regularization for breast cancer diagnosis. Firstly, gradient harmonic loss and cross entropy loss are introduced to improve the imbalance of breast cancer dataset; Secondly, a whale optimization algorithm is designed to iteratively optimize the parameters of the LightGBM model to improve the performance of the breast cancer diagnosis; Then, the Jacobian regularization method proposed in this paper is used to reduce the noise of LightGBM, which solves the problem that the breast cancer diagnosis model is sensitive to noise; Finally, the LightGBM hybrid integrated model is developed to ensure the accuracy and stability of model diagnosis under the diverse and imbalanced dataset of breast cancer. The main contributions can be summarized as follows:

(1) An improved LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization is proposed for breast cancer diagnosis.
(2) Design whale optimization algorithm, gradient harmonic loss, and Jacobian regularization method to continuously optimize the parameters of LightGBM model, improving the problems of data imbalance and noise sensitivity from the perspective of objective function.
(3) Develop LightGBM hybrid integrated model to ensure the accuracy and stability of model diagnosis under the diverse and imbalanced of breast cancer.

The main structure of this paper can be summarized as follows: "Related works" reviews the related work in the field of breast cancer diagnosis. The technical background for breast cancer diagnosis is given in "Background". The "Proposed method" includes a detailed explanation of the proposed method in this paper. The evaluation of the proposed method is completed in the "Performance evaluation" section. Finally, the "Conclusion" section summarizes the paper.

Related works

In recent years, many scholars have conducted relevant research on the prediction or diagnosis of breast cancer, which has made great contributions to the development of modern medicine. The diagnosis model of breast cancer can be divided into traditional model, machine learning model and deep learning model. In recent years, deep learning has made significant breakthroughs in fields such as speech recognition, visual recognition, multimedia processing, medical diagnosis, and biomedical research. Many breast cancer diagnosis systems have been implemented based on data mining and machine learning algorithms. The following will review some work related to breast cancer diagnosis.

Youness et al.[16] compared the use of nonlinear algorithms such as random forest, naive Bayes, support vector machine, and K-nearest neighbor to predict cancer. The author compared data mining algorithms based on the selection of the best classifier using bioinformatics and medical classification techniques, and ultimately concluded that support vector machine (SVM) was more suitable than other algorithms, with an accuracy of 97.9%. Sara et al.[17]considered the data of breast cancer gene expression (GE) and DNA methylation (DM), and selected three different classification algorithms: support vector machine (SVM), decision tree and random forest to create nine models that help predict cancer. Finally, they thought that the scaling SVM classifier in Spark environment was superior to other classifiers because it achieved the highest accuracy and lowest error rate in GE data set.

Huang et al.[18] proposed a new hierarchical clustering random forest (HCRF) model to improve the diagnostic accuracy of breast cancer. Firstly, decision trees are generated using the random forest algorithm, and hierarchical clustering techniques are used to perform similarity analysis and clustering on these decision trees, in order to select representative decision trees from each cluster to construct a random forest with both low similarity and high accuracy. Next, the Variable Importance Measurement (VIM) method is used to optimize feature selection, which calculates the importance of each feature and ranks them to remove less important features and leave the optimal feature subset that can effectively improve classification performance. When building the decision tree, Gini index is used to select the best partition features to ensure the highest purity of subsets, more accurate classification, and ultimately effectively improve the accuracy of breast cancer diagnosis.

Zheng et al.[19] developed a hybrid algorithm of K-means and support vector machine (K-SVM) algorithm, and used the K-means algorithm to identify the hidden patterns of benign and malignant tumors respectively. Finally, on the breast cancer related data set, they thought that the accuracy of the method would be improved to 97.38%. Habib et al.[20]reduced the dimensions of breast cancer data set using the principal component analysis method. On this basis, they used KNN algorithm and naive Bayesian algorithm to predict the diagnosis of

breast cancer patients, and finally believed that the model obtained using KNN algorithm could achieve the best results. Bikku et al.[21] developed an improved quantum neural network algorithm for structural analysis of molecules pertaining to drug discovery. The algorithm was designed for enhancing noise rejection, scalability, and optimization. Lin et al.[22] proposed a deep neural network model (DeepMO) based on the integration of multi omics data, aiming to effectively classify the subtypes of breast cancer. This model includes three coding subnets and one classification subnet. Each encoding subnet is targeted at different types of omics data (mRNA data, DNA methylation data, and CNV data), and features are learned through feedforward networks and optimized using Relu activation functions, dropout, and batch normalization techniques. The features output by the encoding subnet are connected and L2 normalized to form an integrated representation of multiple omics data, which is then input into the classification subnet for predicting cancer subtypes. According to the nature of the problem (binary classification or multi classification), the classification subnet uses the Sigmoid activation function in combination with binary cross entropy loss or SoftMax regression in combination with cross entropy loss to regularize the model to achieve accurate classification of breast cancer subtypes. The whole model is trained end-to-end, which effectively utilizes the complementarity of different kinds of omics data to improve the accuracy of breast cancer subtype classification.

Moloud et al.[23] used a nested ensemble method combining stacking and voting as classifiers to detect benign and malignant breast tumors, and ultimately concluded that the proposed two-layer nested ensemble model was superior to a single classifier and most classifiers. Zhang et al.[24] proposed a new approach to enhance Adaboost through resampling versions. In the local boosting algorithm, local errors are calculated for each training instance, and then used to update the probability of that instance being used in the training set for the next iteration. Diao et al.[25] developed a feature selection approach that promotes the selection of basic models by treating classifiers as features after converting set predictions into training samples. Yu et al.[26] proposed a Hybrid Incremental Ensemble Learning (HIEL) method that considers both feature space and sample space for classifier selection to handle noisy datasets. Tian et al.[27] proposed the XGBoost algorithm based on the GBDT algorithm, which significantly improved the recognition accuracy and running speed of the model on the basis of GBDT.

Background

In this section, the basic concepts of the research are explained, including the LightGBM diagnostic model, swarm intelligent optimization algorithm, which will be further applied to the proposed method.

LightGBM diagnostic model

Light Gradient Boosting Machine (LightGBM) is an efficient gradient lifting framework based on decision tree[28] proposed by the team of Microsoft Research Asia in 2017. It is one of the best algorithms in machine learning algorithms to fit the real distribution. Compared with traditional gradient boosting algorithms, it has the advantages of faster training speed, lower memory consumption, better accuracy, and the ability to quickly process massive industrial data. It has been widely used in various industries to solve prediction problems such as classification and regression. Its main innovations include: Gradient based One Side Sampling (GOSS), Exclusive Feature Bundling (EFB), Histogram based decision tree algorithm, and Leaf-wise leaf growth strategy with depth constraints. The algorithm flowchart of LightGBM is shown in Fig. 1.

(1) Gradient based one side sampling (GOSS)

GOSS is a down-sampling algorithm whose main idea is that sample points with large gradients play a major role in calculating information gain. It can be understood that sample points with large gradients have a greater impact on information gain. Therefore, in order to maintain the accuracy of information gain evaluation,

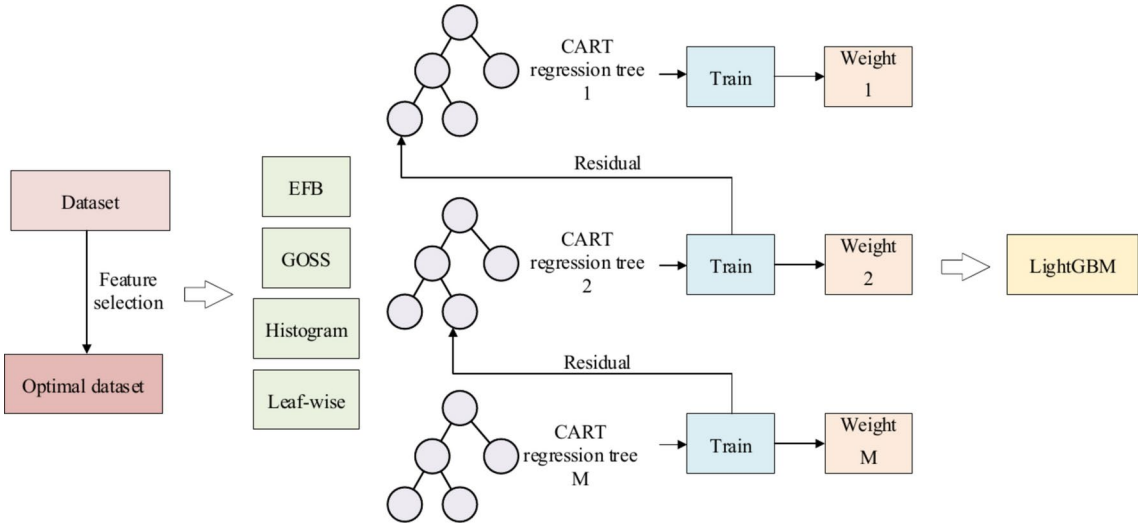

Fig. 1. Basic framework of LightGBM.

#	Feature 1	Feature 2	Feature 3
1	0	2	0
2	0	0	0
3	0	0	0
4	0	0	1
5	3	0	0

Feature New
2
0
0
1
3

Fig. 2. Schematic diagram of EFB.

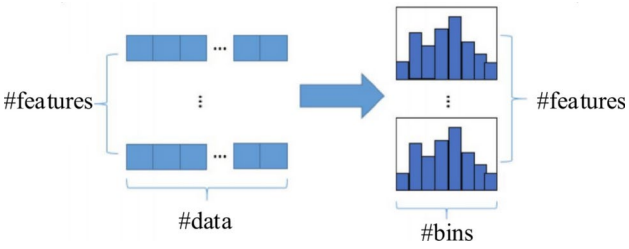

Fig. 3. Histogram algorithm formation process.

the sample points with large gradients are retained during down-sampling, and the sample points with small gradients are randomly sampled proportionally.

(2) Exclusive feature bundling (EFB)

Mutually exclusive feature bundling is a method of reducing feature dimensions, which can be simply understood as bundling mutually exclusive or unrelated features together to form a new feature. This method is very different from the traditional dimension reduction algorithm. The traditional dimension reduction algorithm creates new features that contain as much original information as possible in a linear weighted way based on the correlation between features, while the EFB is different from it. From the perspective of the sparseness of high-dimensional data in practical applications, it makes use of the fact that there are few non-zero values in the feature space at the same time, and binds these mutually exclusive data together as a new feature, which achieves the purpose of dimension reduction without losing feature information. The schematic diagram of EFB algorithm is shown in Fig. 2.

(3) Histogram-based algorithm

To reduce storage capacity and increase computational efficiency, LightGBM introduces histogram algorithm. The basic idea of histogram algorithm is to discretize continuous feature values into K integers and construct a histogram with a width of K. In the process of node splitting in decision trees, only the rough statistical information of the histogram needs to be considered, without traversing all data points, thus significantly reducing the time complexity. The histogram algorithm has many advantages, such as: no need to save pre-sorted results, reducing the computational cost of segmentation gain; Only storing the discretized values of features greatly reduces memory usage; When seeking the leaf nodes of a certain node, the histogram subtraction method can be used to reduce data traversal and improve efficiency. The process of forming the histogram algorithm is shown in Fig. 3.

(4) Leaf-wise leaf growth strategy with depth constraints.

Most decision tree-based algorithms adopt a Level wise strategy, which allows for the simultaneous growth of leaves in the same layer during data traversal without overfitting. However, it treats all leaf nodes in the same layer equally and searches and splits many leaf nodes that do not require splitting or have low splitting gains, thereby increasing computational complexity. The LightGBM algorithm proposes a Leaf wise strategy to address the above shortcomings, as shown in Fig. 4. This method finds the leaf node with the highest splitting gain or loss reduction from all the currently traversed leaves and splits them, continuously reducing the overall model loss through cyclic splitting. Research has shown that under the same number of splits, the Leaf wise strategy reduces errors much more than the Level wise strategy, resulting in higher model accuracy. However, the Leaf wise strategy also has corresponding drawbacks. When traversing multiple leaf nodes, it will always find the node

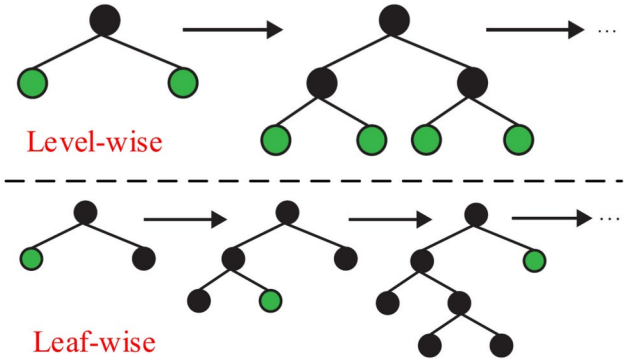

Fig. 4. Schematic diagram of level-wise and leaf-wise.

with the most loss reduction. If the decision tree is not restricted from splitting, it will continue to split according to the point with the most loss reduction, resulting in overfitting. Therefore, researchers have added a maximum splitting depth limit to the Leaf wise growth strategy to prevent overfitting while ensuring high efficiency.

The core idea of LightGBM is to construct a new tree through feature splitting, and iteratively calculate the residual between predicted and actual values based on residual fitting method. Finally, the prediction results of all trees are summed up to obtain the final prediction result. Suppose that the training set Q consists of N breast cancer data samples, where $x_i = [x_i^1, x_i^2, \ldots, x_i^n]$ represents the input sample, n represents the dimension of sample characteristics, and $y_i = [0, 1][0, 1]$ represents the category label, which uses one-hot coding form to represent benign and malignant respectively. The integrated model LightGBM can be represented as:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{1}$$

where $f_k(x_i)$ represents the i-th decision tree; x_i represents the i-th sample feature of the transformer; $\hat{y}_i$ represents the predicted sample category; F represents the tree model; k represents the total number of decision trees. The objective function is constructed as follows:

$$G(x_i) = \sum_{t=1}^{k} \phi\left(y_i, F_{t-1}(x_i) + f_t(x_i)\right) + \sum_{k} \Omega(f_k) \tag{2}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3}$$

where $\phi\left(y_i, F_{t-1}(x_i) + f_t(x_i)\right)$ represents the loss function; $\Omega(f_k)$ represents regularization; $F_{t-1}(x_i)$ represents the top $t-1$ tree models; $f_t(x_i)$ represents the t-th tree model; w represents the weight of each tree leaf; λ, γ is a hyperparameter; T represents the number of leaf nodes in the tree model; y_i represents the true category of the i-th sample.

To minimize the objective function, the LightGBM model uses Newton's method to find a mapping relationship $\overline{G}(x)$ to approximate the objective function $G(x)$, which can be derived as:

$$\overline{G}(x_i) \cong \sum_{i=1}^{n} \left(g_i f_t(x_i) + \frac{h_i f_t^2(x_i)}{2} \right) + \sum_{k} \Omega(f_k) \tag{4}$$

where g_i represents the first derivative of the loss function; h_i represents the second derivative of the loss function.

For the LightGBM classification model, the cross-entropy loss function (CE Loss) is commonly used. Taking binary classification as an example, the cross-Entropy loss can be expressed as:

$$L_{CE}(x_i, y_i) = \frac{-\sum_{i=1}^{N} \left(y_i \log P(\hat{y}_i) + (1 - y_i) \log\left(1 - P(\hat{y}_i)\right)\right)}{N} \tag{5}$$

where N represents the total number of samples; y_i represents the correct category of the i-th sample.

The first and second derivatives corresponding to $\overline{G}(x)$ can be expressed as:

$$g_i = (y_i - P(\hat{y}_i)) \tag{6}$$

$$h_i = (1 - P(\hat{y}_i))P(\hat{y}_i) \tag{7}$$

Due to the imbalance of breast cancer samples and the large proportion of benign samples in the total sample loss, the model will tend to predict those categories with a large proportion and learn too much about the characteristics of benign samples. By hiding the error rate of minority samples through the correctness of the majority class, the overall diagnostic performance of the model shows a high "accuracy", but in reality, it cannot identify minority class samples well.

Swarm intelligent optimization algorithm

For the optimization of parameters in the diagnosis model of breast cancer, many scholars choose to seek solutions from the real models of nature, thus starting the research on nature inspired computing. These algorithms have shown good performance in adaptability, self-learning ability, robustness, and efficiency. In natural heuristic computing, there is a type of swarm intelligence that focuses on individuals with simple behaviors completing complex tasks through self-organization, called swarm intelligence (cluster intelligence), as shown in Fig. 5.

Cluster intelligence leverages all the advantages of a cluster, including self-organization, decentralized control, high robustness, flexibility, and low consumption. It can still provide optimal solutions when facing large-scale complex problems. The earliest cluster intelligence algorithms[29,30], including Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO), have received widespread attention and application, and have achieved great success in many fields, such as the classic traveling salesman problem. They generally have the characteristics of simple structure, few parameters, and easy implementation. Nowadays, it has been widely applied in practical problems such as function optimization, multi-objective optimization, solving integer constraints and mixed integer constraint optimization, neural network training, signal processing, routing algorithms, etc. The practical results have proved the feasibility and efficiency of these algorithms. In recent years, there has been an explosive growth in research on cluster intelligent algorithms for solving complex computing tasks in academia. These studies are mostly related to optimization problems, mainly focusing on fields such as health and hygiene, social networks, transportation, energy and climate, Industry 4.0, etc., and have achieved good results. Cluster intelligent optimization algorithms, in addition to the above-mentioned ones, also include cuckoo algorithm, firefly algorithm, bat algorithm, whale optimization algorithm, etc. Two commonly used swarm intelligent optimization algorithms will be introduced in the following.

Particle swarm optimization (PSO) was proposed by Kennedy and Eberhart in 1995, inspired by the foraging behavior of bird or fish flocks, to solve increasingly complex optimization problems. The two factors that need to be considered in particle swarm optimization are population optimality (g-best) and individual optimality (p-best), and particles update their velocity vectors based on these two factors. Assuming that the problem to be solved is in a D-dimensional search space, at time t, the current position of the i-th particle in the population is represented by a D-dimensional vector $x_i^t = (x_{i1}^t, x_{i2}^t, \ldots, x_{iD}^t)^T$, its velocity is represented by another D-dimensional vector $v_i^t = (v_{i1}^t, v_{i2}^t, \ldots, v_{iD}^t)^T$, the optimal solution position visited by the i-th particle is represented by $p_i^t = (p_{i1}^t, p_{i2}^t, \ldots, p_{iD}^t)^T$, and the index of the optimal particle in the population is g, then in each search iteration, the velocity and position of the i-th particle are updated by the following equations:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 \left(p_i^t - x_i^t\right) + c_2 r_2 \left(p_g^t - x_i^t\right) \tag{8}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{9}$$

The above equations describe the original form of particle swarm optimization algorithm, which has achieved very good results in multiple application scenarios. However, the algorithm itself still has many problems. For example, it is easy to fall into local optima, slow down convergence speed, have random parameter selection, and require multiple adjustments.

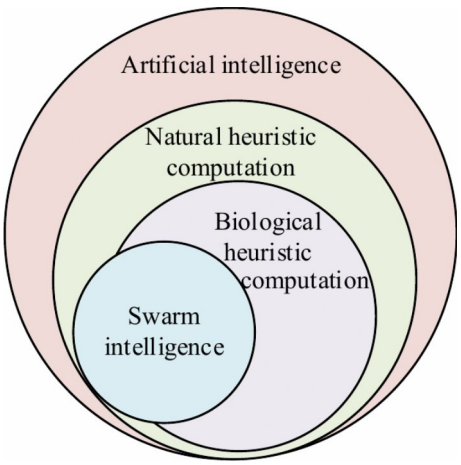

Fig. 5. The relationship between swarm intelligence and artificial intelligence.

Whale Optimization Algorithm (WOA) is a swarm optimization algorithm[31] that simulates the spiral bubble feeding method of humpback whales during the optimization process. Optimization is carried out through three stages: surrounding preys, chasing preys, and attacking preys with bubble nets.

(1) Surrounding the preys. As shown in Fig. 6, the humpback whale uses echo to surround its preys, continuously reducing the encirclement to capture more accurately. Its motion trajectory equation is:

$$\begin{cases} Q^{k+1} = Q^k_{best} - AD \\ D = \left| CQ^k_{best} - Q^k \right| \\ A = 2a\rho_1 - a \\ C = 2\rho_2 \\ a = 2 - \frac{2k}{k_{\max}} \end{cases} \tag{10}$$

where k is the number of iterations; Q^k and Q^k_{best} represent vectors from the starting point to the search agent position and the optimal position, respectively; D is the vector between the whale and its current best individual; A and C are coefficient vectors that control the way whales swim; Set ρ_1 and ρ_2 as random numbers between 0 and 1; a is the convergence factor that linearly decreases from 2 to 0.

(2) Chasing the preys. The spiral ascent mechanism is used to calculate the optimal distance between whales and prey, and its motion trajectory equation is:

$$\begin{cases} Q^{k+1} = Q^k_{best} - D' e^{bl} \cos(2\pi l) \\ D' = Q^k_{best} - Q^k \end{cases} \tag{11}$$

where b is the logarithmic spiral constant, and is a random number between [− 1, 1].

(3) Attacking the preys.

To avoid getting stuck in local optima, random searches will be conducted based on each other's positions. When $|A| > 1$, the search agent is forced to move away from the current agent to update its position, and its motion trajectory equation is:

$$\begin{cases} Q^{k+1} = Q^k_{rand} - AD'' \\ D'' = \left| CQ^k_{rand} - Q^k \right| \end{cases} \tag{12}$$

where Q^k_{rand} represents a randomly selected position vector from the whale population.

Proposed method

The proposed LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization for breast cancer diagnosis involves the following and is shown in Fig. 7 to improve the diagnostic accuracy and stability of the model for breast cancer as a whole. Introducing gradient harmonic loss and cross entropy loss, the model's attention to minority classes in the dataset is enhanced, alleviating the impact of data imbalance on diagnostic results. Designing whale optimization algorithm to improve LightGBM to achieve

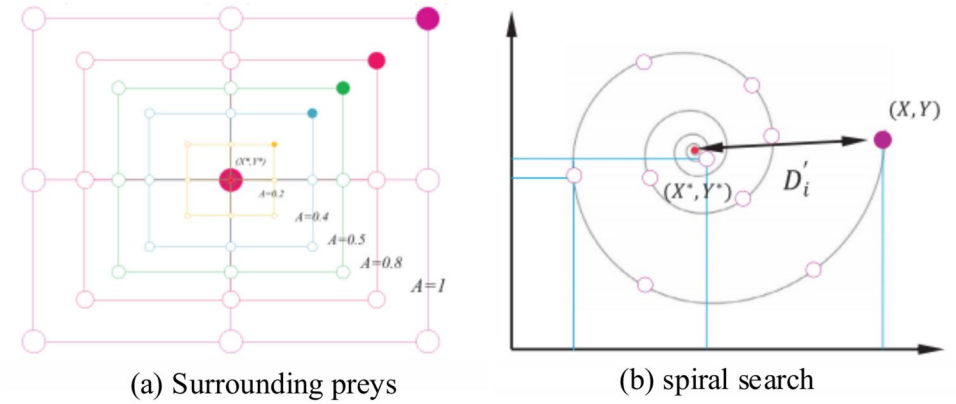

Fig. 6. Search methods for WOA.

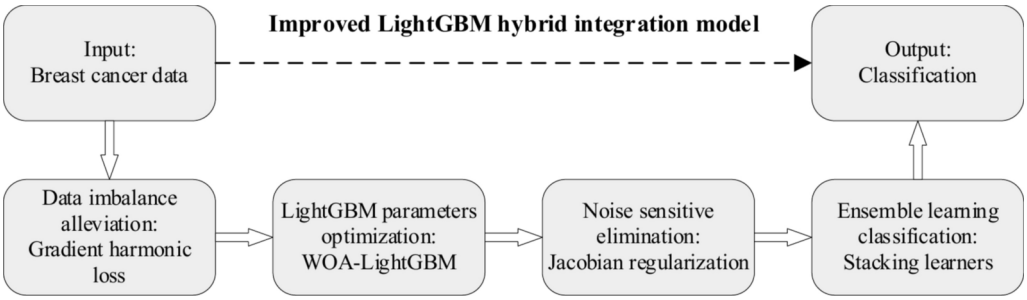

Fig. 7. Flow diagram of the proposed method.

iterative optimization of hyperparameters, enhancing the overall performance of the model. Proposing Jacobian regularization method to denoise LightGBM, solving the problem of model sensitivity to noise. Developing the LightGBM hybrid integrated model to ensure the accuracy and stability of model diagnosis on diverse and imbalanced datasets.

(1) Breast cancer data input: Select the appropriate cancer data based on the research question and download them.
(2) Model training: Preprocess the raw data, and conduct model training, including data imbalance mitigation, LightGBM parameters optimization, noise sensitivity elimination and integrated learning classification.
(3) Output and validation: Output the diagnosis results and validate the findings using dataset.

Data imbalance mitigation: gradient harmonic loss and cross entropy loss

Generally, in the breast cancer dataset, benign samples account for a large proportion, and there is imbalance between different data samples. Using the cross-entropy loss function for training can lead to the model focusing too much on the majority of class samples. Therefore, a gradient harmonic loss function is proposed, which uses a progressive weighted histogram to rank the gradients of samples, and then assigns a dynamic weight to each rank, so that difficult to rank samples (such as extremely small positive samples or highly similar samples) obtain higher weights, thereby promoting the model's learning of these samples. It can effectively alleviate the problem of data imbalance:

$$L_{GHM} = \beta_i L_{CE} \tag{13}$$

where L_{CE} represents the cross-entropy loss function, L_{GHM} represents the gradient harmonic loss function, and β_i is the weight of the i-th sample.

The β_i represents the importance of each sample, calculated by the gradient density $GD(g_i)$ of the i-th sample, which is as follows:

$$\begin{cases} \beta_i = \frac{N}{GD(g_i)} \\ GD(g_i) = \frac{R_{ind(g_i)}}{\zeta} \end{cases} \tag{14}$$

where $R_{ind(g_i)}$ represents the number of samples in the region with g_i as the center and length ζ.

However, through practice, it is found that in the breast cancer diagnosis task, directly using gradient harmonic loss to replace cross entropy loss cannot achieve the best network diagnosis performance. After analysis, it was found that some fault data were mistakenly identified as outliers, resulting in a decrease in weight and ultimately poor diagnostic performance. And cross entropy loss is calculated on a data-by-data basis. Therefore, this article combines gradient harmonic loss and cross entropy loss as the total loss function of the model. The fusion loss function mainly uses gradient harmonic loss to solve the problems of imbalanced data and difficult sample separation, supplemented by cross entropy loss to reduce the misjudgment of outliers by gradient harmonic loss. The fusion loss function can be expressed as:

$$L_{FUS} = (1-\alpha)L_{CE} + \alpha L_{GHM} \tag{14}$$

where α is the hyperparameter proportional coefficient that balances gradient harmonic loss and cross entropy loss. After experience and multiple experimental analyses, it is generally found that the optimal model performance is achieved when the value is 0.8.

According to the analysis, for a small number of samples in the breast cancer dataset—malignant samples, the derivative has fewer samples in the nearby area ζ, which will obtain smaller gradient density $GD(g_i)$, and its loss function will obtain greater weight β_i. On the contrary, for the majority of samples—benign samples, smaller weights will be obtained. Therefore, in order to minimize the objective function, the model will increase its attention to minority class samples, thereby reducing the loss value of minority classes.

LightGBM parameters optimization: WOA-LightGBM

In the LightGBM model, there are numerous hyperparameters included, and different hyperparameter settings have varying impacts on the diagnostic performance of the model. Seven important hyperparameters were selected for optimization in the LightGBM model. These seven parameters have a significant impact on the recognition accuracy, training speed, fitting performance, and diversity of the model, but have a wide range of values, making it difficult to obtain the optimal parameter values based on experience. The Whale Optimization Algorithm (WOA) is a versatile, efficient, and robust optimization method that is well-suited for solving complex, multi-dimensional, and non-linear problems across a wide range of fields. Its ease of implementation and the ability to balance exploration and exploitation make it particularly appealing for real-world applications. In this paper, the WOA algorithm is used to optimize them and establish the WOA LightGBM model. WOA needs to optimize parameter settings as shown in Table 1.

The parameter optimization process of the WOA algorithm for the LightGBM model is essentially to use the above parameters in the model as the position vectors in the WOA algorithm, and through iterative optimization of the WOA algorithm, find the global optimal position in the algorithm and output it as the final parameters of LightGBM. The specific implementation process is as follows:

(1) Initialize the whale population, maximum iteration times, and dimensions in the WOA algorithm, and set the search range for hyperparameters in LightGBM.
(2) Set the fitness function. Firstly, the position of each humpback whale in the WOA algorithm is randomly initialized within the boundary range. The fitness value corresponding to each whale in the current population is calculated through the fitness function, and the position of the humpback whale with the smallest fitness value is selected as the global optimal solution for the current population. In this article, the mean square error function of the LightGBM model is selected as the fitness function, and the specific mathematical expression is:

$$f(x_i) = \frac{1}{n}\sum_{j=1}^{n}(\theta_{i,j} - \hat{\theta}_{i,j})^2 \tag{15}$$

where x_i represents the position of the i-th humpback whale in the current population, $\theta_{i,j}$ is the j-th true value in the i-th individual, and $\hat{\theta}_{i,j}$ is the LightGBM model predicted value corresponding to the true value.

(3) Start iterating. During the iteration process of the WOA algorithm, the individual positions of the humpback whales are updated using three methods: encircling prey, spiral search, and random search, ensuring that the updated individual positions meet the pre-set boundary conditions. If the obtained result exceeds the boundary, randomly generate the position of individual humpback whales within the boundary range;
(4) Calculate the fitness value. After the above three individual position update processes are completed, the obtained position of the humpback whale is input into the fitness function, and the fitness result of each individual is calculated. The position of the individual with the best fitness value is taken as the global optimal solution at this time;
(5) Repeat the above steps until the preset maximum number of iterations is met. At this time, the global optimal solution is the lightGBM model optimal parameter output by the WOA algorithm, and it is brought into the model for breast cancer diagnosis.

The intuitive flowchart corresponding to the above process is shown in Fig. 8.

Noise sensitive elimination: Jacobian regularization

LightGBM adjusts the weights of samples based on the results during each iteration. As the number of iterations increases, the model's bias decreases, making it sensitive to noise. Therefore, this paper proposes a Jacobian regularization method to denoise LightGBM.

Firstly, use the F-norm of the networked Jacobian matrix to regularize it. The idea is to represent the input of the network as a d-dimensional vector and the output as a k-dimensional vector. Assuming that the training dataset consists of N examples, each sample x_i is a D-dimensional vector, the index $l = 1, 2, \ldots, L$ is used to specify a specific layer in the network with L layers. $z^{(l)}$ is the output of the -th layer of the network, and $z_k^{(l)}$ is the output of the k-th neuron in that layer. In addition, λ is used to represent the hyperparameters of the

WOA parameters	Set	LightGBM parameters	Default	Optimization range
Population	10, 30, 50, 100, 150, 200, 250	n_estimators	100	(1, 500)
		num_leaves	31	(2, 2000)
		learning_rate	0.1	(0.001, 1)
max_iter	150	max_bin	255	(2, 255)
		max_depth	-1	(1, 1000)
Dimension	7	subsample	1	(0.1, 1)
		Colsample_bytree	1	(0.1, 1)

Table 1. Description of WOA and LightGBM parameters.

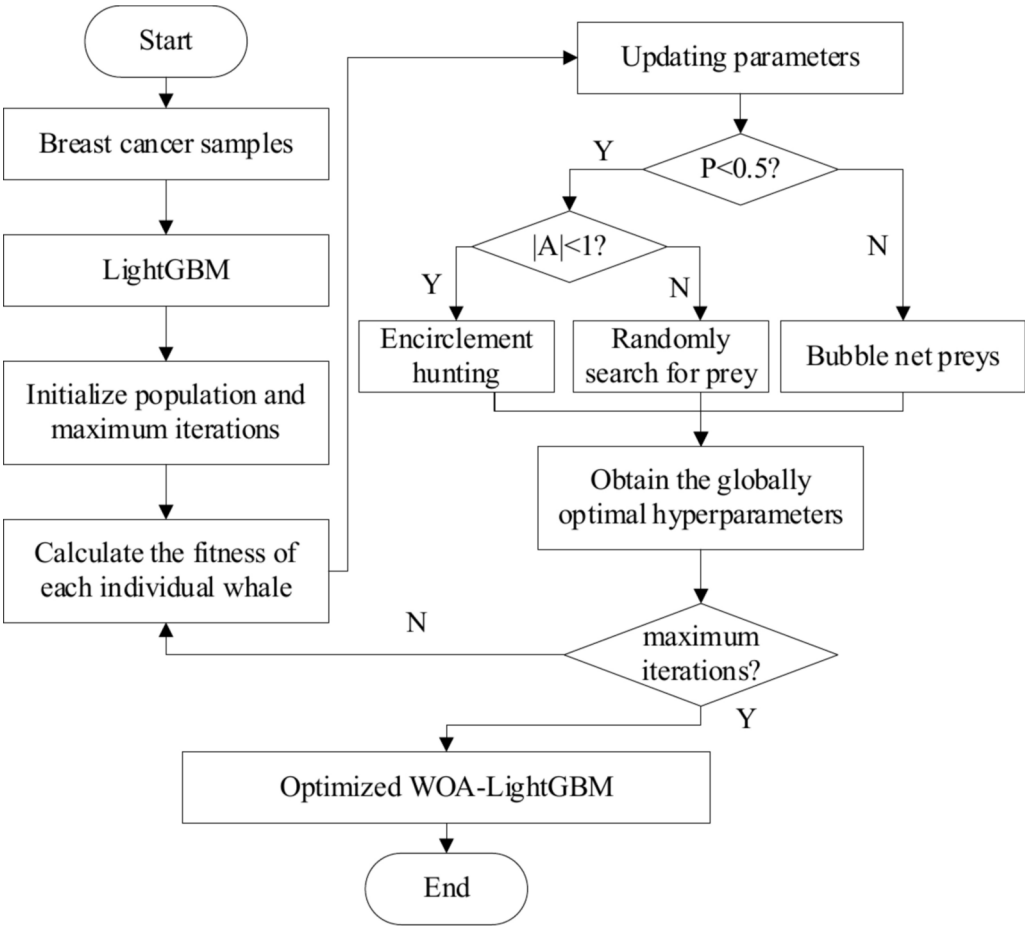

Fig. 8. Flowchart of WOA-LightGBM for breast cancer diagnosis.

regularization penalty term in the control loss function. The input of the network is x_i, the output is $f(x_i)$, and the predicted class of x_i is $k_i^* = \arg\max x_k f_k(x_i)$, where $f(x_i) = soft\max\{z^{(l)}(x_i)\}$ is the output of the last fully connected layer in the network. $\nabla_x z^{(l)}(x_i)$ is the Jacobian matrix of layer L evaluated at point x_i, i.e., $J^{(L)}(x_i) = \nabla_x z^{(l)}(x_i)$. The corresponding $J_k^{(L)}(x_i) = \nabla_x z_k^{(l)}(x_i)$ is the k-th row in matrix $J^{(L)}(x_i)$. The networked Jacobian matrix is:

$$\begin{aligned} J(x_i) &\triangleq J^{(L)}(x_i) \\ &= \begin{pmatrix} \frac{\partial z_1^{(L)}(x_i)}{\partial x_{(1)}} & \cdots & \frac{\partial z_1^{(L)}(x_i)}{\partial x_{(D)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial z_K^{(L)}(x_i)}{\partial x_{(1)}} & \cdots & \frac{\partial z_K^{(L)}(x_i)}{\partial x_{(D)}} \end{pmatrix} \in \mathbb{R}^{K\times D} \end{aligned} \tag{16}$$

Therefore, the Jacobian regularization term for input sample x_i is:

$$\|J(x_i)\|_{\mathrm{F}}^2 = \sum_{d=1}^{D}\sum_{k=1}^{K}\left(\frac{\partial}{\partial x_d} z_k^{(L)}(x_i)\right)^2 = \sum_{k=1}^{K}\left\|\nabla_x z_k^{(l)}(x_i)\right\|_2^2 \tag{17}$$

Therefore, the loss function modified by Jacobian regularization is defined as follows:

$$L_{loss} = L_{FUS} + \lambda\sqrt{\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{n=1}^{N}\left(\frac{\partial}{\partial x_d} z_k^{(L)}(x_i)\right)^2} \tag{18}$$

Ensemble learning classification: stacking learners

The proportion coefficient of the fusion loss function L_{FUS} changes with the change of data samples, and the optimal proportion coefficient α in this paper may not be applicable to other datasets. In order to achieve good performance of LightGBM without the need for experimental determination of the optimal value for practical and variable operating conditions, this paper innovatively integrates LightGBM with different specific values based on the Stacking framework, and proposes a LightGBM hybrid ensemble model that is more suitable for practical and variable operating conditions.

Stacking, also known as stacking generalization[32], is a heterogeneous serial learning method that consists of multiple base learners forming the first layer prediction model and using a meta learner as the fusion model for the second layer. The prediction results of the base learner are used as inputs for the meta learner, and the final prediction results are generated through retraining. The core idea of Stacking is to utilize the prediction results of multiple base learners and improve overall prediction performance through the fusion of meta learners. The performance of the Stacking framework is related to the selection of base and meta learners. The model selection in this paper is as follows:

(1) Selection of base learners: In the base learning layer of the Stacking model, selecting models with significant differences can increase the diversity of the model and reduce the correlation between the output labels of the base learning layer. By using differential models, stacking models can integrate the advantages of various models and model data from multiple perspectives, better adapting to different data distributions and features. This combination of diversity and reduced correlation can bring stronger learning ability and recognition performance to Stacking models. Therefore, when choosing a base learning layer model, priority should be given to models with significant differences to highlight the diversity of the model and reduce the correlation between output labels. This strategy can better utilize the advantages of integrated models and improve overall performance.

This paper selects four LightGBM models with different values as the first layer base learners, namely $\alpha = 0, 0.4, 0.8, 1$. The above models are all based on tree models. To increase the diversity of the first layer base learners, the nonlinear model SVM is also selected as the first layer base learner. SVM[33] can use nonlinear kernel functions $\varphi(x)$ to map samples to high-dimensional feature spaces, making transformer data samples that were originally inseparable in low dimensional spaces separable in high-dimensional spaces. Its principle can be expressed as:

$$\begin{cases} \min \frac{1}{2} \|\omega\|^2 + c \sum_{i=1}^{l} \xi_i \\ \text{s.t.} y_i \left[\omega^T \phi(x_i) + b\right] \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \ldots, l \end{cases} \tag{19}$$

where ω represents the normal vector of the hyperplane, c represents the penalty coefficient vector, ξ_i represents the relaxation coefficient vector, and b represents the bias parameter.

(2) Meta learner selection: The Stacking framework requires the selection of models with strong generalization and learning ability as meta learners. Therefore, this article chooses ELM[34] as the meta learner.

ELM is a machine learning algorithm based on feedforward neural networks, which has the advantages of strong generalization and learning ability. The structure of a single hidden layer ELM is shown in Fig. 9, and its specific principle is as follows:

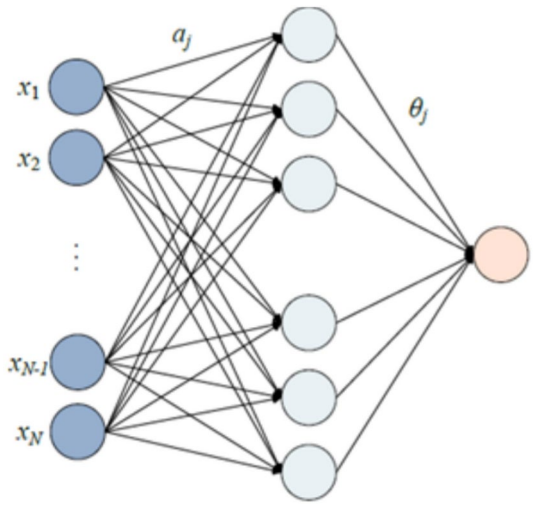

Fig. 9. Single hidden layer ELM structure diagram.

$$\sum_{j=1}^{o} \theta_j \left[\mu \left(a_j \cdot x_i + \gamma_j\right)\right]^{\mathrm{T}} = \hat{y}_j, i = 1, 2, \ldots, N \tag{20}$$

where μ represents the activation function; o represents the number of neurons in the hidden layer; a_j represents the weight vector; γ_j represents the bias vector; represents the inner product of vectors; θ_j represents the weight parameters from the hidden layer to the output layer; N represents the total number of samples.

In summary, this paper selects LightGBM1($\alpha = 0$), LightGBM2($\alpha = 0.4$), LightGBM3($\alpha = 0.8$), LightGBM4($\alpha = 1$), SVM5, a total of 5 models as the base learners, and ELM is used as the meta learner. The proposed LightGBM based hybrid ensemble model framework is shown in Fig. 10.

Performance evaluation

In this section, we first introduce the evaluation criteria, then introduce the breast cancer dataset used, and finally give the diagnosis results of the algorithm in this paper. All experiments were conducted on the Ubuntu system and achieved CUDA acceleration of RTX3090. To improve the reliability of the experiment, 80% of the samples were used as the training set and 20% as the testing set, and all results were reported as the average after 10 independent runs.

Evaluation criteria

The model evaluation metrics used in this article include ten-fold cross validation score, accuracy, precision, recall, etc. The above evaluation indicators are estimated based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Among them, TP represents the number of samples belonging to Class 1 that are also judged as Class 1 by the model; FN represents the number of samples belonging to class 1 that were incorrectly identified as class 0 by the model; FP represents the number of samples belonging to class 0 that were

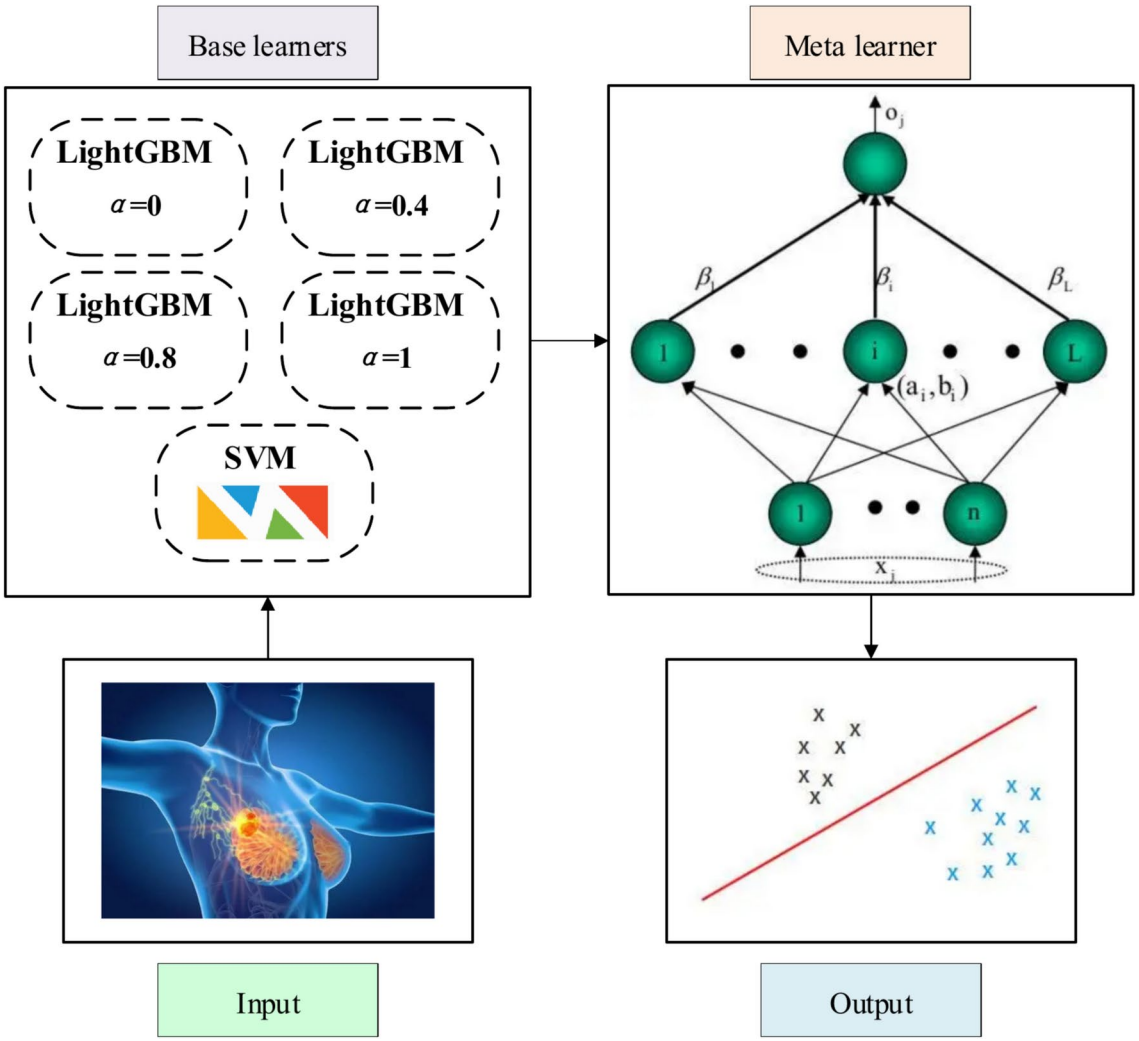

Fig. 10. Hybrid integrated model framework for LightGBM.

incorrectly identified as class 1 by the model; The number of samples belonging to class 0 represented by TN is also determined by the model as the number of samples belonging to class 0.

Accuracy is the most commonly used metric for evaluating models, which represents the proportion of correctly predicted samples (including positive and negative categories) to all samples. The equation for calculating accuracy is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

Precision, also known as precision, refers to the proportion of correctly predicted positive classes among all predicted positive classes. Accuracy refers to the prediction results, which represents how many of the predicted positive class samples are truly positive class samples. The equation for calculating accuracy is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

Recall, also known as completeness rate, refers to the proportion of correctly predicted positive samples to the total actual positive samples. The recall rate is specific to the original sample and represents how many positive class samples in the entire sample were correctly predicted. The calculation equation is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

Dataset

This paper uses the Wisconsin breast cancer diagnosis public dataset[23] on the UCI website, and analyzes the diagnostic data of 569 patients in total. The sample size of the data set is shown in Fig. 11. The data consists of 32 columns, with two columns used to represent the patient's ID and labels indicating whether the patient's tumor is benign (0) or malignant (1). The remaining 30 columns describe the maximum, average, and minimum values of the 10 features of the patient's cell nucleus. The average value was selected as the experimental data for this experiment.

Results and comparisons

Performance analysis of data imbalance methods

In order to compare the effectiveness of the methods proposed in this article for solving data imbalance, several commonly used comparison algorithms were selected, including (1) the method proposed in this article, which introduces gradient harmonic loss and cross entropy loss to enhance the model's attention to minority classes in the dataset and alleviate the impact of data imbalance on diagnostic results. (2) Using Focal loss[35] improved F-LightGBM, referring to the selection of hyperparameters in the paper, $gamma = 2.5$, $alpha = 0.75$. (3) The commonly used oversampling technique for handling imbalanced data, Borderline Synthetic Minority Oversampling Technique (Borderline SMOTE)[36], is an improved BS-LightGBM, in which oversampling uses the maximum number of samples to sample the original data. (4) U-LightGBM, an improved under-sampling technique commonly used for handling imbalanced data, uses under-sampling to sample the raw data with the minimum number of samples. (5) In the proposed method, the removal of gradient harmonic loss LightGBM/G did not address the issue of data imbalance. The experimental results are shown in Fig. 12.

The analysis shows that the proposed method has the best performance in the diagnosis of breast cancer, and is superior to the commonly used under sampling, over sampling, and other data imbalance processing strategies.

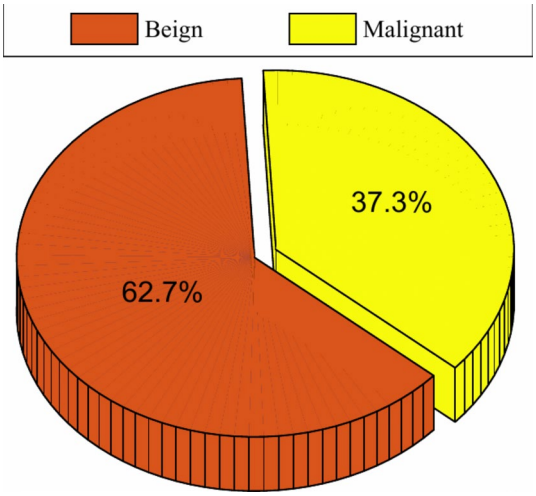

Fig. 11. Breast cancer data distribution.

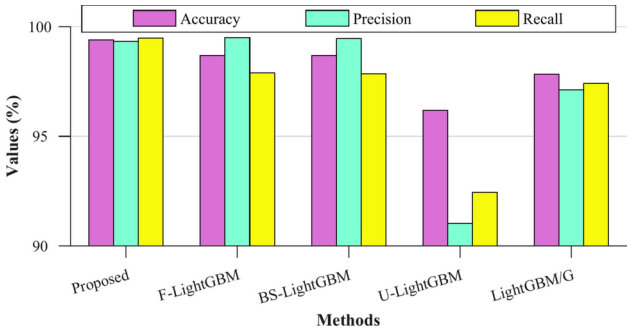

Fig. 12. Diagnostic results of breast cancer with different unbalanced methods.

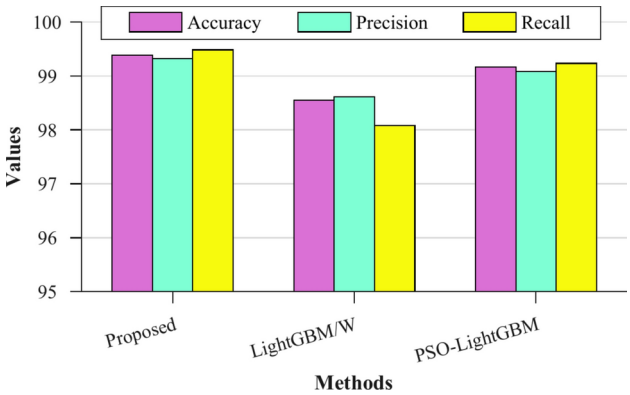

Fig. 13. Diagnostic results of breast cancer with different parameters optimization methods.

Although oversampling techniques can solve the problem of sample imbalance, the overall performance is not as good as the algorithm and F-LightGBM proposed in this paper, indicating that gradient harmonic loss and Focal loss can increase the model's attention to difficult to distinguish samples. However, the gradient harmonic loss function proposed in this paper is more effective in alleviating data imbalance. The under-sampling technique can solve the problem of decreased accuracy caused by imbalanced samples, but under-sampling destroys the integrity of the samples, making it impossible for the model to fully recognize other data, resulting in a decrease in all indicators. It is necessary to adopt an unbalanced processing strategy to further optimize the LightGBM/G model. The above results indicate that the proposed algorithm has superior performance in handling data imbalance.

Performance analysis of parameters optimization methods

In order to compare the effectiveness of the WOA LightGBM method proposed in this paper, this paper selects several commonly used comparison algorithms for comparison, including LightGBM/W, which directly diagnoses breast cancer without optimization algorithm, and the PSO LightGBM model, which is optimized by the commonly used particle swarm optimization algorithm. The experimental results are shown in Fig. 13.

It can be seen from the analysis that the above models can accurately diagnose breast cancer. The LightGBM model without parameter optimization has a slightly worse diagnostic performance than the proposed algorithm and PSO optimized diagnostic model due to the random parameter selection. However, using the WOA algorithm designed in this article, there is a relatively good improvement in the effectiveness of the diagnostic model.

Performance analysis of noise elimination methods

In order to compare the effectiveness of the Jacobian regularization method proposed in this paper, the LightGBM/R method without noise cancellation was selected for comparison. The experimental results are shown in Fig. 14.

It can be seen from the analysis that in terms of breast cancer diagnosis, the LightGBM model after Jacobian regularization has relatively improved its effect. This is because in the process of gradual iteration, the deviation of the model will continue to decrease, which will lead to the model being sensitive to noise. The proposed Jacobian regularization can better reduce noise and improve the performance of the LightGBM model for breast cancer diagnosis.

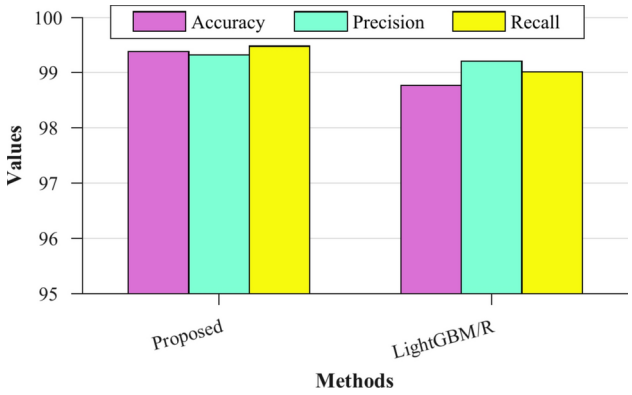

Fig. 14. Diagnosis results of breast cancer with or without noise elimination method.

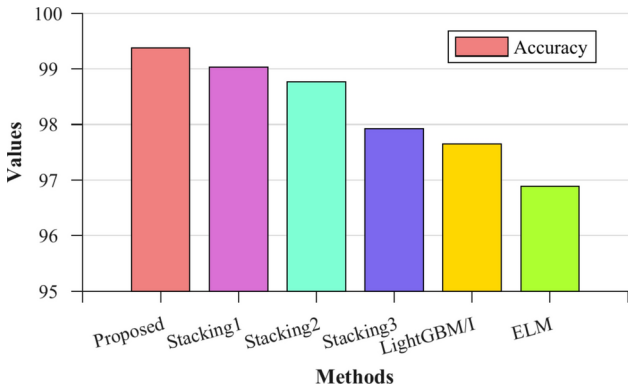

Fig. 15. Diagnostic results of breast cancer with different integrated methods.

Performance analysis of integrated methods

In order to verify the superiority of the LightGBM hybrid ensemble model, this paper uses accuracy as the evaluation index and selects Stacking1, Stacking2, Stacking3, ELM, and the unintegrated LightGBM model (LightGBM/I) for comparative analysis. The Stacking algorithm uses ELM as the meta learner, Stacking1 uses LightGBM, CNN, and SVM as base learners, Stacking2 uses LightGBM and CNN as base learners, and Stacking3 uses LightGBM and SVM as base learners. The experimental results are shown in Fig. 15.

Analysis shows that Stacking1, Stacking2, and Stacking3 are currently the most accurate ensemble methods. From the results, it can be seen that the proposed method, Stacking1, Stacking2, Stacking3 and other recognition models in an integrated manner can integrate the advantages of each basic learner. Compared with a single model, the recognition accuracy is higher. And the method proposed in this article has the highest recognition performance, surpassing traditional integration methods.

Performance comparison of different models

In order to verify the performance of different comparison models, this paper compares the proposed algorithm with diagnostic models such as ICA-MLP[37], SVM-RFE[38], CAD[39], SELF[40], Sun et al.[41], etc. The experimental results are shown in Fig. 16 and Table 2.

It can be seen from the analysis that there are some differences in the diagnosis results of different diagnostic models for breast cancer, but the performance advantages of the methods proposed in this paper are significantly improved after integrating different targeted strategies. Meanwhile, the methods proposed in this paper also achieved the better diagnosis results with low complexity at high speed.

Ablation study

In this paper, gradient harmonic loss and cross entropy loss are introduced, whale optimization algorithm, Jacobian regularization and integrated learning are used to modify the LightGBM model (recorded as K1–K4 respectively), so as to improve the diagnosis effect of breast cancer. To verify the effectiveness of each method, each module was removed separately and its impact was analyzed. The results are shown in Fig. 17.

Analysis shows that the gradient harmonic loss and cross entropy loss proposed in this article, as well as whale optimization algorithm, Jacobian regularization, ensemble learning, and other methods, have shown varying degrees of performance improvement for the LightGBM diagnostic model. In particular, ensemble learning methods have shown significant improvement in the performance of the proposed algorithm. To sum

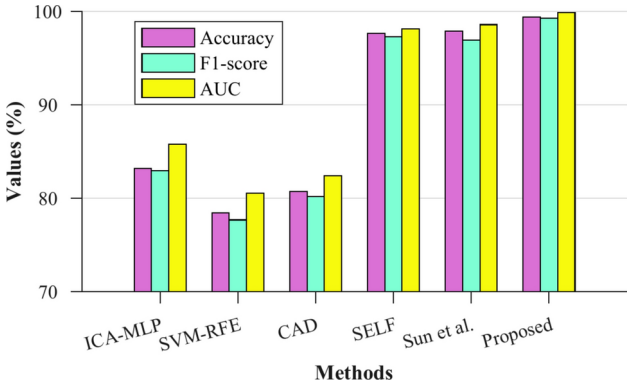

Fig. 16. Diagnostic results of breast cancer in different models.

References	Accuracy	Runtime	Paired *t*-test
ICA-MLP	83.18	302	0.0069
SVM-RFE	78.45	21	0.0018
CAD	80.74	389	0.0043
SELF	97.63	460	0.0315
Sun et al.	97.86	280	0.0924
Proposed	99.38	278	–

Table 2. Diagnostic results of comparisons.

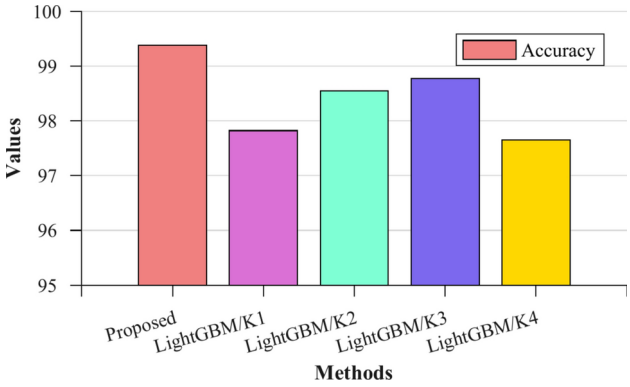

Fig. 17. Diagnosis of breast cancer by ablation study.

up, several improved methods added in this paper are effective and can greatly improve the accuracy of breast cancer diagnosis results.

Conclusion

In order to improve the accuracy of breast cancer diagnosis, this paper proposes an improved LightGBM hybrid integration model combining gradient harmonization and Jacobian regularization for breast cancer diagnosis. Through comparative experiments with other diagnostic models, the results show that the diagnostic effect of this model has been greatly improved, and the problems such as imbalanced data, random parameters selection, noise sensitivity, and poor diagnosis results in breast cancer diagnosis have been solved. It can provide valuable information for prediction and diagnosis of breast cancer, and help doctors to make auxiliary reference for medical diagnosis. Although the effectiveness of the proposed method has been validated on the UCI Wisconsin dataset, future research will focus on experimental validation on more complex data or clinical data due to issues such as small dataset size and insufficient diversity.

Availability of data and materials

The datasets analyzed during the current study are available in the UCI repository: http://archive.ics.uci.edu/dataset/17/breast + cancer + wisconsin + diagnostic. Example from: https://doi.org/https://doi.org/10.1117/12.148698.

Received: 15 October 2024; Accepted: 7 January 2025
Published online: 20 January 2025

References

1. Wang, J. & Wu, S. G. Breast cancer: An overview of current therapeutic strategies, challenge, and perspectives. Breast Cancer Targets Therapy 721–730 (2023).
2. Bikku, T., Ramu, J. & Sekhar, J. C., et al. Optimizing gene expression analysis using clustering algorithms. In *International Conference on Computer & Communication Technologies*, 163–171 (Springer, 2023).
3. Turner, N. C. et al. Capivasertib in hormone receptor–positive advanced breast cancer. *N. Engl. J. Med.* **388**(22), 2058–2070 (2023).
4. Luo, L., Wang, X., Lin, Y., et al. Deep learning in breast cancer imaging: A decade of progress and future directions. IEEE Rev. Biomed. Eng. (2024).
5. Thompson, J. L. & Wright, G. P. Contemporary approaches to the axilla in breast cancer. *Am. J. Surg.* **225**(3), 583–587 (2023).
6. Caswell-Jin, J. L. et al. Analysis of breast cancer mortality in the US—1975 to 2019. *JAMA* **331**(3), 233–241 (2024).
7. An, J., Kwon, H. & Kim, Y. J. The firmicutes/bacteroidetes ratio as a risk factor of breast cancer. *J. Clin. Med.* **12**(6), 2216 (2023).
8. Rahman, M. M. et al. Breast cancer detection and localizing the mass area using deep learning. *Big Data Cognit. Comput.* **8**(7), 80 (2024).
9. Bikku, T. Multi-layered deep learning perceptron approach for health risk prediction. *J. Big Data* **7**(1), 50 (2020).
10. Chugh, G., Kumar, S. & Singh, N. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cognit. Comput.* **13**(6), 1451–1470 (2021).
11. Bikku, T. Fuzzy associated trust-based data security in cloud computing by mining user behaviour. *Int. J. Adv. Intell. Paradigms* **25**(3–4), 382–397 (2023).
12. Singamaneni, K. K., Budati, A. K. & Bikku, T. An efficient Q-KPABE framework to enhance cloud-based IoT security and privacy. Wirel. Pers. Commun. (2024). https://doi.org/10.1007/s11277-024-10908-8.
13. Bikku, T. et al. Enhancing real-time malware analysis with quantum neural networks. *J. Intell. Syst. Internet Things* **12**(1), 57–67 (2024).
14. Praveen, S. P. et al. Enhanced intrusion detection using stacked FT-transformer architecture. J. Cybersecur. Inform. Manag. **13**(2), 19–29 (2024).
15. Singh, A., Tiwari, V. & Tentu, A. N. Ceiling improvement on breast cancer prediction accuracy using unary KNN and binary LightGBM stacked ensemble learning. In *Proceedings of the Seventh International Conference on Mathematics and Computing: ICMC 2021*, 451–471 (Springer, 2022).
16. Khourdifi, Y. & Bahaj, M. Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 1–5 (IEEE, 2018).
17. Sara, A., Heyam, A., Alghunaim, S. & Al-Baity, H. H. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access* **7**, 91535–91546 (2019).
18. Huang, Z. & Chen, D. A breast cancer diagnosis method based on VIM feature selection and hierarchical clustering random forest algorithm. *IEEE Access* **10**, 3284–3293 (2021).
19. Zheng, B., Yoon, S. W. & Lam, S. S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst. Appl.* **41**(4), 1476–1482 (2014).
20. Dhahri, H. et al. Automated breast cancer diagnosis based on machine learning algorithms. *J. Healthc. Eng.* **2019**(1), 4253641 (2019).
21. Bikku, T. et al. Improved quantum algorithm: A crucial stepping stone in quantum-powered drug discovery. J. Electron. Mater. (2024). https://doi.org/10.1007/s11664-024-11275-7
22. Lin, Y. et al. Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes* **11**(8), 888 (2020).
23. Abdar, M. et al. A new nested ensemble technique for automated diagnosis of breast cancer. *Pattern Recognit. Lett.* **132**, 123–131 (2020).
24. Zhang, C. X. & Zhang, J. S. A local boosting algorithm for solving classification problems. *Comput. Stat. Data Anal.* **52**(4), 1928–1941 (2008).
25. Diao, R. et al. Feature selection inspired classifier ensemble reduction. *IEEE Trans. Cybern.* **44**(8), 1259–1268 (2013).
26. Yu, Z. et al. Hybrid incremental ensemble learning for noisy real-world data classification. *IEEE Trans. Cybern.* **49**(2), 403–416 (2017).
27. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
28. Liang, W. et al. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics* **8**(5), 765 (2020).
29. Gupta, E. & Saxena, A. Robust generation control strategy based on grey wolf optimizer. *J. Electr. Syst.* **11**(2), 174–188 (2015).
30. Saputra, R. H. & Prasetyo, B. Improve the accuracy of c4. 5 algorithm using particle swarm optimization (pso) feature selection and bagging technique in breast cancer diagnosis. *J. Soft Comput. Explor.* **1**(1), 47–55 (2020).
31. Alnowibet, K. A. et al. Development and applications of augmented whale optimization algorithm. *Mathematics* **10**(12), 2076 (2022).
32. Naimi, A. I. & Balzer, L. B. Stacked generalization: An introduction to super learning. *Eur. J. Epidemiol.* **33**, 459–464 (2018).
33. Sahu, Y. et al. A CNN-SVM based computer aided diagnosis of breast cancer using histogram K-means segmentation technique. *Multimed. Tools Appl.* **82**(9), 14055–14075 (2023).
34. Junyue, C. et al. Breast cancer diagnosis using hybrid AlexNet-ELM and chimp optimization algorithm evolved by Nelder-mead simplex approach. *Biomed. Signal Process. Control* **85**, 105053 (2023).
35. Yeung, M. et al. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graphics* **95**, 102026 (2022).
36. Omotehinwa, T. O., Oyewola, D. O. & Dada, E. G. A light gradient-boosting machine algorithm with tree-structured parzen estimator for breast cancer diagnosis. *Healthc. Analyt.* **4**, 100218 (2023).
37. Rezaeipanah, A. et al. Design of ensemble classifier model based on MLP neural network for breast cancer diagnosis. *Intel. Artif.* **24**(67), 147–156 (2021).
38. Liu, X. & Tang, J. Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method. *IEEE Syst. J.* **8**(3), 910–920 (2013).
39. Rizzi, M. et al. Health care improvement: Comparative analysis of two CAD systems in mammographic screening. *IEEE Trans. Syst. Man Cybern.-Part A Syst. Hum.* **42**(6), 1385–1395 (2012).

40. Jakhar, A. K., Gupta, A. & Singh, M. SELF: A stacked-based ensemble learning framework for breast cancer classification. *Evolut. Intell.* **17**(3), 1341–1356 (2024).
41. Sun, X. & Qourbani, A. Combining ensemble classification and integrated filter-evolutionary search for breast cancer diagnosis. *J. Cancer Res. Clin. Oncol.* **149**(12), 10753–10769 (2023).

Author contributions
All authors accept public responsibility for the content of the work and have an equal contribution to the study.

Funding
The research leading to these results has received no specific grant.

Declarations

Competing interests
The authors declare no competing interests.

Consent for publication
Freely and informed consent was obtained from all authors to participate in the study.

Additional information
Correspondence and requests for materials should be addressed to X.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.