scientific reports

Check for updates

OPEN

Classification of biomedical lung cancer images using optimized binary bat technique by constructing oblique decision trees

Shobha Aswal[1]✉, Neelu Jyothi Ahuja[2] & Ritika Mehra[3]

Due to imbalanced data values and high-dimensional features of lung cancer from CT scans images creates significant challenges in clinical research. The improper classification of these images leads towards higher complexity in classification process. These critical issues compromise the extraction of biomedical traits and also design incomplete classification of lung cancer. As the conventional approaches are partially successful in dealing with the complex nature of high-dimensional and imbalanced biomedical data for lung cancer classification. Thus, there is a crucial need to develop a robust classification technique which can address these major concerns in the classification of lung cancer images. In this paper, we propose a novel structural formation of the oblique decision tree (OBT) using a swarm intelligence technique, namely, the Binary Bat Swarm Algorithm (BBSA). The application of BBSA enables a competitive recognition rate to make structural reforms while building OBT. Such integration improves the ability of the machine learning swarm classifier (MLSC) to handle high-dimensional features and imbalanced biomedical datasets. The adaptive feature selection using BBSA allows for the exploration and selection of relevant features required for classification from ODT. The ODT classifier introduces flexibility in decision boundaries, which enables it to capture complex linkages between biomedical data. The proposed MLSC model effectively handles high-dimensional, imbalanced lung cancer datasets using TCGA_LUSC_2016 and TCGA_LUAD_2016 modalities, achieving superior precision, recall, F-measure, and execution efficiency. The experiments are conducted in Python to evaluate the performance metrics that consistently demonstrate enhanced classification accuracy and reduced misclassification rates compared to existing methods. The MLSC is assessed in terms of both qualitative and quantitative measurements to study the capability of the proposed MLSC in classifying the instances more effectively than the conventional state-of-the-art methods.

Keywords Binary Bat Swarm Algorithm, Biomedical data, Lung cancer, Imbalanced data, Oblique decision tree

In the past decade, the potent impact of data analysis in the biomedical domain has gained substantial attention for early diagnosis and treatment[1,2]. Technological advancement provides unique availability to high-dimensional datasets, where the complexity of analyzing the data poses a significant challenge[3,4]. The difficulty of accurate classification is made more difficult by the occurrence of an imbalanced dataset in biomedical research[5–7].

The analysis of biomedical data leads the modern healthcare system because of its power in transforming the diagnostic process, treatment, and patient care[8]. The complex ecosystem of cancer research, especially lung cancer, requires sophisticated analytical models to acquire essential information from the data generated from diverse sources, including proteomics, genomics, and imaging[9,10]. The combination of such diverse data types poses a serious challenge to identifying the hidden patterns and linkages within the datasets. This requires a sophisticated algorithm for unraveling the pattern and linkages[11,12]

On the other hand, early and accurate classification is required for lung cancer, as it is the leading cause of mortality globally[13]. Conventional diagnostic techniques limit their classification ability in classifying the subtle data types in heterogeneous datasets[14]. Biomedical data analysis is hence required to harness the power

[1]Department of Computer Science and Engineering, VM Singh Bhandari Uttarakhand Technical University, Dehradun, India. [2]Department of Systemics, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India. [3]Department of Computer Science and Engineering, Dev Bhoomi Uttarakhand University, Dehradun, India. ✉email: shobha.swl@gmail.com

of computational techniques or algorithms and to study these complex data patterns[15–17]. Hence, the goal is to classify the malignancies and subtypes in the dataset, predict prognosis, and personalize treatment for individuals.

With such challenges in lung cancer Computed Tomography (CT) scans image classification, the elimination of imbalanced datasets requires machine learning (ML) models[18]. Due to these imbalanced datasets, where each class outweighs the subsequent classes, biases have a negative impact on ML algorithms[19]. The TCGA_LUSC_2016 and TCGA_LUAD_2016 datasets exhibit significant class imbalance, primarily due to unequal representation of tumor subtypes and stages. This imbalance can bias classifiers toward majority classes, reducing sensitivity to minority class predictions. Proper quantification and mitigation strategies are essential to ensure robust and generalizable model performance. Moreover, the application of high-dimensional datasets in ML increases the challenge of extracting useful information as it involves many features and variables. Thus, to realize the potential of data analysis in improving understanding and lung cancer treatment, it is crucial to overcome such obstacles in ML algorithms.

Thus, the challenges associated with biomedical classification related to lung cancer data are multi-faceted. The limitations of conventional classifiers in handling the imbalanced dataset are due to bias constraints and the high-dimensional representation of minority samples. The feature extraction and classification have therefore become complex, which demands sophisticated machine learning models.

This research aims to resolve the problem of misclassification of minority samples in biomedical datasets associated with lung cancer in CT scans images. The reliability of classification is endangered by its inability to produce accurate results when handling high-dimensional and imbalanced data with the bias problem in ML. Addressing such constraints benefits the ML in classifying the datasets precisely. The research presents a novel approach for the classification of biomedical data pertaining to lung cancer treatment. It aims to mitigate the problems of misclassification of minority samples and handling high-dimensional and imbalanced data with the bias problem in ML using a novel framework.

In this research, a novel framework is employed to enable robust feature extraction and classification of lung cancer datasets. The benefit of using Binary Bat Swarm Algorithm (BBSA) is that it offers a balanced trade-off between exploration and exploitation through frequency modulation and adaptive loudness, making it well-suited for high-dimensional feature selection. Unlike Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), BBSA's binary encoding and random walks enhance its ability to efficiently navigate discrete feature spaces. The BBSA is designed to create an Oblique Decision Tree (OBT) for classifying the instances correctly even in the presence of the above constraints. The creation of OBT using BBSA enables a novel framework that eliminates the problem of bias in OBT to handle the high-dimensional features in imbalanced datasets.

The main contribution of the research involves the following:

- The research introduces a novel ML Swarm Classifier (MLSC) approach that combines the BBSA and ODT for lung cancer classification from CT scans images. This approach uses swarm intelligence for adaptive feature selection in biomedical datasets and then structures the OBT to handle the complexities of high-dimensional and imbalanced biomedical data.
- The proposed MLSC approach addresses class imbalance by enhancing precision and recall through minority-sensitive learning, effectively reducing misclassification of underrepresented samples.
- The proposed research enhances the robustness and accuracy of biomedical data classification using BBSA for optimizing the ODT structure, where it adapts to high-dimensional feature spaces, mitigates bias problems on an imbalanced dataset, and offers accurate classification outcomes.

The outline of the paper is given below: Sect. "Related works" provides the related works. Section "Proposed MLSC biomedical lung cancer data classification" discusses the proposed MLSC framework. Section "Results and discussion" provides a comprehensive evaluation of MLSC with state-of-the-art models. Section "Conclusion" concludes the entire work with possible directions for future scope.

Related works

The following research contribute to the biomedical data classification in several ways:

Swarm based ML on biomedical datasets

A pinhole-imaging-based learning strategy and crossover operator Binary Sand Cat Swarm Optimization (PILC-BSCSO)[20] and Chemotaxis Foraging-Shuffled Frog Leaping Algorithm (BF-SFLA) algorithm[21] are developed for high-dimensional feature selection in biomedical datasets. It increases the exploration and search capabilities for addressing the problem of dimensionality. Likewise, a hybrid Ensemble Enhanced-Grey Wolf Optimization (E-GWO)[22] method is applied to feature selection. This is combined with various classifiers, including Random Forest Bagging and Naive Bayes Bagging, which achieve higher accuracy. In[23], we propose a rank aggregation filtering for hybrid feature selection with an Improved Squirrel Search (ISS) Algorithm for classification. It achieves higher classification accuracy across nine biomedical datasets. In[24], a classifier named Random Vector Functional Link (RVFL) is applied to noisy biomedical datasets, and a ε-insensitive Huber loss function (ε-HRVFL) is applied to achieve higher accuracy in handling noisy datasets. Table 1 shows the detailed summary of swarm-based ML on biomedical dataset for feature selection and classification.

References	Feature selection algorithm	Classification algorithm	Dataset used	Performance metrics	Results
[20]	PILC-BSCSO	Support vector machine	Three public medical datasets, including colon cancer and TCGA-HCC	Classification accuracy	It achieves 100% accuracy for colon cancer
[21]	BF-SFLA	K-NN, C4.5 decision tree	Not specified	Classification accuracy, Classification time	BF-SFLA-based method improves feature subset, accuracy, and reduces classification time compared to other algorithms
[22]	Enhanced-grey wolf optimization (E-GWO)	Naive Bayes Bagging, Random Forest Bagging, Decision Tree Bagging, K-Nearest Neighbors Bagging, Neural Network Bagging, Gradient Boosting, Adaptive Boosting	Combined five biomedical heart disease datasets: Cleveland, Long-Beach-VA, Switzerland, Hungarians, and Statlog	Accuracy, Recall, Precision, F1-score, Specificity, Error Rate, G-mean, False Negative Rate (FNR), False Positive Rate (FPR), Negative Predictive Value (NPV)	RFBT achieves the highest accuracy of 99.26%, 11.90% improvement over the conventional model
[23]	Rank aggregation in hybrid feature selection model	Improved squirrel search algorithm	Nine biomedical datasets	Classification accuracy, Computational time	superior accuracy and reduced computational time
[24]	Random vector functional link (RVFL)	Support vector machine, extreme learning machine with RBF and sigmoid activation functions	Biomedical datasets	Classification accuracy	ε-HRVFL achieves highest accuracy of 96.52% for biomedical datasets and 98.13% for non-biomedical datasets

Table 1. Summary of existing methods.

Advanced biomedical data classification techniques

A complementary Naïve Bayesian (CNB)[25] is developed for noisy and imbalanced biomedical instances, and it uses double learning for model construction. This improves the computational time and results in a significant enhancement in accuracy compared to boosted naïve Bayesian approaches. A multi-category classification in biomedical datasets[26] uses various algorithms like decision trees (DT), random forests (RF), support vector machines (SVM), and artificial neural networks (ANN) for accurate classification in large rectangular biomedical datasets. The accuracy is further improved using information entropy, and this further enhances the efficacy of dimensionality reduction with similar performance on all ML models. The Classification and Regression Tree (CART) algorithm is applied for medical data classification in[27] to address redundant data and missing values. It is then combined with the Boruta method for dimensionality reduction. A soft computing and ML technique using Radial Basis Function Neural Networks (RBFNN) for prostate cancer prediction is developed in[28]. It identifies the cancer in its early stages with diverse symptoms collected from the DBCR dataset. The hit rate, mean square error rate, accuracy, and selectivity account for the effective classification of abnormal features. The neural computing combined with a soft computing approach like Discrete AdaBoost Optimized Ensemble Learning Generalized Neural Networks for Lung Cancer Detection are developed in[29] to avoid dimensionality and overfitting. The data collected from the ELVIRA Biomedical Data Set Repository helped in the successful analysis of the model. A one-class classification is developed in[30] for deep kernel learning with small biomedical datasets. The Kernel Regularized Least Squares (KRL) method-based deep architecture improves the generalization capability by embedding minimum variance. This proves to be performing better than the state-of-the-art methods on diverse biomedical benchmark datasets. Hybrid greedy ensemble feature selection with greedy parameter-wise optimization and genetic algorithm (GA) classification is conducted in[31]. The former uses a weighted occurrence frequency and penalty scheme on high-dimensional biomedical datasets. The research outperforms prediction accuracy by 15% compared to existing methods by overcoming irrelevant feature sets and predictive accuracy issues. Table 2 shows the advanced biomedical data classification techniques.

As reported in[32], BBSA outperforms various optimization tasks, especially in biomedical data classification with high-dimensional features. The BBSA has the capability to intelligently search for relevant features. Unlike the methods in Sect. "Advanced biomedical data classification techniques", ODTs have the ability to adaptively adjust their structure with relevance to the data patterns. While dealing with complex and diverse datasets, its adaptability is crucial, and it further establishes non-axis-aligned decision boundaries to invoke intricate relationships between the features. With such advantages, the research uses BBSA for feature selection and ODT for classification purposes to classify the biomedical datasets.

Proposed MLSC biomedical lung cancer data classification

In this section, the MLSC uses BBSA with the creation of OBT to resolve the inherent challenges of biomedical lung cancer data classification. The architecture of which is given in Fig. 1.

Paper	Algorithm	Dataset	Performance Metrics	Results
[25]	Complement naïve Bayesian (CNB)	Single photon emission computed tomography, Indian liver patient, wilt, tic-tac-toe endgame	Computational time, accuracy	CNB showed 19.58% and 10.51% accuracy enhancement on first and second biomedical data. Outperformed most boosted Naïve Bayesian approaches
[26]	Random forest (RF), Decision tree (DT), Artificial neural network (ANN), Support vector machine (SVM), Multinomial logistic regression	Female breast cancers (surveillance, epidemiology, and end results-18 database)	Accuracy	RF, DT, ANN, and SVM had similar accuracy for multi-category outcomes. Dimension reduction based on information gain increased model efficiency while maintaining accuracy
[27]	Classification and regression tree (CART), Boruta method	UCI dataset	Accuracy	Improved accuracy of CART algorithm with the Boruta method
[28]	Radial basis function neural network (RBFNN)	DBCR prostate cancer dataset	Mean square error rate, hit rate, selectivity, accuracy	Efficient classification of abnormal prostate features
[29]	Discrete AdaBoost optimized ensemble learning generalized neural networks	ELVIRA biomedical data set repository	Error rate, precision, recall, G-mean, F-measure, prediction rate	Successful analysis and classification of biomedical lung data
[30]	Kernel regularized least squares (KRL) method-based deep architecture	Benchmark datasets (13 biomedical, 5 others)	F1-score	More than 5% F1-Score improvement compared to state-of-the-art methods for biomedical datasets
[31]	Hybrid greedy ensemble approach (HGEA)	High-dimensional biomedical datasets	Prediction accuracy	Out performed base feature selection techniques by 4.17%–15.14% in terms of prediction accuracy

Table 2. Summary of existing methods.

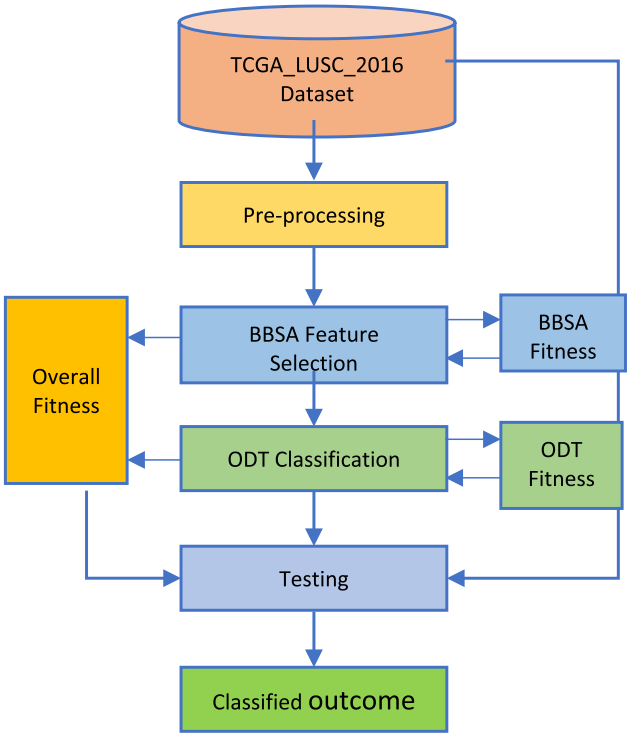

Fig. 1. Proposed MLSC Method with BBSA feature selection and OBT Classification Models.

(1) BBSA Feature Selection: The BBSA is inspired from the echolocation bats behavior. In case of biomedical data classification, BBSA handles the imbalanced datasets for feature selection by mimicking the bat echolocation process to search and optimize the solutions present in a multidimensional space.
(2) ODT Classifier: ODT is employed, which enables the flexible decision boundaries to establish complex linkage between high-dimensional data, and this enables the DT to capture the essential patterns for classification.
(3) BBSA-ODT Classification: In MLSC, BBSA optimizes the structure and parameters of the ODT and while performing the optimization process, the BBSA-ODT algorithm refines the DT boundaries to increase the classification accuracy. This specifically aims at eliminating the problem of misclassification associated with

minority samples. This algorithm combines the optimization capabilities of BBSA and the ODT flexibility to handling the challenges on high-dimensional and imbalanced datasets.

Input: Biomedical lung cancer dataset (features and labels)
Output: Trained Classifier

a) Initialize BBA parameters (frequency range, population size, exploration rate and loudness).

b) Generate an initial binary solution that represents the potential features.

c) For each iteration:

   i) Evaluate each solution fitness using classifier performance.

   ii) Update position and velocity of bats.

   iii) Use frequency modulation to explore the solution space.

   iv) Update loudness and explore the solution space with levy random walk.

   v) Select the best solutions for feature selection.

d) Use the selected features from BBA to create original feature subset.

e) Initialize ODT parameters (maximum depth, number of decision trees, learning rate, minimum samples per leaf, and regula strength).

f) For each decision tree:

   i) Train an ODT on the subset of selected features.

   ii) Apply ODT for feature weighting and decision boundary flexibility.

   iii) Repeat until the specified decision trees are created.

g) Form a Dual-Dimensional Classifier that combines the BBSA selected features with the DTs from ODT.

h) Apply BBSA-ODT optimization for biomedical data classification:

   i) Evaluate the classifier.

   ii) Update BBA parameters w.r.t classifier performance.

   iii) Adjust ODT parameters for improved decision tree structure.

   iv) Repeat until convergence.

i) Fine-tune BBA-ODT parameters based on validation performance.

j) Optimize the structure/parameters of the classifier.

k) Train the classifier on the dataset using optimized parameters.

l) Evaluate the performance of the trained MLSC model

Algorithm 1. MLSC

Proposed BBSA with levy movements

The BBSA (Fig. 2) is utilized in this research to solve the optimization problems associated with feature selection in binary domains, where it mimics the foraging bats behavior with its exploration and exploitation behavior of bats for their prey using echolocation.

The following process applies for the BBSA behavior in the solution space:

- Echolocation and Levy Foraging Behavior: In former, the bats listen to the echoes to navigate and locate the features and this process is referred as echolocation. In latter, the random exploration using levy flight movement and directed movement towards the features is gathered via echolocation.
- Binary Encoding: The operation in binary domain makes it suitable to resolve the issues associated with the selection of binary decision variables for feature selection task. The solutions are represented in binary string, with each bit belonging to a decision variable.

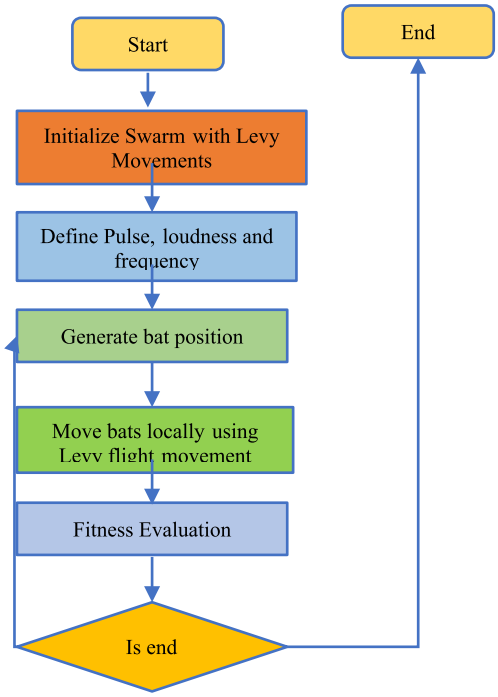

Fig. 2. BBSA flow chart.

In solution space, the BBSA process involves random seed generation from the binary string population to acquire possible solution. To initialize the position and velocity, the following considerations are made using Eq. (1):

$$X_i^0, V_i^0 \sim \text{Levy Flights} \tag{1}$$

where, X_i represents the position of bat *i*, and V_i represents its velocity. X_i^0 is the initial position of bat *i*. V_i^0 is the initial velocity of bat *i*.

Bats obtain the current and global best solution using echolocation behavior via exploration–exploitation balance. The bats positions are updated at each iteration using echolocation and frequency tuning, which leads to the optimal solution space. The echolocation and frequency tuning for conducted using the expression presented in Eq. (2):

$$f_i = f_{\min} + (f_{\max} - f_{\min}) \times \beta \tag{2}$$

where f_i represents the frequency of bat *i*. $f_{\min}$ and $f_{\max}$ are minimum and maximum frequencies. $0 \leq \beta \leq 1$ is a random value.

Further, the velocity update is conducted using the expression in Eqs. (3) and (4):

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{3}$$

$$LF(V_i(t+1)) = V_i(t) + (X_{tr} - X_i(t)) \times fi \times r_{\text{LF}} \tag{4}$$

where X_i is updated based on the current position, velocity, and echolocation. r_{LF} is a Levy Flight value in the range [0,1). X_{tr} is the current best solution in the population.

The loudness of the pulses (Eq. (5)) is updated using the solution fitness and the bats emits louder pulses for better solutions, which influences the search process largely.

$$A_i = \alpha A_i \tag{5}$$

where A_i represents the loudness of bat *i*. $0 \leq \alpha < 1$ is a constant controlling loudness decay.

The position update based on the pulse emission is expressed as in Eq. (6):

$$LF(X_{new}) = X_{old} + A_i \times \sin(f_i) \times r_{LF} \tag{6}$$

where $\boldsymbol{X}_{old}$ = Current position of bat r_l is the random values.

X_{new} is the updated position.

The proposed method introduces levy flight walk to improve the exploration and simulates the stochastic bat movement. This helps to acquire optimal features related to the biomedical domain, which helps in reducing the misclassification problem.

Objective function

This fitness function allows the BBSA to find the features set that maximizes the classification accuracy while reducing the selected features. Then, the BBSA iteratively searches the optimal feature subset with binary value adjustment in the feature selection vector using the fitness function. In this section, the objective function is estimated for the feature selection process using BBSA and it is formulated by maximizing the classification performance while reducing the total selected features. The research considers the binary vector X as a feature selection vector, which is defined in Eq. (7):

$$X_i = \begin{cases} 1 & \text{if feature } i \text{ is selected} \\ 0 & Otherwise \end{cases} \tag{7}$$

Let $f(X)$ be the fitness function, which is defined as a combination of total selected features and the feature selection accuracy. Thus, $f(X)$ is defined in Eq. (8):

$$f(X) = w \cdot AC(x) - (1-w) \cdot \frac{K}{Ns} \tag{8}$$

$AC(x)$ is the accuracy of the selected features. *Ns*—Number of Selected Features

K is the total number of features.

w is the weight parameter that balances the number of selected features and the importance of accuracy. Adjusting weight parameter w makes the BBSA to prioritize feature selection accuracy over sparsity of feature or vice versa.

ODT for biomedical data classification

An ODT is a variant of DT that permits the decision boundaries to aligned with the feature space axis. The ODT allows for splitting the feature space at arbitrary angles rather than splitting along the coordinate axes. It enables flexibility for the ODT to capture complex linkage between the data, especially when the boundaries fail to align with the coordinate axes. The ODT applies this splitting criterion at each decision tree nodes that finds how the feature space is partitioned.

Consider a feature vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ with the decision boundary parameters i.e. weight vector as $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ and bias as b.

Linear Decision Boundary: The node decisions are determined based on the condition $w \cdot x + b$ as stated in Eq. (11), which defines the linear decision boundary.

$$Child = \begin{cases} \text{Left} & \text{if } w \cdot x + b \geq 0 \\ \text{Right} & \text{if } w \cdot x + b < 0 \end{cases} \tag{9}$$

Splitting Criteria: This splitting criterion involves finding the optimal weight w and bias b to maximize information gain or minimize impurity. Consider an impurity measure I (Gini impurity) for the dataset D at a node.

$$OS: \; w^*, b^* = \arg\min_{w,b} \sum_{d_{left}} I(D_{left}) + \sum_{d_{right}} I(D_{right}) \tag{10}$$

where D_{left} and D_{right} are the datasets present in the left child nodes and right child nodes, respectively. d_{left} and d_{right} are the data points in the left child nodes and right child nodes, respectively.

The recursive learning process applies the splitting criteria to split the feature space till the maximum depth is reached.

//Initialization

1) Initialize the Root Node and allow the features selected to it.

//Splitting Criteria Selection

2) Use Gini impurity measure to assess the splitting quality.
   a) Validate the optimal linear decision boundary.

// Optimize Decision Boundary

3) For a node
   a) Find the optimal parameters w^* and b^* to maximize information gain and minimize impurity.

// Split the Node

4) Split the selected node into left and right child using $w*$ and $b*$.

//Recursive Splitting

5) Apply the algorithm recursively to the left and right child.
6) For each child, repeat steps 2-4 if minimum samples per node and maximum depth is reached.

//Stopping Criterion

7) Create a leaf node
8) Assign the leaf node to the majority class or the class that minimizes the node impurity.

//Build the Tree

9) Continue the recursive splitting until the tree is built.

//Predict Class for New Data

10) Traverse the tree from root node → leaf node using decision boundaries
11) Assign the leaf node class as the predicted class.

Algorithm 2. ODT Algorithm for Biomedical Data Classification

The flexibility across the decision boundaries in ODT is defines the decision made for classification. The linear decision boundary is defined in a binary classification task is expressed in Eq. (11):

$$f\left(x; w, b\right) = w \cdot x + b \tag{11}$$

where x is the feature vector, w is the weight vector, b is the bias term.

This linear expression defines the decision made in the hyperplane present in the feature space. The decision is hence made if $f(x; w, b) \geq 0$ (on one side of a hyperplane) or $f(x; w, b) < 0$ (on the other side of the hyperplane).

Thus, the optimization objective for the ODT classification process is expressed in Eq. (12):

$$\mathrm{Min}J\left(w, b\right) = H\left(p, q\right) + \lambda \cdot \|w\|_2^2 \tag{12}$$

The research uses cross-entropy function $H(p,q)$ (refer Eq. (13)) to control the decision boundary complexity with binary classification.

$$H\left(p, q\right) = -[p \cdot \log\left(q\right) + (1 - p) \cdot \log\left(1 - q\right)] \tag{13}$$

where: p is the true probability of the positive class (class 1). q is the predicted probability of the positive class.

The combination of BBSA and ODT requires to find a balance between optimized ODT structure and BBSA feature selection to capture the complexity of the biomedical data classification. The research defines the objective function to combine the BBSA optimization with performance metrics of ODT. It involves minimizing a cross-entropy function and weighted sum of classification loss as in Eq. (14):

$$\mathrm{Min}J = \alpha \cdot J_{\mathrm{BBSA}} + (1 - \alpha) \cdot J_{\mathrm{ODT}} \tag{14}$$

where, J_{BBSA}—optimization criteria of BBSA, J_{ODT}—performance metrics of ODT, and α—weighting parameter.

1) Initialize BBSA
   a) Initialize a population of binary solutions that represents the DT structures using BBSA.
   b) Set parameters such as population size, loudness, frequency range, and exploration rate.
2) BBSA-ODT Objective Function:
   a) Define the function (Eq.(14)) for the optimization criteria of BBSA with ODT for biomedical data classification.
3) Fitness Evaluation:
   a) For each binary solution in BBSA population:
      i) Evaluate the fitness using Eq.(12)
4) BBA Optimization Iterations:
   a) While termination criteria met:
      i) For each binary solution in BBSA population:
         (1) Update binary solution using BBSA optimization rules.
         (2) Evaluate the updated solution fitness using Eq.(12).
5) Convert Binary Solutions to ODTs:
   a) For each binary solution in the final population:
      i) Convert the binary solution into an ODT structure.
      ii) Determine the decision boundaries at each node based on the features.
6) Biomedical Data Classification
   a) For each ODT
      i) Evaluate the classification metrics.
7) Select Best DT
   a) Identify the ODT with best performing DT
8) Termination Criteria

Algorithm 3. ODT Algorithm for Biomedical Data Classification

Results and discussion

In this section, the proposed MLSC is validated quantitatively and qualitatively based on the parameters given in Table 3. The experiments are conducted in a python simulator, which runs on a i7 processor with 16 GB of RAM. The proposed method is evaluated on two different datasets that includes TCGA_LUSC_2016[33] and TCGA_LUAD_2016[29], where the former contains 552 samples and 576 samples, respectively. The MLSC model uses BBSA parameters like population size, frequency, loudness, and exploration rate to effectively search and select optimal feature subsets from biomedical lung cancer data. These selected features are then used by ODT, which leverages parameters such as tree depth, learning rate, and regularization strength to build an ensemble of efficient decision trees. Together, this integration enables robust classification by combining global feature optimization with localized decision boundary learning.

Parameter	Value/range
Population size (BBSA)	50
Frequency range (BBSA)	[0, 1]
Loudness (BBSA)	[0, 1]
Exploration rate (BBSA)	[0, 1]
Number of DTs (ODT)	5
Maximum depth (ODT)	10
Minimum samples per leaf (ODT)	5
Learning rate (ODT)	0.1
Regularization strength (ODT)	0.001
Train-test	80:20
Cross-validation	Fivefold stratified
BBSA hyperparameters	tuned for binary feature selection
ODT implemented using	Gradient boosting with limited estimators and regularization
Evaluation metrics	Accuracy and F1-score for robust performance measurement

Table 3. Experimental setup.

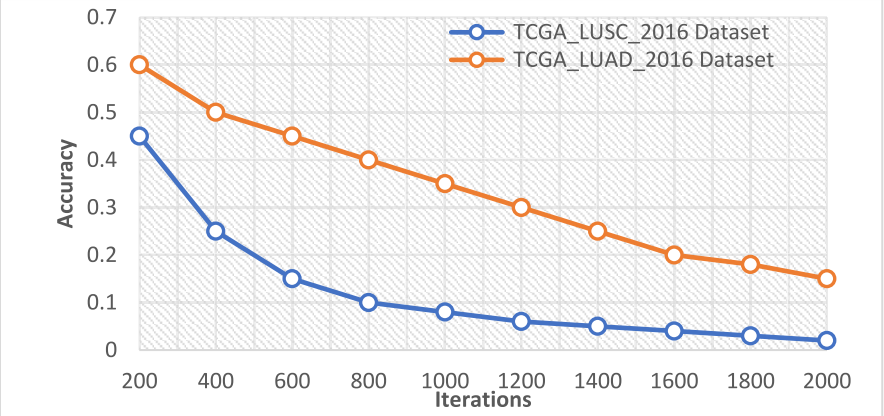

Fig. 3. Convergence vs. Iterations to study the stability of BBSA in feature selection process.

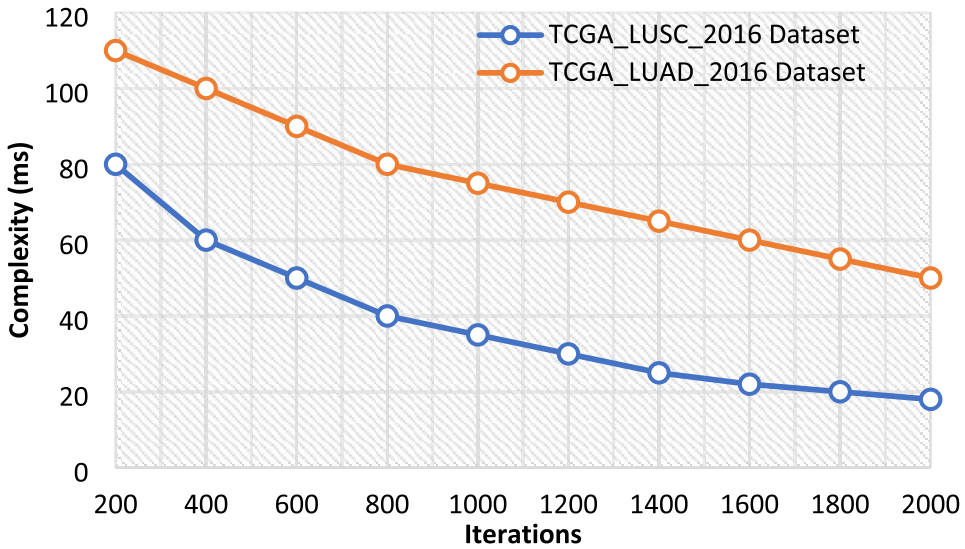

Fig. 4. Complexity vs. Iterations to study the stability of ODT structures in classification process.

Qualitative analysis

In this section, the proposed BBSA and OBT is analyzed qualitatively on biomedical lung cancer datasets. This includes algorithm behavior of BBSA and ODT, feature importance of BBSA, classification outcomes of ODT based on BBSA, robustness and computational cost of BBSA and OBT classifier.

Analysis on feature selection stability

Convergence Patterns in Fig. 3 shows the ability of BBSA to converge during the feature selection process. The convergence patterns are reported on two different datasets over 2000 iterations that includes TCGA_LUSC_2016 and TCGA_LUAD_2016. Both the datasets reported a decreasing trend in convergence over iterations, which shows BBSA convergence over time. The convergence rate for TCGA_LUSC_2016 is higher than TCGA_LUAD_2016, which shows that TCGA_LUAD_2016 encounters more complexity. The convergence rate reaches a relatively low level, which stabilizes the optimization process. The slow rate of convergence with increasing iterations further suggests BBSA approaches a stable solution.

Analysis on classification stability

This section analyses the decision tree structure, which analyze the structure of the ODT and it examines the decision boundaries for lung cancer classification on the two datasets: TCGA_LUSC_2016 and TCGA_LUAD_2016 as in Fig. 4. The computational complexity is measured in milliseconds, as it is the measure of time and memory consumption in forming the DT structure for accurate classification. The results shows that there exists a reducing complexity with increasing iterations, which improves the computational efficacy of ODT classification. The results indicate the formation of ODT structure for TCGA_LUSC_2016 is more

computationally demanding. The complexity stabilizes with increasing iterations, which suggests that the ODT reaches a point where the increasing iterations have minimal impacts on complexity.

Quantitative analysis

In this section, the biomedical lung cancer data is classified using the proposed MLSC method that combines BBSA feature selection and ODT classifier. It is then compared with existing state-of-art methods like M-BMIRC, TL-DLE, MMF, Uniform-Net, PulDi in terms of accuracy, precision, recall, f-measure, classification loss and computational efficiency.

Accuracy is defined as the ratio of total number of correctly predicted instances to the total number of instances (*TI*) as expressed in Eq. (15):

$$\text{Accuracy (AC)} = \frac{TP + TN}{TI} \tag{15}$$

Precision defines the ratio of predicted positive instances to the total predicted positives instances as expressed in Eq. (16):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

Recall denotes the ratio of positive observations that are correctly predicted to the overall observations in actual class as expressed in Eq. (17):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

F1-Score denotes the harmonic mean of precision and recall as expressed in Eq. (18):

$$F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

Computational Cost is defined as the computational resources (time and memory) required to train the MLSC model. Classification Loss is defined as the measures that defines the ability of the MSLC model in classifying the instances and it uses cross-entropy loss for binary classification as expressed in Eq. (19):

$$L = \frac{1}{N} \sum_{i=1}^{N} [y_i \log (p_i) + (1 - y_i) \log (1 - p_i)] \tag{19}$$

where *N*—number of instances, y_i—true label, and *pi*—predicted probability.

Precision is crucial metric for biomedical data analysis, where MLSC precision improvements contribute well to diagnostic accuracy. Figure 5 shows the accuracy of positive predictions and it minimizes the false positives

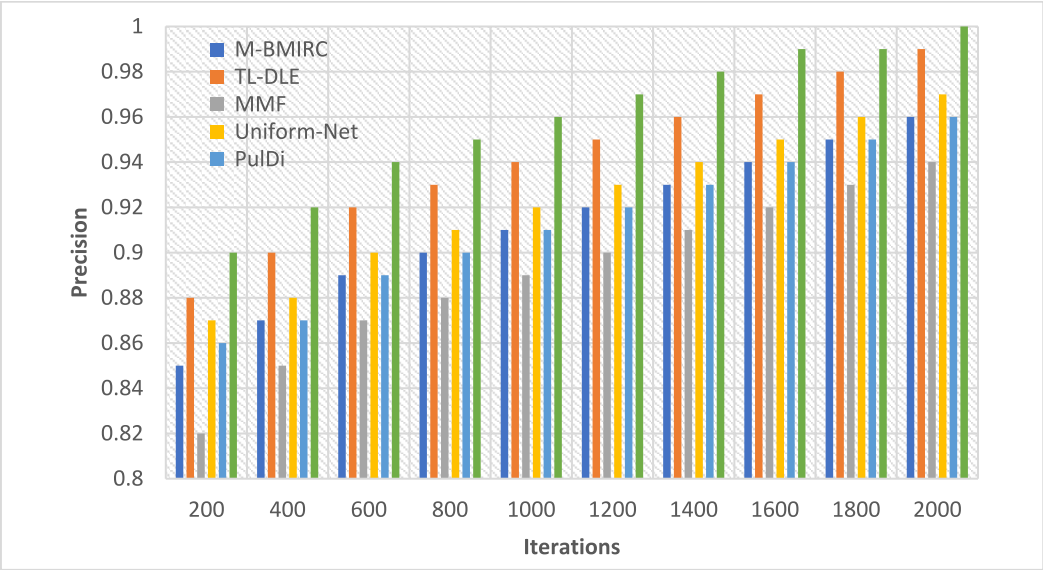

Fig. 5. Precision between the proposed MLSC and existing biomedical classifiers on lung cancer datasets.

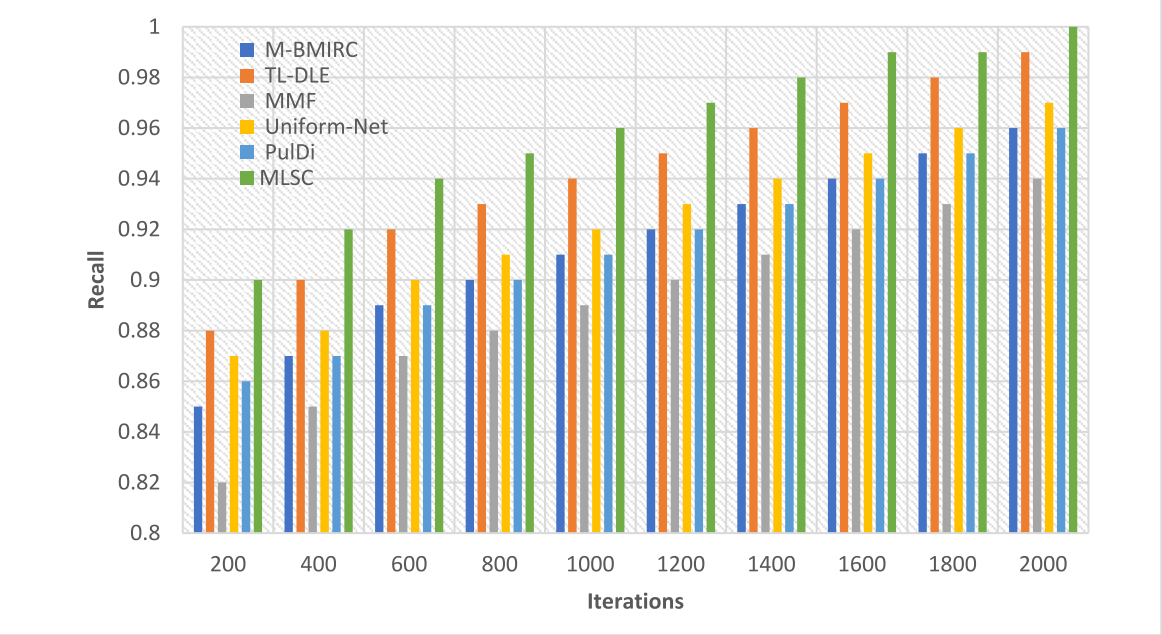

Fig. 6. Recall between the proposed MLSC and existing biomedical classifiers on lung cancer datasets.

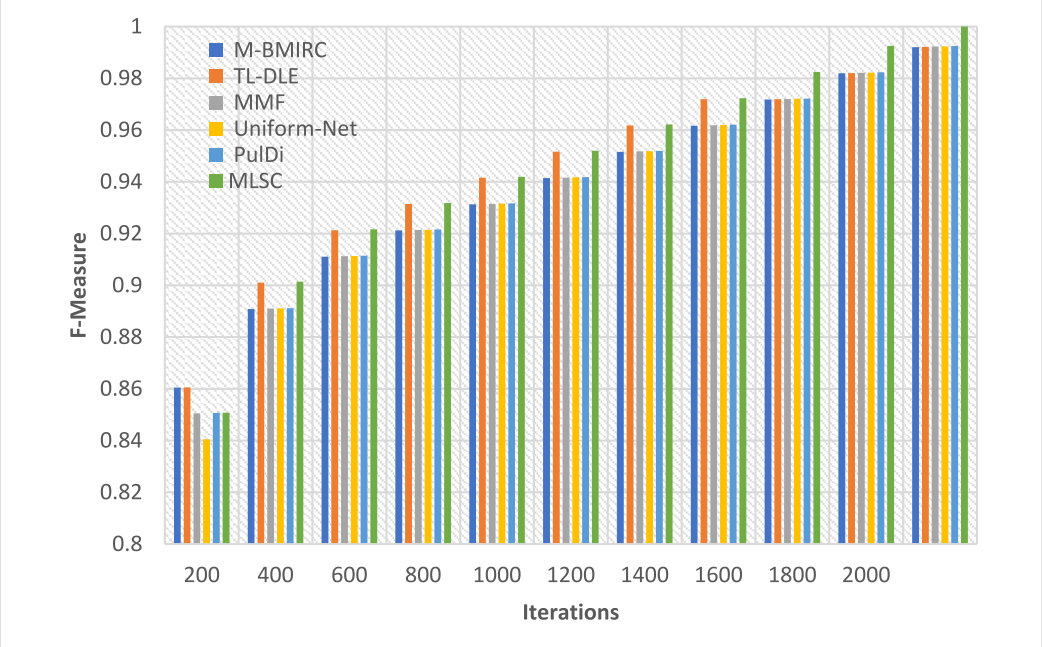

Fig. 7. F-Measure between the proposed MLSC and existing biomedical classifiers on lung cancer datasets.

using proposed MLSC. The proposed MLSC shows a higher precision rate compared to conventional methods, which shows its ability to reduce the minority sample misclassification rate.

Figure 6 shows the ability of MLSC to capture the positive instances and reducing or eliminating the false negatives. From the graph, it is found that the MLSC achieves higher recall rate, which shows its efficacy in identifying minority samples than the existing methods. However, it is found that the recall rate of TL-DLE method reports a marginal difference with the proposed method.

Figure 7 shows the results of balanced results with trade-off between the improved precision and recall rates. The Fig. 7 shows that the proposed MLSC method achieves higher F-Measure rate with increasing iterations. The balance acquired by MLSC is critical in biomedical datasets as it have significant implications on false positives and false negatives.

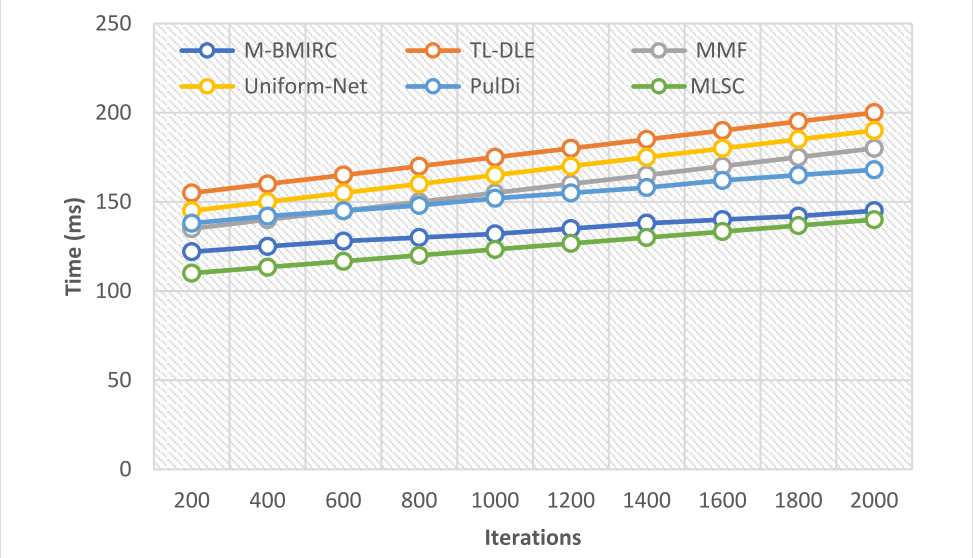

Fig. 8. Execution Time between the proposed MLSC and existing biomedical classifiers on lung cancer datasets.

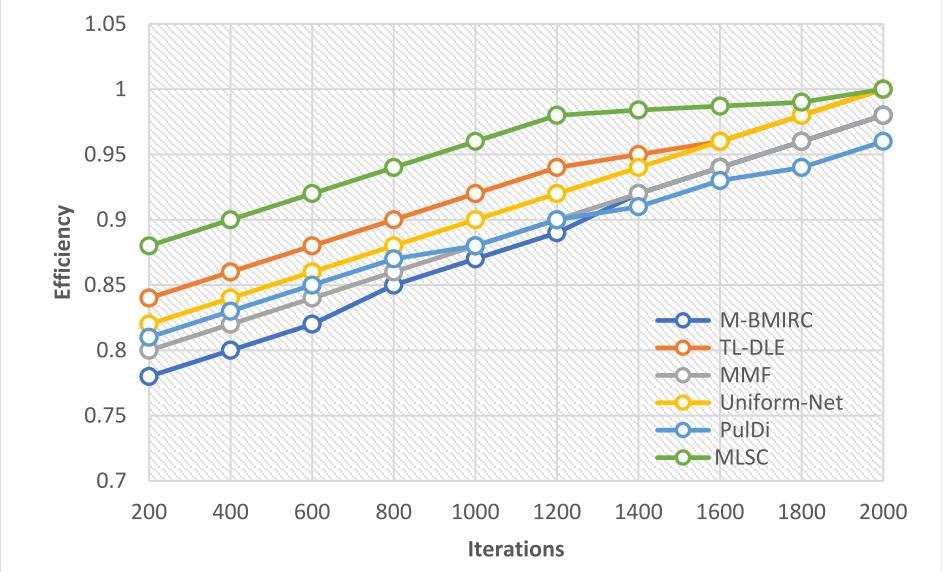

Fig. 9. Computational Efficiency between the proposed MLSC and existing biomedical classifiers on lung cancer datasets.

Figure 8 shows the execution time, which is the measure of the efficacy of proposed MLSC in terms of resource usage and training time. It is seen that the proposed method achieves a reduced execution time and this increases marginally with increasing iterations. However, the training time and its corresponding resource usage is reported to be high in existing methods.

From Fig. 9, it is seen that the efficiency quantifies the difference occurred between the predicted and true labels. The proposed MLSC shows an improvement in computational efficiency, which achieves higher accuracy while maintaining the reasonable execution rate. From the improvement of computational efficiency, the research reflects the ability of the model to deliver accurate results in a resource-efficient manner than the existing methods.

Figure 10 shows the results of accuracy to find the overall correctness of an imbalanced lung cancer datasets. The MLSC method outperformed consistently than its predecessors like M-BMIRC, TL-DLE, MMF, Uniform-Net, and PulDi. It shows a steady increase in this classification accuracy with increasing iterations. Thus, it is inferred that the MLSC performed with superior accuracy in handling the imbalanced, and high-dimensional

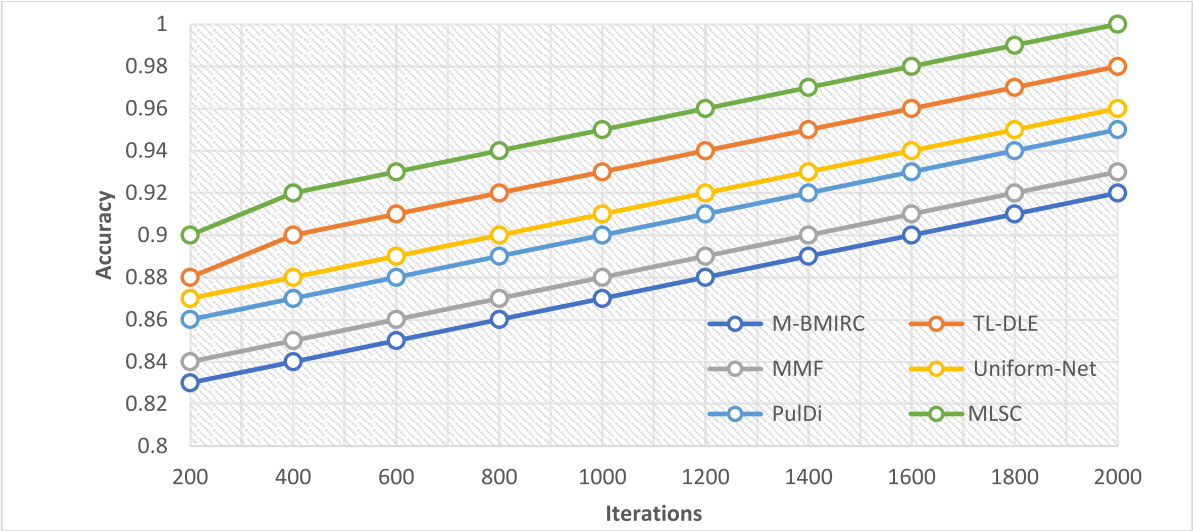

Fig. 10. Accuracy between the proposed MLSC and existing biomedical classifiers on lung cancer datasets.

dataset. It further reduces the problem of misclassification of minority samples with increasing iterations as reported in Fig. 10.

Lastly, to address interpretability concerns of oblique decision trees, our method limits tree depth and uses BBSA for selecting concise, clinically relevant features. While interpretability remains a challenge, future work will incorporate explainability tools and clinician feedback for practical adoption.

Conclusion

In this research, the proposed MLSC method leverages BBSA to normalize ODT for effective classification of biomedical lung cancer data. The MLSC resolves the problem of misclassification of minority samples in high-dimensional and imbalanced datasets. The features selected using BBSA construct the OBT in an adaptive manner that allows for oblique decision boundaries, enabling flexible and expressive modeling of complex relationships between the biomedical data. Thus, the MLSC optimization process ensures that the BBSA feature selection and ODT classification processes are fine-tuned for optimal classification of biomedical data. From the experimental results, it is found that the outcomes are promising when compared with state-of-the-art approaches. MLSC demonstrated optimal performance over accuracy, precision, recall, and F-measure while assessing its ability on biomedical lung cancer data. The MLSC achieves an average improvement of 4.6% in accuracy, 4.5% in precision, 7.5% in recall, and 7.5% in F-measure compared to its predecessors. Also, better computational efficiency is reported, which indicates the ability to achieve a favorable balance between accuracy and classification tasks with reduced computational cost. A reasonable execution time shows its enhanced capability in handling high-dimensional and imbalanced datasets and accurately identifying minority samples. The real-world applications of proposed work can be considered as the early detection of lung cancer using CT Scan data and personalized treatment planning-based on electronic health records. In the future, the research tends to explore the robust nature of deep learning swarm intelligence with oblique boundaries on diverse biomedical datasets and its scalability on complex scenarios and larger datasets. Several additional clinical factors such as patient demographics or tumor staging can be used, and proposed work can apply in several clinical tasks such as recurrence and survival prediction.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 5 January 2025; Accepted: 16 May 2025
Published online: 29 May 2025

References

1. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **6**(1), 1–25 (2019).
2. Su, Y., Shi, Q. & Wei, W. Single cell proteomics in biomedicine: High-dimensional data acquisition, visualization, and analysis. *Proteomics* **17**(3–4), 1600267 (2017).
3. Fernández, A., del Río, S., Chawla, N. V. & Herrera, F. An insight into imbalanced big data classification: Outcomes and challenges. *Complex Intell. Syst.* **3**, 105–120 (2017).
4. Ma, N. et al. Enhancing the sensitivity of spin-exchange relaxation-free magnetometers using phase-modulated pump light with external Gaussian noise. *Opt. Express* **32**(19), 33378–33390. https://doi.org/10.1364/OE.530764 (2024).
5. Dong, Q. & Jiang, Z. Platinum-iron nanoparticles for oxygen-enhanced sonodynamic tumor cell suppression. *Inorganics* **12**(12), 331. https://doi.org/10.3390/inorganics12120331 (2024).

6. Zhao, C., Song, W., Wang, J., Tang, X. & Jiang, Z. Immunoadjuvant-functionalized metal–organic frameworks: synthesis and applications in tumor immune modulation. *Chem. Commun.* **61**(10), 1962–1977. https://doi.org/10.1039/D4CC06510G (2025).
7. Jiang, C., Sun, T., Xiang, D., Wei, S. & Li, W. Anticancer activity and mechanism of xanthohumol: A prenylated flavonoid from hops (*Humulus lupulus* L.). *Front. Pharmacol.* **9**, 530. https://doi.org/10.3389/fphar.2018.00530 (2018).
8. Remark, R. et al. The non–small cell lung cancer immune contexture. A major determinant of tumor characteristics and patient outcome. *Am. J. Respiratory Critic. Care Med.* **191**(4), 377–390 (2015).
9. Senthilkumar, S. A., Rai, B. K., Meshram, A. A., Gunasekaran, A. & Chandrakumarmangalam, S. Big data in healthcare management: A review of literature. *Am. J. Theor. Appl. Bus.* **4**(2), 57–69 (2018).
10. Jain, P., Gyanchandani, M. & Khare, N. Big data privacy: A technological perspective and review. *J. Big Data* **3**, 1–25 (2016).
11. Grisci, B. I., Feltes, B. C. & Dorn, M. Neuroevolution as a tool for microarray gene expression pattern identification in cancer research. *J. Biomed. Inform.* **89**, 122–133 (2019).
12. Joshua, N., Stephen, E., Bhattacharyya, D., Chakkravarthy, M. & Kim, H. J. Lung cancer classification using squeeze and excitation convolutional neural networks with grad cam++ class activation function. *Traitement Du. Signal* **38**(4), 1103 (2021).
13. Bilgen, I., Guvercin, G. & Rekik, I. Machine learning methods for brain network classification: Application to autism diagnosis using cortical morphological networks. *J. Neurosci. Methods* **343**, 108799 (2020).
14. Rauschert, S., Raubenheimer, K., Melton, P. E. & Huang, R. C. Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification. *Clin. Epigenetics* **12**(1), 1–11 (2020).
15. Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J. & Shah, S. P. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* **22**(2), 114–126 (2022).
16. Guo, K. et al. Artificial intelligence-driven biomedical genomics. *Knowl. Based Syst.* **4**(279), 110937 (2023).
17. Alsinglawi, B. et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci. Rep.* **12**(1), 607 (2022).
18. Weiss, G. M. Foundations of imbalanced learning. *Imbalanced Learn. Found. Algorithms Appl.* **24**, 13–41 (2013).
19. Pashaei, E. An efficient binary sand cat swarm optimization for feature selection in high-dimensional biomedical data. *Bioengineering* **10**(10), 1123 (2023).
20. Dai, Y., Niu, L., Wei, L. & Tang, J. Feature selection in high dimensional biomedical data based on BF-SFLA. *Front. Neurosci.* **16**, 854685 (2022).
21. Chakraborty, C., Kishor, A. & Rodrigues, J. J. Novel enhanced-grey Wolf optimization hybrid machine learning technique for biomedical data computation. *Comput. Electr. Eng.* **99**, 107778 (2022).
22. Nagarajan, G. & DhineshBabu, L. D. A hybrid feature selection model based on improved squirrel search algorithm and rank aggregation using fuzzy techniques for biomedical data classification. *Netw. Model. Anal. Health Inform. Bioinform.* **10**(1), 39 (2021).
23. Hazarika, B. B. & Gupta, D. Random vector functional link with ε-insensitive Huber loss function for biomedical data classification. *Comput. Methods Programs Biomed.* **215**, 106622 (2022).
24. Anagaw, A. & Chang, Y. L. A new complement naïve Bayesian approach for biomedical data classification. *J. Ambient. Intell. Humaniz. Comput.* **10**, 3889–3897 (2019).
25. Deng, F., Huang, J., Yuan, X., Cheng, C. & Zhang, L. Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. *Lab. Invest.* **101**(4), 430–441 (2021).
26. Tang, R., & Zhang, X. (2020, May). CART decision tree combined with Boruta feature selection for medical data classification. In *2020 5th IEEE International Conference on Big Data Analytics (ICBDA)* (pp. 80–84). IEEE.
27. Shakeel, P. M. & Manogaran, G. Prostate cancer classification from prostate biomedical data using ant rough set algorithm with radial trained extreme learning neural network. *Heal. Technol.* **10**, 157–165 (2020).
28. Shakeel, P. M., Tolba, A., Al-Makhadmeh, Z. & Jaber, M. M. Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks. *Neural Comput. Appl.* **32**, 777–790 (2020).
29. Gautam, C. et al. Alzheimer's disease neuroimaging initiative. Minimum variance-embedded deep kernel regularized least squares method for one-class classification and its applications to biomedical data. *Neural Netw.* **1**(123), 191–216 (2020).
30. Gangavarapu, T. & Patil, N. A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets. *Appl. Soft Comput.* **81**, 105538 (2019).
31. Shrivastava, P., Shukla, A., Vepakomma, P., Bhansali, N. & Verma, K. A survey of nature-inspired algorithms for feature selection to identify Parkinson's disease. *Comput. Methods Programs Biomed.* **139**, 171–179 (2017).
32. TCGA_LUAD_2016, https://lce.biohpc.swmed.edu/lungcancer/datasetsearch.php?datasetid=61, Accessed on 01. 09. 2023
33. TCGA_LUAD_2016, https://lce.biohpc.swmed.edu/lungcancer/datasetsearch.php?datasetid=60, Accessed on 01. 09. 2023.

Author contributions

Shobha Aswal contributed as the main author who proposed the paper's main idea. Where Dr. Neelu Jyothi Ahuja (Supervisor) helped in stream lining the complete text and Dr. Ritika Mehra (Co-Supervisor) had a complete revision of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.A.

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.