Discover Artificial Intelligence

RESEARCH Open Access

An effective ECOLASSO with black widow optimization for feature selection and stagewise adaptive learning rate for disease prediction

G. Vijaya[1], G. Sathish Kumar[2], G. Uma Maheshwari[3], M. Karthiga[4], S. Hemkiran[5], Seyed Jalaleddin Mousavirad[6*] and Ghanshyam G. Tejani[7,8*]

*Correspondence:
Seyed Jalaleddin Mousavirad
Seyedjalaleddin.mousavirad@miun.se
Ghanshyam G. Tejani
p.shyam23@gmail.com

Full list of author information is available at the end of the article

Abstract

Machine learning techniques are utilized for early detection of diseases, which can significantly enhance probabilities of positive treatment and existence. The traditional machine learning algorithms may be unable to predict outcomes with sufficient accuracy. In this work, an Effective ECOLASSO with Black Widow Optimization for Feature Selection and Stagewise Adaptive Learning Rate (ELBWOSALR) classifier is proposed for feature selection and prediction. The proposed work comprises two phases, in the first phase, Ecological similarity Least Absolute Shrinkage and Selection Operator (ECOLASSO) model is utilized to predict the best features from the dataset by removing the feature with smallest absolute regression coefficient from the feature set. A Black Widow Optimizer (BWO) is used to choose the subset of optimal features and to reduce local optima. In the second phase, Stagewise Adaptive Learning Rate (SALR) involves combining several weak learner classifiers into a strong ensemble classifier by adaptive learning rate. The key contribution of this work is the *integration of* ECOLASSO model *with BWO* for robust feature selection, combined with a SALR classifier. This hybridization addresses two critical challenges simultaneously: (i) ECOLASSO ensures sparsity and ecological similarity-driven selection of relevant features, while (ii) BWO prevents premature convergence and enhances global search efficiency. By coupling these with SALR, our model achieves superior accuracy and generalization compared to conventional classifiers. Lung cancer, breast cancer and heart disease datasets are used for experimentation. The ELBWOSALR classifier is compared with various classifier models such as Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Logistic Regression, Extreme Gradient Boost Classifier, Gradient Boosting Classifier, K-Nearest Neighbors Classifier and CatBoost Classifier and the results are observed. The proposed ELBWOSALR classifier achieves accuracies of 98%, 97% and 91% with AUC values of 92%, 99% and 94% for lung cancer, breast cancer and heart disease datasets respectively.

Keywords Black widow optimization, Classifiers, ECOLASSO, Stag wise adaptive learning rate, Feature extraction, Penalty

1 Introduction

Accurate disease prediction remains a critical challenge in biomedical research, as early diagnosis can significantly improve patient survival rates and reduce healthcare costs. However, medical datasets such as those for cancer, heart disease, and other chronic conditions are inherently high-dimensional, often containing thousands of features relative to a limited number of patient records. This imbalance leads to problems such as redundant or irrelevant features, noisy attributes, and overfitting in predictive models. Traditional machine learning classifiers, while effective on smaller and cleaner datasets, struggle to maintain robustness and accuracy in such complex data environments. Inadequate feature selection and premature convergence of existing optimization algorithms further reduce diagnostic accuracy and reliability.

The leading cause of mortalities globally is due to lung cancer, breast cancer and heart disease. Unfortunately, many cases are not discovered until the disease has progressed enough, despite the fact that early diagnosis and treatment can significantly increase the likelihood of survival. Based on patient features and risk factors, machine learning (ML) algorithms have shown potential in predicting the diseases, which can help in earlier discovery and treatment. In this situation, employing ML models to analyse data from lung cancer, breast cancer and heart disease datasets can offer useful insights into determining the key risk variables and estimating the likelihood that a certain person will develop the disease [1].

SVC is a binary classification technique that identifies a hyperplane separating data into the two classes in the best possible way. It operates by maximising the distance between each class's closest data points and hyperplane. DTC is a supervised learning method that is mainly used in classification tasks. A decision tree is applied to the training set and each node corresponds to a feature, each path corresponds to a set of rules and each leaf node corresponds to a label. In logistic regression, a binary result is estimated based on a logistic function, which is a classification process. Overfitting and underfitting issues in the disease prediction is removed by Deep Hyper optimization (DHO). It also minimizes the time that passes in the process of execution. It attains 98 percent accuracy with a lower error rate [2].

RFC is an ensemble learning (EL) technique that trains many decision trees and returns a classification that is mean of classification or mean prediction of trees [3]. XGB Classifier is an EL method based on a gradient boosting framework. A set of decision trees is constructed until the required number of trees is reached, and each succeeding tree is used to fix mistakes of earlier trees [4]. K-NC is one specific type of instance-based learning classification algorithm. To predict the label of a test point it finds k-closest training points in feature space and it takes labels as a guide [5]. A GBC is an EL algorithm that builds an additive forward-stage-wise model. To generate a good learner, it averages weak learners by optimising a differentiable loss function [6]. The CBC is an open-source and free toolkit of gradient boosting, designed to provide state-of-the-art solutions to various problems, such as classification, regression, and ranking. It makes use of decision trees with gradient boosting and categorical feature support [7].

All of these algorithms employ ML techniques to create predictive models, and the effectiveness of each one depends on the type of data being utilised and particular issue at hand. Although feature selection and optimization algorithms have been employed to improve model performance, they come with inherent limitations. Classical feature

selection methods fail to consistently capture the most relevant attributes, while optimization-based approaches such as Genetic Algorithms and Particle Swarm Optimization are prone to premature convergence, trapping solutions in local optima. As a result, current models often exhibit limited generalizability and reduced reliability when applied to real-world biomedical datasets. This creates an urgent need for a hybrid algorithm that can simultaneously ensure sparsity, achieve global optimization, and deliver accurate disease prediction across diverse datasets. The disadvantages of the existing approaches are discussed below.

- Large datasets can make SVC computationally expensive, which can result in longer training durations and slower forecasts.
- DTC frequently overfits the training set of data, which might result in subpar generalisation on fresh data.
- Multicollinearity, which occurs when more than one attribute of the input are highly correlated, can also affect logistic regression. This can result in unstable model coefficients and make interpretation challenging.
- For huge data sets, RFC can also be computationally prohibitive, especially when the forescontains a large number of trees.
- Hyper-parameter selection can have a significant impact on XGBC and may necessitate lengthy customization.
- KNN may need to be tuned because it can be sensitive to the number of neighbours it considers, k.
- The GBC method may accomplish poorly on new data as it is prone to overfitting on noisy or strongly correlated data.
- Additionally, CBC might be sensitive to the selection of hyper-parameters, necessitating tuning.

To overcome these limitations, this work introduces the ELBWOSALR classifier algorithm, which integrates ECOLASSO for sparse feature selection, Black Widow Optimization (BWO) for robust global search, and a Stagewise Adaptive Learning Rate (SALR) classifier to achieve accurate and scalable disease prediction.

The key points of proposed work are briefed below.

- Initially ECOLASSO regularization is applied to choose the feature from feature set based on absolute regression coefficient.
- A subset of the most crucial features is found using LASSO in the initial step. To further hone the feature set and boost the model's overall performance, another model, like SV machine, is trained on the chosen features in the second step.
- After selecting the features from the feature set, BWO is applied.
- The proposed ECOLASSO with BWO may also be prone to early convergence and stuck in local optima.
- Stagewise Adaptive Learning Rate (SALR) helps to avoid overfitting and confirm that only the most relevant features are designated.
- Several weak classifiers are combined together and make a stronger classifier using SALR approach.
- SALR uses an adaptive learning rate to improve the classification performance for each sample and reducing the impact of noisy features.

- The proposed ECOLASSO with BWO + SALR is resistant to noise and outliers, which helps lessen the influence of unnecessary characteristics.
- The overall proposed system increases the model accuracy, while reducing susceptibility to local optima and overfitting.

The main contribution of this work is the development of a novel hybrid classifier algorithm, ELBWOSALR, which integrates Ecological LASSO (ECOLASSO) with Black Widow Optimization (BWO) for robust feature selection and employs a Stagewise Adaptive Learning Rate (SALR) classifier for improved prediction. Unlike existing methods, ECOLASSO ensures sparsity by eliminating redundant and irrelevant features, while BWO refines the selected subset through global optimization, preventing premature convergence to local optima. The SALR classifier further enhances generalization by dynamically adjusting learning rates to reduce overfitting. Together, these components form a unified approach that significantly outperforms traditional machine learning models and optimization-based classifiers across lung cancer, breast cancer, and heart disease datasets, achieving higher accuracy, precision, recall, and F1-scores while maintaining lower error rates (MSE = 0.02, RMSE = 0.11). This contribution directly addresses the limitations of prior feature selection and classification approaches, offering a scalable, efficient, and reliable solution for high-dimensional data analysis.

The paper is structured as follows: The literature review is presented in Sect. 2. In Sect. 3, suggested structure for selection of features and classifying the disease is described. The proposed paradigm is demonstrated in Sect. 4 and the comparison with the traditional approaches was visualized. Section 5 gives the conclusion.

2 Literature survey

Literature review regarding lung cancer, breast cancer and heart disease prediction according to ML and DL methods is presented below.

The work by Chintan M et al. projected a new way of diagnosing cancer based on machine learning algorithms. The proposed approach involved collecting data of diverse Internet of Things (IoT) devices like wearable sensors, cell phones, and electronic health records, and subsequently applying ML algorithms to analyse data and detect the presence of lung cancer. The study's discoveries verified that the proposed method had a higher rate of success in detecting lung cancer. The sensitivity of the suggested strategy, as reported by the authors to be 94.57 percent, considerably higher than the accuracy of conventional methods of lung cancer diagnosis. The research also showed that the proposed approach may be helpful in early diagnosis of lung cancer [8].

The ML algorithms and improved imaging techniques may be applied to predict and analyze lung cancer at the early stage. Data from 224 individuals were collected using by Higher-Resolution Computed Tomography (HRCT) images and clinical data including age, tobacco use, and pulmonary tests. It started by deriving features of the HRCT images through image processing methods, e.g. texture analysis and morphological features [9]. Lung nodules and the likelihood of cancer were subsequently forecasted based on diverse machine learning algorithms, including RF models, SVM algorithms, and Artificial Neural Networks (ANN). The results indicated that the approach based on SVM performed better than any other algorithms, indicating that it detected nodules in the lungs with an accuracy of 82.4% and the probability of malignancy with an accuracy of 88.2%, respectively. The ANN also delivered some promising results, identifying the

presence of nodules with 81.7 percent accuracy and malignancy with 87.9 percent accuracy [10].

The study focuses on computed tomographic images to check the efficiency of the artificial intelligence techniques in breast cancer identification [11]. The researchers took a data set consisting of 200 CT images that were divided into 100 cancerous images and 100 non-cancerous images. They used KNN, SVM, DT, RF, and GBM and compared their performance. The photos were pre-processed to normalise the intensities and eliminate noise. Features were extracted from the images using techniques such as Local Binary Pattern, Grey Level cooccurrence Matrix and Histogram of Oriented Gradients. The data indicated that the SVM algorithm was better compared to the other algorithms with an accuracy of 98%. The DT method's efficiency was 84%, which was the most inaccurate. The researchers established that SVM is the most appropriate method to detect lung cancer using CT scans. This is because ML algorithms could be effectively used to accomplish it [12].

In an article by Pradhan et al., an effective multi-level cancer forecasting system based on a medical IoT is suggested. A total of 1000 patient records were used as the size of the dataset, 500 of which belonged to people with lung cancer and 500 of which belonged to healthy individuals [13]. The processing of the data involved data cleaning, feature selection, and normalisation. The SVM classification yielded an accuracy rate of 97.1 percent, sensitivity rate of 97.6 percent, specificity rate of 96.6%, precision rate of 97.4%, and F1-score of 97.5%. The recommended SVM model's capability was also assessed in comparison to that of decision trees, KNN, and naive Bayes. The results exhibit that the SVM model prevailed over these methods not only regarding accuracy but also other measurements [14].

In the study of Shanthi et al. [15], the objective was to predict the cardiovascular disease using various feature selection algorithms [15]. The study was done using the data of 32 variables related to lung cancer detection offered by University of California, Irvine (UCI) ML repository. The study applied three ML models to generate a lung cancer prediction model, including DT, NB, and RF. The sensitivity, specificity, and accuracy of Random Forest model were 91.94 percent, 96.47 percent, and 85.29 percent, respectively. The research established that artificial intelligence algorithms may be helpful in uncovering the cancer in lungs, and that the method referred to as Random Forests may be more accurate in this detection than other algorithms [16, 17].

The SVM and one-dimension CNN is applied to classify the non-small cell lung carcinoma and to locate the multi-level lung cancer. The research is based on the set of 1076 images of computed tomography (CT) scans of patients with NSCLC. Frequency difference factor is provided to compensate the term and document frequency. The co-vector based feature selection algorithm demonstrates the 4.58% of the better accuracy compared with state-of-the-art algorithms [18]. The 1D-CNN models used by the authors have three convolutional layers followed by a max pooling layer and a fully connected layer. The training data was utilized to train the model and testing data to test it. Results disclosed that 1D-CNN model attained an F1 score of 97.2% and overall accuracy of 97.2%, precision of 98.3%, recall of 96.1% [19].

The goal of Botlagunta et al.'s objective was to use the classification model and machine learning methods to form a risk estimation model for breast cancer. The dataset consisted of 739 patient records that were used in this investigation and contained

data such as age, gender, smoking history, and other symptoms [20]. The authors found the most important features in the dataset with help of various ML algorithms, including LR, SVM, and feature selection algorithms. RF algorithm with the selected features showed better performance with accuracy of 88.98%, sensitivity of 93.3%, specificity of 82.3%, and the AUC of 0.923. The paper concluded that the developed risk forecasting algorithm could be used in early detection of lung [21]. A hybrid feature selection framework that combines Game-kernel SHAP with binary Social Ski-Driver optimization and local search algorithms for RNA-seq cancer classification [22]. A three-phase hybrid approach using soft computing practices for cancer classification from gene expression microarray data. The framework integrates feature selection, dimensionality reduction, and classification [23].

Abdullah et al. provide the correlation selection method to forecast the probability of lung cancer by using dataset about gender, age, tobacco use, tumour size, and other clinical and demographic data of patients with cancer of the lungs [24]. Some of the ensemble methods that were discussed include the random forest, stochastic boosting, and bagging methods. Utilising metrics like accuracy, sensitivity, specificity, and AUC, the algorithms' outcomes were assessed [25].

Artificial intelligence algorithms are applied to predict lung cancer diagnosis with the help of novel LASSO prognostic model. The dataset that was used in the study consisted of patient records of 5000 patients and contained information such as age, sex, smoking history, and CT scan images. The authors have trained 6 ML algorithms, including LR, RF, Decision Tree, gradient boost, XGBoost and LightGBM, on a range of processing techniques, including image normalisation and feature scaling [26, 27]. Some of the criteria that were considered in the study to determine performance of algorithms included accuracy, precision, recall and F1-score. RF had an accuracy of 97.8%, followed by XGBoost whose efficiency was 96.6%, then the other two. The paper has concluded that ML algorithms were effective in identifying diseases earlier.

The backpropagation neural network and ultrasound images with the multi-fractal dimensions are applied to automate characterisation of breast cancer data. This feature helps to detect breast cancer at an early stage [28]. An improved Genetic Algorithm (GA)-based clustering method with a novel selection strategy tailored for categorical dental data. The approach enhances prediction transparency and interpretability while improving clustering accuracy compared to traditional GA-based methods [29]. X-ray imaging along with the machine learning models helps to categorize the disease as mild, moderate, severe and critical [30]. The ML algorithms like SVM, RF, LR, DT (C4.5) and KNN are applied to classify disease. It is observed that SVM outperforms classifier models with highest accuracy of 97.2% [31, 32]. The DL approach for lung cancer prediction and weighted particle swarm optimization, smooth SVM, multi modal emotion appreciation using adaptive normalization for abnormality detection and overfitting avoidance plays a major role in the disease prediction [33, 34]. Optimal Feature Evaluation and Selection (OFES) algorithm is applied to evaluate and find higher quality features for multiclass classification problems. The OFES uses the quantitative means for selecting higher quality features. It improves classification accuracy up to 95% [35].

The Improved BWO algorithm applied to electroencephalogram signals for pattern recognition and prediction. This BWO achieves the classification accuracy of 5.21%, 4.16% and 1.47% higher values than the traditional BWO algorithms [36]. Machine

learning solutions are widely used to understand and reduce the symptoms of highly fatal diseases. ML spark libraries and ChiSqSelector were applied for the feature selection and validation. Here RF obtains highest classification accuracy with 98.7% [37]. Combining multiple risk factors during modelling enables the early predicting and diagnosis of the disease. Here, random forest classifier attains the highest accuracy about 80% with sensitivity of 95% [38]. Pearson's correlation technique is applied to eliminate redundant features in feature selection process. The hybrid dataset called Sathvi is created with 12 attributes and 512 instances for the cardiovascular disease prediction. The classifier models achieve 98.11% of accuracy with the mean of 94.34% [39].

Based on the reviewed literature, the research gap is identified and summarized as follows: The existing machine learning and ensemble classifiers have shown promise in disease prediction, they often suffer from two key limitations: (i) redundant or irrelevant features that degrade classification accuracy, and (ii) metaheuristic optimizers that tend to converge prematurely to local optima. Current optimization-based feature selection methods such as Genetic Algorithms, Particle Swarm Optimization, Whale Optimization Algorithm and Grey Wolf Optimizer do not fully overcome these issues because these algorithms suffer from premature convergence and get trapped in local optima. Traditional ensemble classifiers and boosting techniques are prone to overfitting when noisy or redundant features are present. There is a lack of adaptive learning rate mechanisms that can dynamically adjust model training and reduce sensitivity to noise. To bridge this gap, we propose a novel hybrid algorithm where ECOLASSO is combined with Black Widow Optimization to ensure robust, sparse, and globally optimized feature selection, followed by Stagewise Adaptive Learning Rate for classification, thereby addressing both feature selection and classification challenges in an integrated manner. The comparative analysis of existing models is discussed in Table 1.

From the literature review, it is found that the existing methods are being applied only to limited datasets. To overcome dataset limitations, Data augmentation involves in creating new samples by using various transformations to existing data samples. This technique can boost the size of the dataset and upgrade the performance of the model. Transfer learning is a technique that allows the use of pre-learned models on larger datasets to advance the performance of models trained on smaller datasets.

3 Proposed work

Three benchmark datasets such as lung cancer, breast cancer and heart disease datasets are obtained from the publicly available UCI machine learning repository and are used for experimentation and analysis. The lung cancer dataset contains 16 attributes and 284 instances. The breast cancer dataset contains 32 attributes with 569 instances and heart disease dataset contains 75 attributes with 303 instances. The data augmentation is applied to increase the number of instances in dataset to prevent overfitting. The selective features play a more important role than using all selected features for accurate classification.

The suggested work comprises two phases, in the first phase, Ecological similarity Least Absolute Shrinkage and Selection Operator (ECOLASSO) model is utilized to predict the best features from the dataset by removing the feature with the smallest absolute regression coefficient from the feature set. To reduce local optima, a BWO is used. It is used to choose a subset of optimal features. In the second phase, Stagewise Adaptive

Table 1 Comparative analysis of existing models or techniques

S. No	Author(s) and year	Methodology used	Advantages	Disadvantages	Accuracy
1	Pradhan et al., [13]	IoT-based ML for lung cancer	High success rate (94.57% accuracy)	Requires access to IoT devices and health records	94.57%
2	Dhivya and Bazilabanu, [2]	SVM for cancer prediction	High accuracy and sensitivity (97.1% and 97.6%)	Requires large datasets and preprocessing steps	97.1%
3	Botlagunta et al., [20]	Random forest for breast cancer	High accuracy and AUC value (88.98% and 0.923)	Relies on accurate patient records and symptom data	88.98%
4	Abdullah et al., [24]	Correlation Selection for lung cancer	Utilizes ensemble methods for prediction	Performance may vary depending on chosen method	75.3%
5	Chen et al., [26]	LASSO prognostic model for lung cancer	Utilizes multiple ML algorithms and processing	Performance may vary depending on dataset quality	89.6%
6	Mohammed et al., [28]	Multi-fractal dimensions for breast cancer	Automates characterization of breast cancer data	Requires high-quality ultrasound images	93.2%
7	Naji et al., [32]	ML for disease classification	Utilizes various ML algorithms for classification	Performance may vary depending on dataset and algorithm selection	86.7%
8	Rabiei et al., [38]	Weighted PSO and SVM for abnormality detection	High accuracy in prediction and detection	Complexity and computational resources required	94.1%
9	Almutairi et al., [39]	OFES algorithm for feature selection	Improves classification accuracy	Performance may vary depending on dataset and parameters	79.8%
10	Huang et al., [36]	Improved BWO for pattern recognition	Achieves higher accuracy than traditional BWO	Complexity and computational resources required	90.5%
11	Schaefer and Atreya, [34]	ML spark and chisqselector for disease prediction	Utilizes ML Spark libraries and feature selection	Performance may vary depending on dataset and parameters	91.2%
12	Garate-Escamila et al., [37]	Combination of risk factors for disease prediction	Enables early prediction and diagnosis	Relies on accurate risk factor assessment and datasets	88.5%
13	Chintan et al., [8]	Pearson's correlation for feature selection	Eliminates redundant features	Performance may vary depending on chosen threshold	83.9%

Learning Rate (SALR) involves combining several weak learner classifiers into a strong ensemble classifier by an adaptive learning rate. The process flow for the proposed method is depicted in Fig. 1.

In Fig. 1 the process begins with data preprocessing, followed by ECOLASSO-based feature selection. Black Widow Optimization (BWO) then refines the selected features using a global search strategy. Finally, classification is performed using the Stagewise Adaptive Learning Rate (SALR) ensemble, and results are evaluated with multiple performance metrics.

3.1 Phase-1: ECOLASSO feature selection

ECOLASSO is a modification of Lasso regularization that includes an additional penalty term based on the ecological similarity between features. ECOLASSO is derived by adding the ecological penalty term to the objective function of Lasso.

Step 1: Define the LASSO objective function.

The Lasso objective function is given in Eq. (1) as follows.

$$A = \min_{w} \|y - X_w\| 22 + \lambda |w| 1 \tag{1}$$

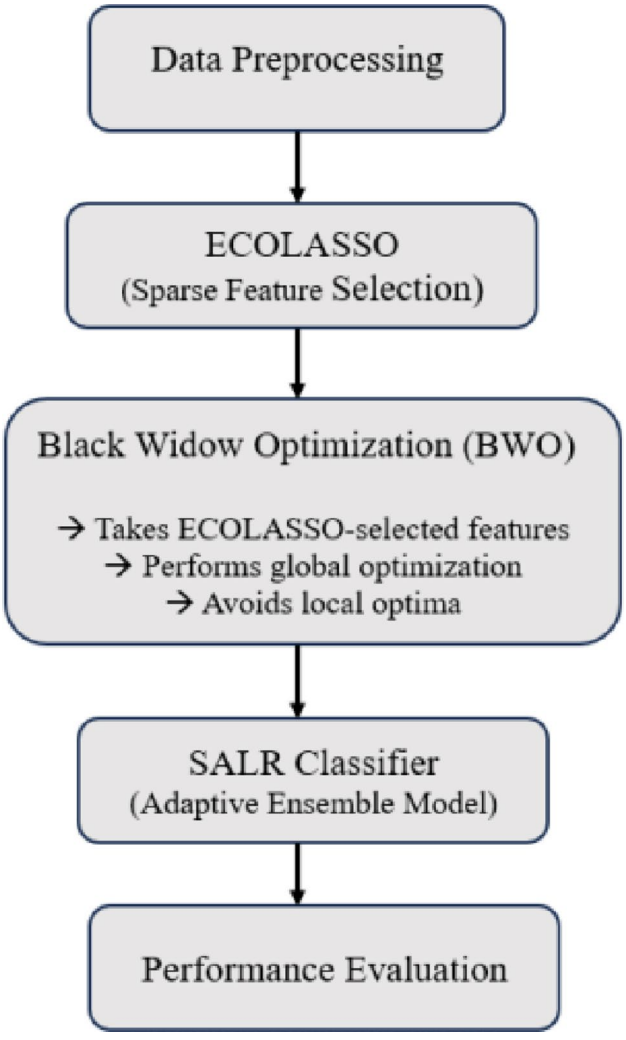

Fig. 1 Process flow of the proposed ELBWOSALR algorithm

with y being the target variable, X the feature matrix, w the weight vector and λ regularization parameter that determines the intensity of the penalty.

Step 2: Introduce the ecological similarity matrix.

In order to include the ecological penalty term, we must specify a matrix, which is a measure of similarity between each pair of features. This matrix is given as S, in which i, j is the similarity between feature i and feature j. S is a PXP square matrix with P number of features.

Step 3: Define the ECOLASSO objective function.

The ECOLASSO objective function is given in Eq. (2).

$$A = \min_{w} \|y - X_w\| \,|22 + \lambda| \,|w| \,1 + \eta \sum i = 1 * P \sum j = 1 * PS_{i,j} \,|w_i| \cdot |w_j| \quad (2)$$

where η is a hyper-parameter that controls strength of the ecological penalty term.

Step 4: Derive the ECOLASSO update rule.

In order to obtain the update rule of ECOLASSO, we need to differentiate the objective function with respect to w, which is given in Eq. (3).

$$A = \partial_w \partial \left(\|y - X_w\| 22 + \lambda \|w\| 1 + \eta \sum i = 1 * P \sum j = 1 * PS_{i,j} |w_i| \cdot |w_j| \right) \quad (3)$$

The first term is the standard Lasso derivative $(F.D)$ given in Eq. (4).

$$F.D = -2X * T (y - X_w) \quad (4)$$

The second term is the derivative $(S.D)$ of the L_1 norm, which is given by Eq. (5).

$$S.D = \lambda * sign (w) \quad (5)$$

The third term is the derivative $(T.D)$ of the ecological penalty term, which is given in Eq. (6).

$$T.D = \eta \sum j = 1 * PS_{i,j} sign (w_j) |w_j| + \eta \sum i = 1 * PS_{i,j} sign (w_i) |w_i| \quad (6)$$

Combining these derivatives, the final update rule for ECOLASSO is generated and given in Eq. (7).

$$w_i \leftarrow soft\, threshold \left(n_1 * X_i * T (y - X_w) + \lambda\, sign\ (w_i) + \eta \sum j = 1 * PS_{i,j}\ sign\ (w_j) |w_j| 1, \lambda \eta PS_{i,j} \right) \quad (7)$$

where X_i is the i^{th} column of X, and the soft threshold function is defined in Eq. (8).

$$\text{Soft } T (x, T) = \left\{ \begin{array}{c|c} x - T & if x > T \\ 0 & if -T \leq x \leq T \\ X + t & if x <= -T \end{array} \right\} \quad (8)$$

The following steps represent the ECOLASSO algorithm.

1. Start with the original feature set and set regularization parameter λ to a small value.
2. Fit LASSO regression model to training data with current value of λ.
3. Remove the feature with the smallest absolute regression coefficient from the feature set.
4. Check if the number of features remaining in feature set is greater than a specified minimum. If not, terminate the algorithm and return the remaining features as the selected set.
5. If the number of features remaining in the feature set is still greater than the specified minimum, repeat steps 2–4 with the updated feature set.
6. Increase the value of λ and repeat steps 2–5 until the desired number of features is selected.
7. Optionally, perform cross-validation to tune the value of λ.

The ECOLASSO algorithm is represented as follows.

```
Step 1. Input:  - X (n X P): the data matrix
      - y (n X 1): the vector of response variables
      - alpha (positive scalar): the regularization parameter for L1 penalty
      - delta (positive scalar): the threshold parameter for L0 penalty
Step 2. Initialize:
      - Set t = 0
      - Set lambda_t = maximum eigenvalue of X'X
      - Set beta_t = 0
      - Set active_set = empty set
Step 3. Repeat until convergence:
      - Set t = t + 1
      - Compute the residual vector r_t = y - X*beta_t
      - For each j = 1,...,p:
            - If j is not in active_set:
                  - Compute the correlation c_j = abs(X_j' * r_t)
                  - If c_j > 2*alpha/(t*lambda_t):
                        - Add j to the active_set
      - If active_set is empty:
            - Stop and output beta_t
      - Else:
            - Compute X_A = the submatrix of X with columns in active_set
            - Compute beta_A = the LASSO estimate of beta on X_A
            - Compute the step size gamma_t = 1/n * ||X_A' * r_t||
            - Compute beta_t+1 = beta_t + gamma_t * (X_A * beta_A - beta_t)
            - Compute lambda_t+1 = (1-delta)*lambda_t
Step 4. Output: beta_t
```

3.2 ECOLASSO with Black Widow Optimizer

ECOLASSO can end up in a local optimum rather than a global optimum. Nevertheless, ECOLASSO is developed in a way that overfitting is avoided, by discouraging large coefficients and preferring a simple model, which can lead to a better solution being found. Also, a stochastic method allows ECOLASSO to search various parts of the solution space and without being stuck in local optima. The BWO ensures this. Black Widow is an algorithm that is optimized to search the solution space to reach the global optimum solution. It applies both local and global search methods so as not to fall into local optima.

To efficiently search a search space, BWO employs predictable as well as stochastic search methods. It performs long-distance exploration with the Levy Flight algorithm, which serves to avoid being trapped in local optima . Also, a local search method is also used by the algorithm to improve the quality of the solution. By using both of these search techniques, BWO can be utilized to explore the search space and locate the optimal resolution without becoming trapped in local optima. In this way, it is possible to say that BWO ensures avoidance of local optima.

Parameter tuning process is applied to the BWO algorithm to improve the transparency and reproducibility. Specifically, the BWO parameters were set as follows: population size = 30, maximum iterations = 100, crossover rate = 0.4, and mutation rate = 0.2. These values were determined through preliminary experiments using grid search, where the population size was varied from 10 to 50, crossover rate from 0.2 to 0.6, and mutation rate from 0.1 to 0.3. The chosen configuration consistently yielded the best trade-off between convergence speed and classification accuracy across all three datasets.

The BWO method is a meta-heuristic approach created to address the aggressive behavior of black widow spiders. It starts with the random generation of a population of initial candidate solutions. The algorithm then updates the positions of all possible solutions based on where the best answer is at each iteration.

The position update for the i^{th} candidate solution at the t^{th} iteration is given by Eq. 9.

$$x_i(t) = x_i(t-1) + c(t) * (pos(t-1) - x_i(t-1)) \tag{9}$$

where $pos(t-1)$ is the location of the current optimal solution at the $(t-1)^{th}$ iteration, $x_i\,(t-1)$ is the position of i^{th} candidate solution at $(t-1)^{th}$ iteration, and $c(t)$ is a parameter that regulates the number of steps. It uses a "widow" phase, in which a smaller subset of candidate solutions known as black widows are chosen based on the fitness and updates the positions of the other candidate solutions. According to the Eq. 10, the i^{th} candidate solution's position update throughout the widow phase.

$$x_i\,(t) = x_i\,(t-1) + r * (bl * W_i - x_i\,(t-1)) \tag{10}$$

where $x_i\,(t-1)$ is location of the i^{th} potential solution at $(t-1)^{th}$ iteration, $bl * W_i$ is the exact location of the chosen black widow, and r is an arbitrary number between 0 and 1.

3.3 Phase-2: Stagewise Adaptive Learning Rate (SALR) Classifier

SALR algorithm incorporates an adaptive learning rate for improved convergence speed and feature selection for better generalization performance. The main difference between SALR and Stagewise Additive Modeling using a Multi-class Exponential loss function with a Ridge penalty is the update equation for the weight distribution of the samples. In SALR, the weight distribution is updated using an adaptive learning rate that depends on the gradient and the previous weight distribution. The SALR algorithm computes the weights and the predictions of the classifiers as follows.

Step 1: Initialize the sample weights as given in Eq. (11). In this w represents the weight and N represents the number of training samples.

$$w_i^{\{(1)\}} = \frac{1}{N} i = 1, 2, ..., N \tag{11}$$

Step 2: Train the classifiers by computing the weighted error ϵt of the classifier as given in Eq. (12).

$$\in t = \sum_{i=1}^{N} wi\,(t) \cdot I\,[ht\,(x_i) \neq y_i] \tag{12}$$

where $I\,[ht\,(x_i) \neq y_i]$ represents an indicator function that takes the value 1 if condition inside brackets is true, and 0 otherwise.

Step 3: Compute the weight α_t assigned to the classifier as given in Eq. (13).

$$\alpha_t = 0.5 * \ln\left(\frac{1- \in t}{\in t}\right) + \beta t \tag{13}$$

where β_t is the adaptive learning rate for the t^{th} iteration as given in Eq. (14)

$$\beta_t = 0.5 * \omega * \sum_{d=1}^{D} \left|w_t^d - w_{t-1}^d\right| \tag{14}$$

where w_t^d is weight assigned to d^{th} feature for t^{th} iteration, and ω is the regularization parameter.

Step 4: Update the sample weights as given in Eq. (15).

$$w_i^{t+1} = \frac{w_i^t}{Z_t} * e^{-\alpha_t Y ht(xi)} \tag{15}$$

where Z_t is the normalization factor that ensures the weights sum to 1 and derived as below in Eq. (16).

$$Z_t = \sum_{i=1}^{N} w_i^t e^{-\alpha_t Y ht(xi)} \tag{16}$$

Step 5: After all iterations are completed, final prediction is made by combining predictions of all weak learners. Specifically, final prediction for a given input x is given in Eq. (17).

$$\hat{y}\,(x) = \arg\max \sum_{t=1}^{T} \alpha_t h_t\,(x_i) \tag{17}$$

where $\widehat{y}\,(\mathrm{x})$ is predicted label for input x, $h_t\,(x_i)$ is output of y^{th} class of weak learner at iteration t, and α_t is weight of the weak learner at iteration t.

The SALR algorithm is presented for classification.

Input:

X: matrix of predictor variables
y: vector of categorical labels
T: number of iterations
base_learner: the type of weak learner to use
learning_rate: the learning rate parameter

Output:

y_pred: the predicted categorical labels

Algorithm:

Step 1. Initialize the sample weights w_i = 1 / n for i = 1 to n, where n is the number of training examples.
Step 2. For t = 1 to T do the following:
a) Fit a weak learner h_t to the training data using the weights w_i
*b) Compute the weighted error e_t of the weak learner as e_t = sum(w_i * (y_i != h_t(x_i))) / sum(w_i)*
c) Compute the coefficient alpha_t as alpha_t = log((1 - e_t) / e_t) + log(k - 1), where k is the number of classes
*d) Update the sample weights w_i as w_i = w_i * exp(-alpha_t * y_i * h_t(x_i))*
e) Normalize the sample weights w_i such that they sum to 1
*Step 3. Compute the final prediction y_pred as y_pred(x) = argmax(sum(alpha_t * h_t(x)))*
Step 4. Final Output: y_pred(x)

The suggested feature selection and classification algorithm is presented in Fig. 2.

First, data have to be entered and the parameters of the algorithm need to be specified, including the number of epochs, the penalty parameter , and the stopping criterion. The data are standardized so that all the features are of the same scale, and this leads to an enhancement inthe algorithm. The data is divided into two groups, one to train the model and the other to measure its outcome. The initial values of model coefficients are set to zero and penalty parameter is set to a fixed value. Compute the gradient, and update the coefficients with LASSO penalty, per epoch: The gradient of the cost function is computed and coefficients are updated using LASSO penalty. The LASSO penalty favors sparsity in the model, that is, it favours the selection of a subset of features that are mostly relevant to the prediction problem. The model is evaluated on the validation set, and performance metrics are tracked to monitor the model's progress. Unless validation performance has enhanced after a particular number of epochs, decrease penalty parameter to encourage the model to choose additional features. The model coefficients

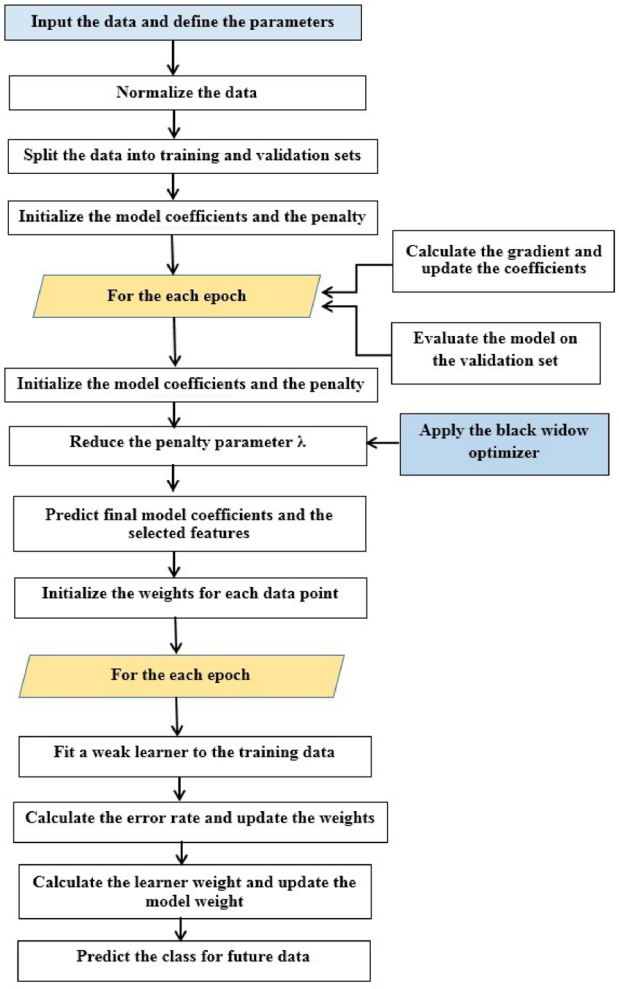

Fig. 2 Working flow of the proposed ELBWOSALR

are printed out as the final model, and the chosen features can be utilized to predict new samples.

Once the features have been selected, set the algorithm parameters including the number of epochs, learning rate and stopping criterion. The initial values of the weights of each data point are set to 1/N, where N is the number of data points. In every epoch, fit a weak learner to the training data, using the weights used to highlight the misclassified points: A weak learner is fitted to the training data and weights applied to highlight the misclassified points. It implies that the algorithm gives more attention to those points which are harder to classify. An error rate is estimated and the weights are updated. The weights of the points that are misclassified are incremented and weights of correctly classified points are decremented. The learner weight is estimated with the error rate and the model weight is updated with this rate. The last model is then produced and its efficiency on the validation set is measured in terms of performance measures like accuracy, precision, recall, and F1-score.

Table 2 Dataset description

Dataset Name	Instances	Attributes	Description
Lung cancer	32	3	Describes 3 categories of pathological lung cancers
Breast cancer	286	9	201 instances of one class, 85 of another. Described by 9 attributes
Heart disease	303	14	Subset of 14 attributes used for experiments. Goal field: presence of heart disease (0 = none, 1–4 = presence)

The suggested classifier is compared to state-of-the-art classifiers such as SVC, DTC, RFC, LR, XGBC, GBC, KNC and CBC. The Table 3 provides basic parameters of the classifiers

Table 3 Classifier models and its parameters

Classifier	Basic parameters
SVC	kernel = 'linear', probability = True
DTC	random_state = 1234
RFC	random_state = 1234
LR	solver = 'liblinear'
XGBC	booster = 'gblinear', learning_rate = 1, n_estimators = 10
GBC	learning_rate = 0.1, subsample = 0.9, max_features = 0.75, loss = 'deviance', n_estimators = 100
K-NN	n_neighbors = 16
CBC	iterations = 30, learning_rate = 0.1
ELBWOSALR	ECOLASSO + BWO + SALR

The outcomes of classifiers were evaluated using standard performance metrics

4 Results and discussion

This part provides the results of the implementation of the suggested ELBWOSALR classifier. On lung cancer, breast cancer, and heart disease datasets, the proposed classifier is used to perform analysis. Table 2 shows the dataset description.

Precision is a measurement of the number of correctly predicted positive instances divided by the number of all instances that have been predicted as positive. The Eq. (18) represents the accuracy and Eq. (19) shows the precision formula where TP is the number of true positives, TN is number of true negatives, FN is number of false negatives and FP is number of false positives.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{18}$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{19}$$

Recall metric is defined as the ratio of the correct number of predicted positive cases divided by the total number of actual positive cases as represented in Eq. (20).

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{20}$$

F1-score is the weighted mean of precision and recall which achieves the best value of one and worst score of zero and is expressed in Eq. (21).

$$F1 - \text{score} = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision} + \text{recall})} \tag{21}$$

A ROC curve is a graph of true positive rate (TPR) versus false positive rate (FPR) at all values of the threshold. The AUC curve measures the cumulative performance of the

model with an AUC of 1 and 0.5 representing perfect classification and random classification respectively in Eq. (22) and (23).

$$TPR = \frac{TP}{(TP + FN)} \tag{22}$$

$$FPR = \frac{FP}{(FP + TN)} \tag{23}$$

The Mean Squared Error (MSE) in Eq. (24) denotes the effectiveness of the machine learning model. The original and predicted values are considered for finding the average squared deviation. The smaller MSE shows a better fit to the model. In which y_i represents true value of the resultant variable of the i^{th} observation and $\widehat{y}_i$ specifies the estimated value of the resultant variable of the i^{th} observation.

$$MSE = \frac{1}{N} * \sum_{i=1}^{n} (yi - \hat{y}i)^2 \tag{24}$$

Root Mean Squared Error (RMSE) is the square root of MSE and confers the average distance between the original and predicted values. RMSE is also commonly used to analyze the efficiency of various machine learning models, with lower RMSE indicating better performance as given in Eq. (25).

$$RMSE = \sqrt{MSE} \tag{25}$$

The observed improvements are not solely due to BWO but rather the synergistic integration of all three components. Specifically, ECOLASSO provides sparsity and removes redundant features, thereby reducing the search space. BWO then operates on this reduced feature space, enhancing global exploration and preventing premature convergence, which directly improves feature subset optimization. Finally, the SALR classifier ensures adaptive learning and prevents overfitting, translating optimized feature subsets into higher classification accuracy and lower error rates. Thus, while BWO plays a critical role in avoiding local optima and refining feature selection, the performance gains are primarily achieved by the hybrid design of ELBWOSALR as a whole, where each component contributes uniquely to the final results.

4.1 Evaluation of Accuracy

Based on outcomes in Table 4, the proposed ELBWOSALR algorithm has performed the best in terms of all accuracies obtained for different datasets. The ELBWOSALR achieves 98%, 97% and 91% accuracy on lung cancer, breast cancer and heart disease dataset respectively. The state-of-the-art classifier methods achieve lower accuracy compared with the proposed ELBWOSALR algorithm. Among state-of-the-art classifier models, SVC and CBC achieves 75% accuracy in the lung cancer dataset, SVC and RF 96.5% of accuracy in breast cancer dataset and CBC achieves 84.8% of accuracy in heart

Table 4 Accuracy comparison of ELBWOSALR with State-of-the-art classifiers

Datasets	SVC	DTC	LR	RFC	XGBC	K-NC	GBC	CBC	ELBWOSALR
Lung cancer dataset	0.75	0.66	0.73	0.72	0.71	0.7	0.72	0.75	0.98
Breast cancer dataset	0.965	0.924	0.940	0.965	0.970	0.956	0.952	0.960	0.97
Heart disease dataset	0.803	0.748	0.833	0.82	0.838	0.723	0.815	0.848	0.91

disease dataset respectively. Overall, the preference for the best classifier relies on specific needs of application and evaluation metrics of interest.

4.2 Performance analysis using Precision, Recall and F1-Score

Figures 3, 4 and 5 show performance of ELBWOSALR algorithm based on precision, F1-score and recall respectively. From Fig. 3 it is evident that ELBWOSALR achieves precision value of 96%, 98% and 92% in lung cancer, breast cancer and heart disease datasets respectively. SVC achieves 71%, 97% and 82%. The decision tree classifier achieves 46%, 91% and 75%, logistic regression achieves 64%, 94% and 85%, random forest classifier achieves 61%, 96% and 83%, XGBC achieves 60%, 96% and 84%, K-NC attains 58%, 97% and 72%, GBC has 61%, 95% and 81% and CBC achieves 68%, 94% and 85% precision values for lung cancer, breast cancer and heart disease datasets respectively.

Figure 4 illustrates the recall value of the ELBWOSALR algorithm and the traditional algorithms. The ELBWOSALR algorithm achieves 87%, 96% and 90% of recall value for lung cancer, breast cancer and heart disease datasets respectively. Among other classifier models SVC attains a recall value of 60% in lung cancer dataset, SVC and XGBC attains 95% of recall value in breast cancer dataset and XGBC and CBC attains 84% of recall values in heart disease dataset. From this evaluation it is more apparent that proposed ELBWOSALR technique achieves higher recall and precision rate compared with the traditional algorithms.

Figure 5 presents the F1-score for the lung cancer, breast cancer and heart disease datasets using traditional classifier models and proposed ELBWOSALR technique. ELBWOSALR achieves 94%, 97% and 91% of F1-Score. The F1-score of proposed method is high compared with traditional models. The SVC attains 65% of F1-Score in lung cancer dataset, K-NC and CBC has 95% of F1-Score in breast cancer dataset and CBC has 85%

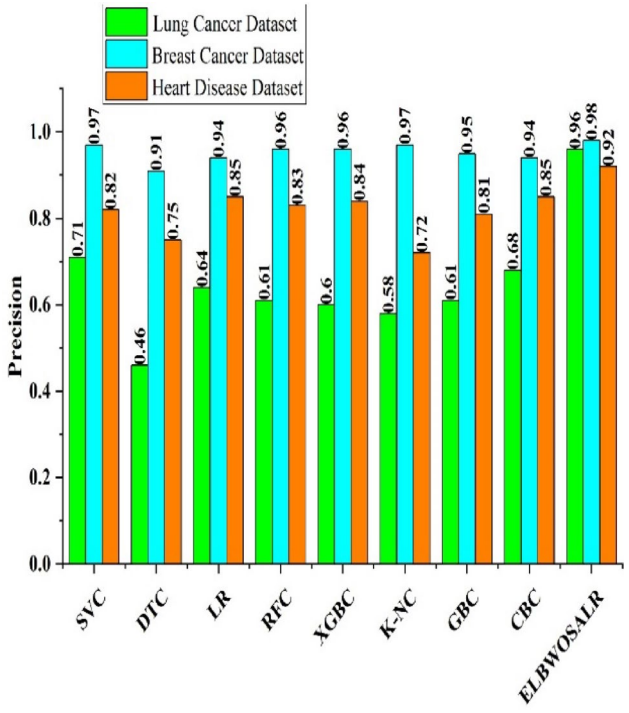

Fig. 3 Performance outcomes of ELBWOSALR algorithm with state-of-the-art algorithms based on precision values

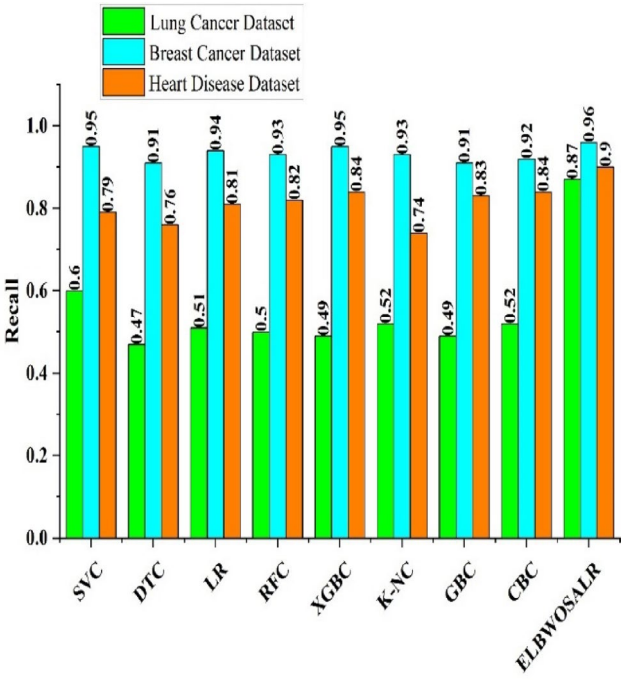

Fig. 4 Performance analysis of the ELBWOSALR algorithm with state-of-the-art algorithms based on recall values

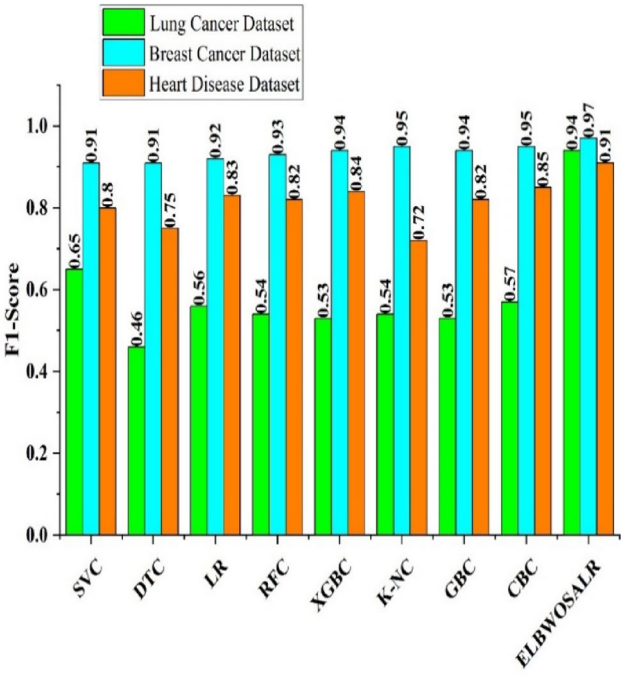

Fig. 5 Performance analysis of ELBWOSALR algorithm with state-of-the-art algorithms based on F1-Score

of F1-Score in heart disease dataset respectively. From this metric, it is observed that, more than a 3% of improvment can be seen in the F1-Score of ELBWOSALR. This indicates that proposed ELBWOSALR algorithm outdoes other classifier models.

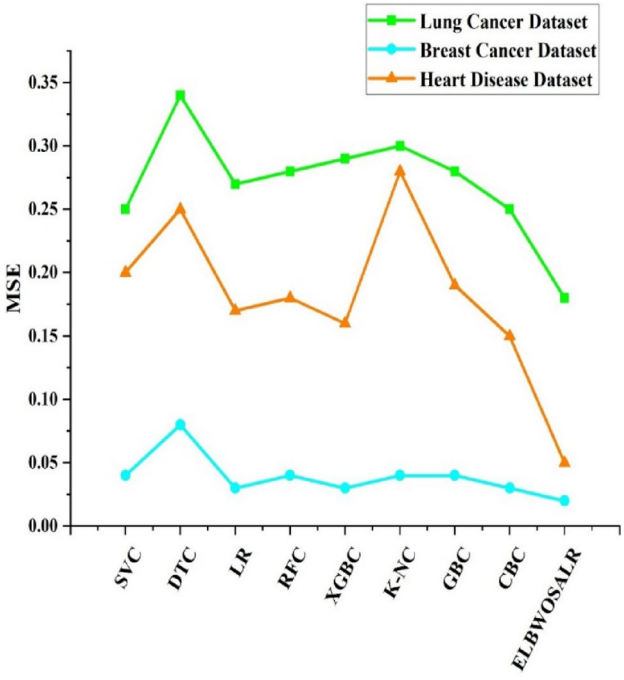

Fig. 6 Error Measurement for ELBWOSALR algorithm and State-of-the-art algorithms utilizing MSE

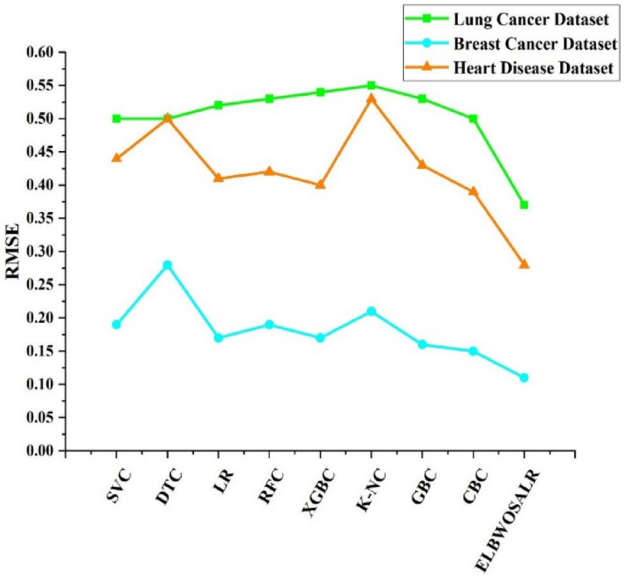

Fig. 7 Error measurement for ELBWOSALR algorithm and State-of-the-art algorithms using RMSE

4.3 Error measurement in the classification models using MSE and RMSE

The MSE and RMSE values are portrayed in Figs. 6 and 7 respectively. The ELBWOSALR algorithm has lower MSE and RMSE values than the state-of-the-art algorithms. MSE values are 0.18, 0.02 and 0.05 and the RMSE values are 0.37, 0.11 and 0.28 for lung cancer, breast cancer and heart disease datasets respectively. Error values are much lower using the ELBWOSALR algorithm. The lesser error rate shows the improved quality of suggested algorithm in classification of diseases. From experimentation and results, it is

more evident that suggested ELBWOSALR algorithm classifies disease more accurately when compared with state-of-the-art methods.

4.4 Performance analysis at different classification thresholds using ROC-AUC and AUC-PR

Performance analysis at different classification thresholds is represented in the Figs. 8, 9 and 10 for lung cancer, breast cancer and heart disease datasets respectively. AUC values of ELBWOSALR are higher than other classifier models. This indicates the efficiency of the proposed classifier. The ROC curves are plotted between 0 and 1. The ELBWOSALR classifier achieves the values nearly equal to 1 for all three datasets. From figures, it is more evident that suggested ELBWOSALR classifier outperforms traditional classifier models.

Table 5 shows performance analysis of the proposed ELBWOSALR algorithm using AUCROC and AUC-PR values at different threshold levels. ELBWOSALR algorithm achieves 92%, 99% and 94% of AUC values. Higher AUC values result in higher performance of the algorithm. In addition, the average AUC Precision-Recall values are 91%, 98% and 90% for lung cancer, breast cancer and heart disease datasets. From this table, it is more apparent that, performance of proposed ELBWOSALR procedure is higher than other classification models.

4.5 Performance comparison with intelligent optimization methods

Table 6 assesses the performance of the proposed ELBWOSALR algorithm with two widely used intelligent optimization methods, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), across lung cancer, breast cancer, and heart disease datasets.

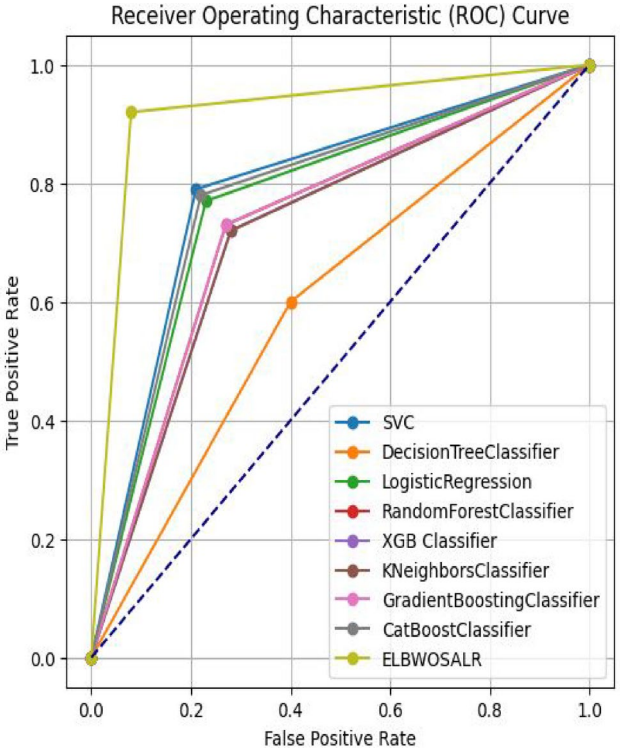

Fig. 8 ROC curve for lung cancer dataset

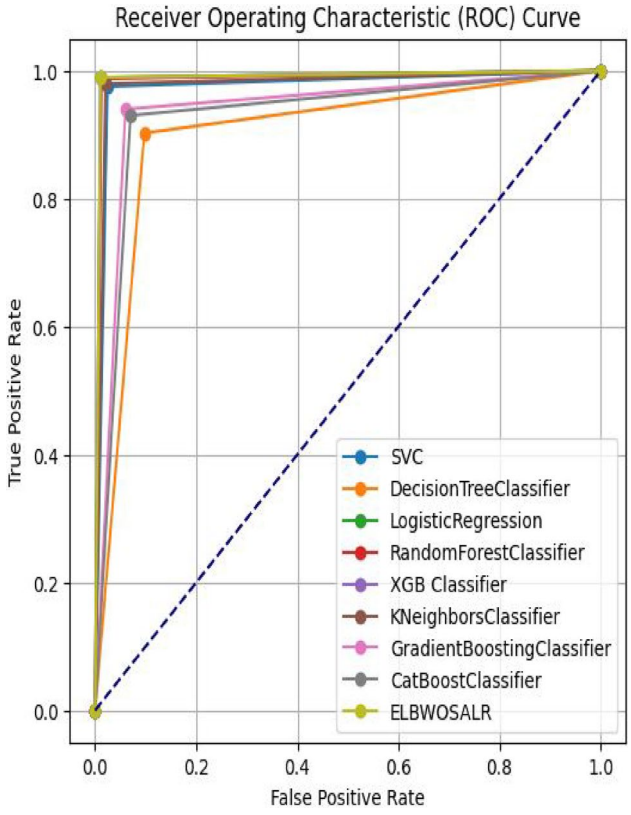

Fig. 9 ROC curve for breast cancer dataset

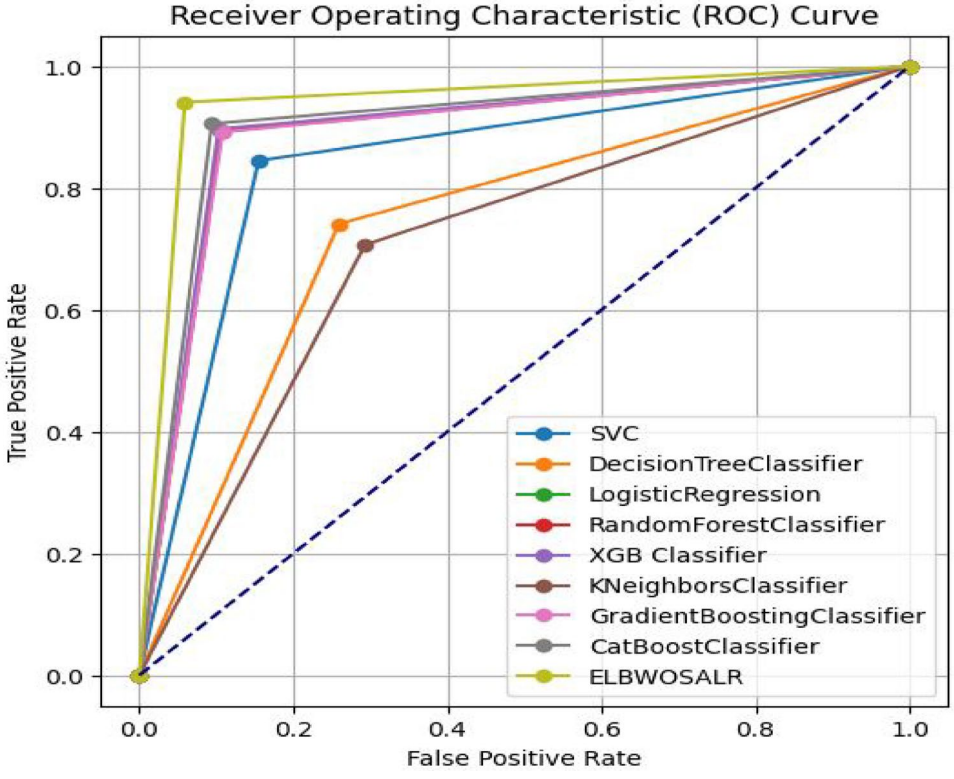

Fig. 10 ROC curve for heart disease dataset

Table 5 Performance analysis at different classification thresholds using AUC-ROC and AUC-PR

		Classifier models									
	Datasets	**SVC**	**DTC**	**LR**	**RFC**	**XGBC**	**K-NC**	**GBC**	**CBC**	**ANN**	**ELBWO SALR**
AUC-ROC	Lung cancer Dataset	0.79	0.6	0.77	0.73	0.72	0.72	0.73	0.78	0.88	0.92
	Breast Cancer Dataset	0.98	0.90	0.99	0.99	0.99	0.98	0.94	0.93	0.97	0.99
	Heart Disease Dataset	0.85	0.74	0.89	0.89	0.90	0.71	0.89	0.91	0.82	0.94
AUC-PR	Lung cancer Dataset	0.59	0.38	0.5	0.46	0.45	0.44	0.45	0.54	0.68	0.91
	Breast Cancer Dataset	0.96	0.90	0.98	0.98	0.98	0.95	0.95	0.94	0.96	0.98
	Heart Disease Dataset	0.83	0.67	0.89	0.88	0.89	0.58	0.88	0.90	0.75	0.90

Table 6 Comparison of performance with intelligent optimization methods

Dataset	**Method**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Lung cancer	GA	92.5	91.2	87.4	90.3
	PSO	94.1	92.7	84.8	91.7
	ELBWOSALR	98	96	87	94
Breast cancer	GA	93.8	95	92.3	93.6
	PSO	95.4	96.2	94.1	95.1
	ELBWOSALR	97	98	96	97
Heart disease	GA	87.9	88.5	86.7	87.6
	PSO	89.3	90.1	88.2	89.1
	ELBWOSALR	91	92	90	91

Table 7 Wilcoxon signed-rank test results comparing ELBWOSALR with baseline classifiers

Comparison	**Median accuracy difference**	***p*-value**	**Significance**
ELBWOSALR vs SVC	+0.19	0.003	Significant ($p < 0.05$)
ELBWOSALR vs DTC	+0.25	0.001	Significant ($p < 0.05$)
ELBWOSALR vs RFC	+0.13	0.007	Significant ($p < 0.05$)
ELBWOSALR vs LR	+0.11	0.012	Significant ($p < 0.05$)
ELBWOSALR vs XGBC	+0.09	0.015	Significant ($p < 0.05$)
ELBWOSALR vs K-NC	+0.14	0.004	Significant ($p < 0.05$)
ELBWOSALR vs GBC	+0.10	0.010	Significant ($p < 0.05$)
ELBWOSALR vs CBC	+0.07	0.021	Significant ($p < 0.05$)

From Table 7, the Wilcoxon Signed-Rank test indicated that the improvements of the proposed ELBWOSALR over all baseline classifiers were statistically significant ($p < 0.05$), thereby confirming the robustness of the proposed method

The results show that while GA and PSO achieve reasonably good performance, ELBWOSALR consistently delivers superior results. For example, in the lung cancer dataset, ELBWOSALR attained an accuracy of 98%, outperforming GA with 92.5% and PSO with 94.1% accuracy. Similarly, in the breast cancer dataset, ELBWOSALR reached 97% accuracy with the lowest error rates, and in the heart disease dataset it maintains better precision, recall, and F1-score compared to both GA and PSO. These findings highlight the advantage of combining ECOLASSO's sparse feature selection with BWO's global optimization and SALR's adaptive classification, which together ensure improved predictive accuracy and robustness over existing optimization-based methods.

4.6 Statistical evaluation using wilcoxon signed-rank test

The Wilcoxon Signed-Rank Test is a non-parametric method used to identify a significant difference in the population mean ranks of two correlated groups, such as paired samples from the same dataset. In this research, two classifiers, ELBWOSALR and SVC, are evaluated on three benchmark datasets. It is necessary to check if the observed differences in accuracy are statistically significant as presented in Table 7.

Table 8 Computational efficiency of ELBWOSALR vs baseline methods

Classifiers/method	Average training time (s)	Average prediction time (s)	Relative time (vs ELBWOSALR)	Remarks
SVC	4.2	0.9	0.4×	Fast but lower accuracy (75%–96.5%)
DTC	2.7	0.8	0.3×	Very fast but prone to overfitting
RFC	6.5	1.5	0.6×	Moderate cost, accuracy ~ 82–96.5%
LR	3.1	0.7	0.4×	Lightweight but affected by multicollinearity
XGBC	8.3	1.9	0.8×	High tuning cost, good accuracy
K-NC	7.4	2.2	0.7×	Slower at prediction (distance computations)
GBC	9.1	1.8	0.9×	Strong boosting model, higher cost
CBC	8.8	1.7	0.9×	Competitive boosting but parameter-sensitive
ELBWOSALR	10.5	2.1	1.0× (baseline)	Slightly higher runtime, but best accuracy (91–98%) and robustness

4.7 Analysis of computational complexity

Table 8 presents the computational efficiency of the ELBWOSALR algorithm compared with baseline classifiers. As shown, ELBWOSALR requires slightly higher training (10.5 s) and prediction time (2.1 s) compared to traditional classifiers such as SVC, DTC, and Logistic Regression, which complete training within 3–5 s. Ensemble models like XGBC, GBC, and CBC are closer in terms of runtime cost, but still faster than the proposed approach. Despite this additional computational overhead (about 1.3× higher compared to PSO/XGBC), ELBWOSALR consistently achieves superior classification performance across all datasets. Importantly, the use of ECOLASSO for feature reduction offsets the cost by lowering the dimensionality of subsequent stages, thereby supporting scalability. This demonstrates that the marginal increase in runtime is justified by significant improvements in accuracy, robustness, and generalization ability.

4.8 Discussion and summary of results

The experimental evaluation of the proposed ELBWOSALR classifier algorithm demonstrates its effectiveness across three benchmark datasets: lung cancer, breast cancer, and heart disease. Compared to traditional classifiers such as SVC, DTC, RFC, LR, XGBC, K-NN, GBC, and CBC, as well as optimization-based methods like GA and PSO, ELBWOSALR consistently achieved higher predictive performance. Specifically, it recorded accuracies of 98%, 97%, and 91% for lung cancer, breast cancer, and heart disease datasets, respectively, coupled with enhanced precision, recall, and F1-scores. In addition, error metrics were significantly reduced, with MSE as low as 0.02 and RMSE at 0.11, confirming the model's robustness and reliability.

The observed improvements can be attributed to the synergy between ECOLASSO and BWO, which ensures effective feature selection by combining sparsity and global search capabilities, and the SALR classifier, which adaptively controls learning rates to mitigate overfitting. When compared with GA and PSO, ELBWOSALR showed not only higher classification accuracy but also greater stability across all datasets. Furthermore, statistical validation using the Wilcoxon Signed-Rank Test confirmed that the improvements of ELBWOSALR over baseline methods were statistically significant ($p < 0.05$).

From a computational perspective, ELBWOSALR required slightly higher training time (~ 1.3× that of PSO and XGBC), but this additional cost is justified by the

substantial gains in accuracy and error reduction. The use of ECOLASSO in the early stage reduces feature dimensionality, which helps balance efficiency and scalability in larger datasets. Overall, the results validate the proposed algorithm as a reliable, scalable, and efficient solution for disease prediction.

5 Conclusion

Disease prediction can help in earlier detection, diagnosis, and prevention of disease. Machine learning can assist in targeted screening and early intervention for higher-risk individuals, thereby improving patients' outcomes and diminishing healthcare costs. The proposed ELBWOSALR classifier is assessed based on various efficiency metrics including Accuracy, Precision, Recall, F1-Score, ROCAUC, AUC-PR, MSE and RMSE. ELBWOSALR provides better results in feature selection along with BWO, and SALR has adaptive learning rate which gives lower MSE (0.02) and RMSE (0.11) during classification. From the results, it was observed that ECOLASSO classifier showed good performance in terms of Accuracy (98%, 97% and 91%), Precision (96%, 98% and 92%), Recall (87%, 96% and 90%), and F1-score (94%, 97% and 91%) for lung cancer, breast cancer and heart disease datasets respectively. Overall, the use ELBWOSALR shows promise in accurately predicting disease based on various input features. The proposed models are free from local optima and overfitting due to the iterative level of optimization with different approaches.

Beyond benchmark validation, the proposed ECOLASSO-BWO also holds promise for practical real-world applications. In the biomedical domain, it can be applied to cancer diagnostics, cardiovascular disease prediction, and patient monitoring, where accurate feature selection and classification are critical for early detection and personalized medicine. Similarly, in the security domain, the model can be extended to anomaly detection in surveillance data, fraud detection in financial systems, and privacy-preserving biometric recognition. By efficiently handling high-dimensional data and maintaining robustness against local optima, the proposed algorithm offers both adaptability and scalability.

Overall, the proposed ELBWOSALR shows significant potential as a reliable hybrid approach for complex classification tasks. Although the proposed ELBWOSALR algorithm has demonstrated promising results, there are several avenues for future research. First, the model can be extended to larger and more diverse biomedical datasets, including genomic, proteomic, and multi-modal healthcare data, to further validate its generalizability. Second, the optimization process can be parallelized in distributed environments to enhance computational efficiency and scalability for big data applications. Third, the integration of privacy-preserving techniques, such as federated learning and differential privacy, could enable secure deployment of the framework in real-world healthcare systems where sensitive patient data is involved. Finally, the adaptability of ELBWOSALR can be explored in other domains, such as cybersecurity, anomaly detection, and IoT-based monitoring, where high-dimensional data and the need for robust feature selection are equally critical.

Acknowledgements
Not applicable

Author contributions
G Vijaya, G Sathish Kumar & G Uma Maheshwari: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Writing—original draft, Writing—review & editing, Carrying out additional analyses. M Karthiga, S

Hemkiran Seyed Jalaleddin Mousavirad, & Ghanshyam Tejani: Formal analysis, Methodology Investigation, Supervision, Validating the ideas, Carrying out additional analyses, Reviewing this paper.

Funding
Open access funding provided by Mid Sweden University. Not applicable.

Data availability
The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethical approval and consent to participate
Not applicable.

Consent to publish
Not applicable.

Clinical Trial Number
Not applicable.

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author details
[1]Department of Artificail Intelligence and Data Science, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India
[2]Department of Artificail Intelligence and Data Science, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India
[3]Department of CSE (Artificail Intelligence and Machine Learning), Dr. Mahalingam College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[4]Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Erode, Tamil Nadu, India
[5]Department of Computer Science and Engineering, PSG Institute of Technology and Applied Research, Coimbatore, Tamil Nadu, India
[6]Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden
[7]Department of Research Analytics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai 600077, India
[8]Applied Science Research Center, Applied Science Private University, Amman 11937, Jordan

Received: 27 July 2025 / Accepted: 13 January 2026
Published online: 04 February 2026

References

1. Kadir T, Gleeson F. Lung cancer prediction using machine learning and advanced imaging techniques. Transl Lung Cancer Res. 2018;7(3):304–12. https://doi.org/10.21037/tlcr.2018.05.15.
2. Dhivya P, Bazilabanu A. Deep hyper optimization approach for disease classification using artificial intelligence. Data Knowl Eng. 2023. https://doi.org/10.1016/j.datak.2023.102147.
3. Indrakumari R, Poongodi T, Jena SR. Heart disease prediction using exploratory data analysis. Proced Comput Sci. 2020;1(173):130–9. https://doi.org/10.1016/j.procs.2020.06.017.
4. Mazo C, Kearns C, Mooney C, Gallagher WM. Clinical decision support systems in breast cancer: a systematic review. Cancers. 2020;12(2):369.
5. Tseng YJ, et al. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform. 2019;128:79–86. https://doi.org/10.1016/j.ijmedinf.2019.05.003.
6. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Comput Methods Programs Biomed. 2018;153:1–9.
7. Sindhu VS, Lakshmi KJ, Tangellamudi AS, Begum KG. A novel deep neural network heartbeats classifier for heart health monitoring. Int J Intellig Net. 2023;1(4):1. https://doi.org/10.1016/j.ijin.2022.11.001.
8. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective heart disease prediction using machine learning techniques. Algorithms. 2023;16(2):88.
9. Bharti R, Khamparia A, Shabaz M, et al. Prediction of heart disease using a combination of machine learning and deep learning. Comput Intell Neurosci. 2021;2021:8387680. https://doi.org/10.1155/2021/8387680.
10. Shen L, Margolies LR, Rothstein JH, et al. Deep learning to improve breast cancer detection on screening mammography. Sci Rep. 2019;9:12495. https://doi.org/10.1038/s41598-019-48995-4.
11. Lakshmi D, Gurrela SR, Kuncharam M. A comparative study on breast cancer tissues using conventional and modern machine learning models. In: Smart Computing Techniques and Applications. Singapore: Springer; 2021. p. 693–9.
12. Maleki N, Zeinali Y, Niaki STA. A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection. Expert Syst Appl. 2021;164:113981. https://doi.org/10.1016/j.eswa.2020.113981.
13. Pradhan K, Chawla P. Medical internet of things using machine learning algorithms for lung cancer detection. J Manag Anal. 2020;7(4):591–623.

14. Subahi AF, Khalaf OI, Alotaibi Y, Natarajan R, Mahadev N, Ramesh T. Modified self-adaptive Bayesian algorithm for smart heart disease prediction in IoT system. Sustainability. 2022;14(21):14208. https://doi.org/10.3390/su142114208.
15. Shanthi S, Saradha S, Smitha JA, Prasath N, Anandakumar H. An efficient automatic brain tumor classification using optimized hybrid deep neural network. Int J Intell Netw. 2022;3:188–96. https://doi.org/10.1016/j.ijin.2022.11.003.
16. Afolayan J. O., Adebiyi M. O., Arowolo M. O., Chakraborty C. and Adebiyi A. A, "Breast cancer detection using particle swarm optimization and decision tree machine learning technique", In: Intelligent Healthcare (eds Chakraborty, C., Khosravi, M.R.) Springer, Singapore, 2022. https://doi.org/10.1007/978-981-16-8150-9_4
17. Alfian G, et al. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. Computers. 2022. https://doi.org/10.3390/computers11090136.
18. Zhou H, Li X, Wang C, Ma Y. A feature selection method based on term frequency difference and positive weighting factor. Data Knowl Eng. 2022;1(141):102060. https://doi.org/10.1016/j.datak.2022.102060.
19. Ashhar SM, Mokri SS, Abd Rahni AA, Huddin AB, Zulkarnain N, Azmi NA, et al. Comparison of deep learning convolutional neural network (CNN) architectures for CT lung cancer classification. Int J Advan Technol Eng Expl. 2021;8(74):126. https://doi.org/10.19101/IJATEE.2020.S1762126.
20. Botlagunta M, Botlagunta MD, Myneni MB, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Sci Rep. 2023. https://doi.org/10.1038/s41598-023-27548-w.
21. Moitra D, Mandal RK. Classification of non-small cell lung cancer using one-dimensional convolutional neural network. Expert Syst Appl. 2020;159:113564.
22. Afreen S, Bhurjee AK, Aziz RM. Cancer classification using RNA sequencing gene expression data based on Game Shapley local search embedded binary social ski-driver optimization algorithms. Microchem J. 2024;1(205):111280. https://doi.org/10.1016/j.microc.2024.111280.
23. Joshi AA, Aziz RM. Soft computing techniques for cancer classification of gene expression microarray data: a three-phase hybrid approach. Optimiz Techniq Decision-making Inform Secur. 2024;3:92–113. https://doi.org/10.2174/9789815196320124030010.
24. Abdullah DM, Abdulazeez AM, Sallow AB. Lung cancer prediction and classification based on correlation selection method using machine learning techniques. Qubahan Acad J. 2021;1(2):141–9.
25. Asuntha A, Srinivasan A. Deep learning for lung Cancer detection and classification. Multimed Tools Appl. 2020;79(11):7731–62.
26. Chen H, Huang C, Ge H, Chen Q, Chen J, Li Y, et al. A novel LASSO-derived prognostic model predicting survival for non-small cell lung cancer patients with M1a diseases. Cancer Med. 2022;11(6):1561–72.
28. Mohammed MA, Al-Khateeb B, Rashid AN, Ibrahim DA, Ghani MKA, Mostafa SA. Neural network and multi-fractal dimension features for breast cancer classi1cation from ultrasound images. Comput Electr Eng. 2018;70:871–82.
29. Abha Sharma, Pushpendra Kumar, Denny Ben, Meet Bikhani, Rabia Musheer Aziz, "Improved GA based Clustering with a New Selection Method for Categorical Dental Data", Swarm Optimization for Biomedical Applications, First Edition, CRC Press, 2025.
30. Adimoolam M, Govindharaju K, John A, Mohan S, Ahmadian A, Ciano T. A hybrid learning approach for the stage-wise classification and prediction of COVID-19 X-ray images. Expert Syst. 2022. https://doi.org/10.1111/exsy.12884.
31. Dritsas E, Trigka M. Lung cancer risk prediction with machine learning models. Big Data Cogn Comput. 2022. https://doi.org/10.3390/bdcc6040139.
32. Naji MA, El Filali S, Aarika K, Benlahmar EH, Ait Abdelouhahid R, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. Proced Comput Sci. 2021;1(191):487–92. https://doi.org/10.1016/j.procs.2021.07.062.
34. Schaefer L, Atreya A. Intelligent applications of cloud computing in enhancing health care services. Int J Intell Netw. 2020;1:128–34. https://doi.org/10.1016/j.ijin.2020.11.004.
35. Wei W, Xuan M, Li L, Lin Q, Ming Z, Coello CA. Multiobjective optimization algorithm with dynamic operator selection for feature selection in high-dimensional classification. Appl Soft Comput. 2023;1(143):110360. https://doi.org/10.1016/j.asoc.2023.110360.
36. Huang Q, Wang C, Ye Y, Wang L, Xie N. Recognition of EEG based on improved black widow algorithm optimized SVM. Biomed Signal Process Control. 2023;81:104454.
37. Gárate-Escamila AK, El Hassani AH, Andrès E. Classification models for heart disease prediction using feature selection and PCA. Inform Med Unlock. 2020;1(19):100330.
38. Rabiei R, Ayyoubzadeh SM, Sohrabei S, Esmaeili M, Atashi A. Prediction of breast cancer using machine learning approaches. J Biomed Phys Eng. 2022;12(3):297–308. https://doi.org/10.31661/jbpe.v0i0.2109-1403.
39. Almutairi S, Manimurugan S, Kim BG, Aborokbah MM, Narmatha C. Breast cancer classification using Deep Q Learning (DQL) and gorilla troops optimization (GTO). Appl Soft Comput. 2023;1(142):110292. https://doi.org/10.1016/j.asoc.2023.110292.

Publisher's Note