scientific reports

OPEN

Greylag goose optimization and multilayer perceptron for enhancing lung cancer classification

El-Sayed M. Elkenawy[1], Amel Ali Alhussan[2], Doaa Sami Khafaga[2], Zahraa Tarek[3] & Ahmed M. Elshewey[4]✉

Lung cancer is an important global health problem, and it is defined by abnormal growth of the cells in the tissues of the lung, mostly leading to significant morbidity and mortality. Its timely identification and correct staging are very important for proper therapy and prognosis. Different computational methods have been used to enhance the precision of lung cancer classification, among which optimization algorithms such as Greylag Goose Optimization (GGO) are employed. These algorithms have the purpose of improving the performance of machine learning models that are presented with a large amount of complex data, selecting the most important features. As per lung cancer classification, data preparation is one of the most important steps, which contains the operations of scaling, normalization, and handling gap factor to ensure reasonable and reliable input data. In this domain, the use of GGO includes refining feature selection, which mainly focuses on enhancing the classification accuracy compared to other binary format optimization algorithms, like bSC, bMVO, bPSO, bWOA, bGWO, and bFOA. The efficiency of the bGGO algorithm in choosing the optimal features for improved classification accuracy is an indicator of the possible application of this method in the field of lung cancer diagnosis. The GGO achieved the highest accuracy with MLP model performance at 98.4%. The feature selection and classification results were assessed using statistical analysis, which utilized the Wilcoxon signed-rank test and ANOVA. The results were also accompanied by a set of graphical illustrations that ensured the adequacy and efficiency of the adopted hybrid method (GGO + MLP).

Keywords GGO, Optimization, Lung cancer, Feature selection, MLP, Classification

Lung cancer is a highly significant and fatal illness globally. According to the latest estimates from the World Health Organization (WHO), over 7.6 million fatalities occur annually globally as a result of lung cancer. Furthermore, the global incidence of cancer is projected to increase, reaching around 17 million cases by 2030[1]. Early detection of cancer is crucial since it has a tendency to metastasize and becomes incurable when it spreads extensively. Diagnosing lung cancer is challenging due to the manifestation of symptoms at the advanced stage, making it very difficult to achieve successful treatment outcomes, Fig. 1 shows Comparison between healthy lung and cancer lung. Imaging methods such as Computed Tomography (CT), Positron Emission Tomography (PET), Magnetic resonance imaging (MRI), and X-ray are used to take images of lungs for assessment. Lung cancer detection employed image processing and deep learning techniques where precision can be enhanced by implementing these methodologies. Detecting and determining the form, size, and location of a tumor is a challenging undertaking. Early detection is crucial for efficient time management, as it allows for necessary medical interventions[2]. Nevertheless, radiologists continue to have challenges in discerning between malignant and benign nodules. Distinguishing between malignant nodules and benign ones by visual inspection is subjective and the outcomes vary across various observers and instances. Typically, proficient radiologists have superior

[1]Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura 35111, Egypt. [2]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia. [3]Department of Computer Sciences, Faculty of Computers and Information, Mansoura University, Mansoura 35561, Egypt. [4]Department of Computer Science, Faculty of Computers and Information, Suez University, P.O.BOX:43221, Suez, Egypt. ✉email: ahmed.elshewey@fci.suezuni.edu.eg

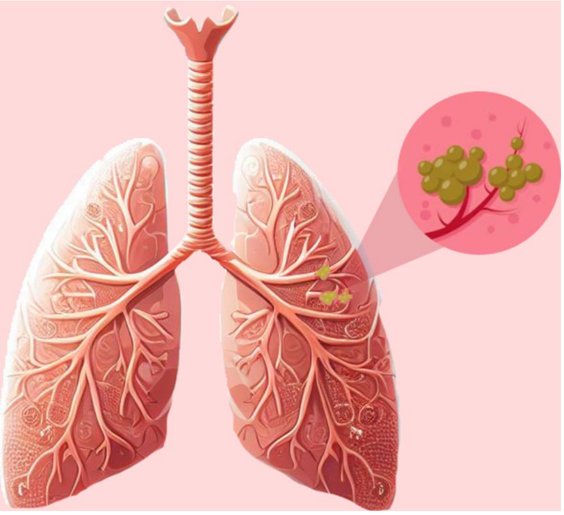


Fig. 1. Comparison between healthy lung and cancer lung.

accuracy in categorizing nodules compared to novice radiologists. The need for dependable and unbiased analysis has spurred the creation of computer-assisted systems[3].

The use of Machine Learning (ML) and Neural Networks (NN) in the categorization of Lung Cancer is not a new[4]. Employing machine learning techniques, such as optimizer and ensemble regression models, can be advantageous in attaining promising outcomes for both classification and regression problems[5]. Medical image recognition techniques can be applied as an initial diagnostic tool that hints about a potential illness[6]. Choosing appropriate features and building a highly efficient classifier for pulmonary nodules are crucial in the development of reliable Content-Based Image Retrieval (CBIR) and Computer-Aided (CAD) systems. Typically, CAD systems have two main stages: feature extraction and categorization. CBIR often incorporates a substantial number of visual elements, such as texture, form, and granulometry, in order to construct the search index[7]. Constructing machine learning architectures necessitates a proficient amalgamation of hyperparameters to enhance classification performance and precision. Tackling combinatorial issues with manual approaches can be daunting and hampers efficiency. Nevertheless, metaheuristic algorithms have been suggested as a means to optimize the process of acquiring the most optimal combination of hyperparameters needed for enhanced performance. Metaheuristic algorithms are optimization methods that draw inspiration from nature. They aim to discover optimal solutions characterized by local search, global search, and sometimes randomization. These algorithms are known for their great performance in finding acceptable optimization solutions. Swarm intelligence algorithms have effectively addressed intricate real-world challenges in engineering, medical sciences, and other scientific fields by achieving minimal computational power[8]. Metaheuristic algorithms are optimization approaches that employ iterative strategies to explore a vast search space and identify the optimal solution. The metaheuristic method is utilized to get the optimal combination of weights necessary for solving the feature extraction and classification challenge[9].

Greylag Goose Optimization (GGO) algorithm has demonstrated encouraging outcomes in many optimization problems, encompassing feature selection and parameter optimization in several sectors, such as healthcare, finance, and engineering. We chose GGO as a metaheuristic optimization technique because of its proven track record in comparable optimization problems and its distinctive characteristics that may offer benefits over alternative optimization approaches. The purpose of integrating ML methods with a metaheuristic algorithm is to optimize the performance of these approaches in terms of accuracy. Enhancing performance in illness detection, such as lung cancer, can lead to improved diagnostic precision and prompt treatment. In this paper, incorporating the GGO metaheuristic algorithm with MLP can enhance parameter optimization and increase its capacity to learn and categorize intricate patterns within the data. Consequently, this work seeks to integrate the GGO-MLP algorithm with specific preprocessing strategies in order to enhance the classification accuracy of lung cancer textual data. Initially, the input data is subjected to preprocessing, which involves scaling, normalization, and removal of null values. Following pre-processing, the subsequent task involves extracting the ideal collection of features that might augment the accuracy of lung cancer categorization. The GGO method is built using a binary format for the purpose of extraction to find the cancerous lung states. The subsequent step is categorizing the lung illness according to the extraction of distinctive features. The classification step utilizes many classifiers such as SVC, DTC, RFC, KNC, and MLP. The findings indicated that the Multilayer Perceptron (MLP) is the most optimal classifier. The hyperparameter of the MLP model is tuned using GGO, and the performance is evaluated against six other optimizers (SC, MVO, PSO, WOA, GWO, and FOA). The GGO with MLP model yielded the most optimal outcomes for the classification of lung cancer.

The specific enhancements proposed in our methodology are:

1. *Normalization and scaling* Data normalization and scaling were performed to standardize the input features. This step is necessary to ensure that all features contributed equally to the model training process and to improve the convergence speed of the optimization algorithms.
2. *Choosing optimization techniques* We used seven different binary optimization algorithms, Binary Greylag Goose Optimization (bGGO), Binary Sine Cosine Algorithm (bSC), Binary Mean Variance Optimization (bMVO), Binary Particle Swarm Optimizer (bPSO), Binary Whale Optimization Algorithm (bWOA), Binary Gray Wolf Optimizer (bGWO) and Binary Falcon Optimization Algorithm (bFOA). Each technique was used to identify the most relevant features, thus reducing the dimensionality of the data and improving classification performance.
3. *Feature selection using bGGO algorithm* Specifically, the bGGO algorithm was used to select the optimal subset of features. The goal was to remove irrelevant and redundant features, thus simplifying the input data and improving the efficiency and accuracy of the classification models.
4. *Choosing machine learning classifiers* We evaluated several machines learning classifiers, including Support Vector Classifier (SVC), Decision Tree (DT), Random Forest Classifier (RFC), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). This diverse selection allowed us to compare the performance of different models with the selected features.
5. *Parameter tuning using GGO* For MLP classifier, we proposed to use the GGO algorithm to optimize its parameters. The GGO algorithm iteratively adjusts the parameters to find the optimal configuration that gives the highest classification accuracy. The algorithm starts by generating a set of solutions, each representing a different set of MLP parameters. Each solution was evaluated based on its performance on the validation set and assigned a fitness score.
6. *Statistical analysis* The results of feature selection and classification were evaluated using statistical tests such as Wilcoxon signed-rank test and ANOVA test. These tests provided an accurate evaluation of the performance improvements achieved by our proposed method.

The structure of this paper is as follows: Section "Related work" provides an overview of relevant state-of-the-art literature. Section "Material and methods" provides an elaborate depiction of the proposed approach. The experimental results as well as discussion are presented in Section "The proposed framework". Section "Experimental results" presents the conclusions and future directions.

Related work

Due to the potential for increased survival rates, researchers are primarily focused on developing novel methods for the automated identification and diagnosis of significant lung nodules, as early detection is crucial in the case of lung cancer. This section presents research findings relating to lung cancer classification, based on ML/DL and textural/image analysis. Mohamed et al.[9] suggested a hybrid convolutional neural network (CNN) and metaheuristic algorithm. Initially, they constructed a CNN structure and subsequently calculated the solution vector of the model. The obtained solution vector was utilized in the Ebola optimization search algorithm (EOSA) to determine the optimal combination of weights and bias for training the CNN model to effectively address the classification challenge. Upon completing extensive training of the EOSA-CNN hybrid model, the researchers achieved the best configuration, resulting in good performance. The lung cancer dataset from the Iraq-Oncology Teaching Hospital/National Center for Cancer Diseases (IQ-OTH/NCCD), which is available to the public, was used for experimentation. The results indicated that the EOSA metaheuristic algorithm achieved a classification accuracy of 0.9321. In addition, they computed and reported the performance comparisons of EOSA-CNN with other approaches, including GA-CNN, MVO-CNN, LCBO-CNN, WOA-CNN, SBO-CNN, and the traditional CNN. The findings indicated that EOSA-CNN had a specificity of 0.7941, 0.9328, and 0.97951, as well as a sensitivity of 0.9038, 0.9071, and 0.13333 for normal, malignant, and benign cases, respectively. Ren et al.[10] introduced is a novel hybrid framework named LCGANT, including of two primary components. The initial component is a deep convolutional GAN (Generative Adversarial Network), which is capable of generating synthetic lung cancer images. The second component is a VGG-DF model, which incorporates regularization and transfer learning techniques, to accurately categorize lung cancer images into three distinct classifications. Their attained result for accuracy was 99.84% ± 0.156%, for precision was 99.84% ± 0.153%, for sensitivity was 99.84% ± 0.156%, and for F1-score was 99.84% ± 0.156%. Ananya Bhattacharjee et al.[11], outlined a precise approach for identifying cancerous nodules in computed tomography imaging. They utilized a combination of an optimized random forest classifier and a K-means visualization tool to achieve the most accurate findings. The model's hyperparameters were fine-tuned to ensure optimal performance, and the visualization tool was used to distinguish between malignant and non-malignant clusters. Among the four experiments conducted for hyperparameter optimization, the most successful model accurately identified cases as either malignant or non-malignant and obtained a 10-Fold cross-validation accuracy of 92.14% on the LIDC-IDRI dataset. Furthermore, the visualization configuration that yielded the best results achieved an inertia score of 16.21, which was the lowest, and a silhouette score of 0.815, which was the greatest. Vijh et al.[12] designed a pulmonary neoplasm diagnostic system utilizing a range of image processing methodologies. The Whale Optimization Algorithm (WOA) is utilized to pick the most optimal and conspicuous subset of features. They proposed the WOA_SVM, a novel approach for automating the diagnosis of lung CT images to determine their normality or abnormality. The suggested methodology's performance was assessed by measuring accuracy, sensitivity, and specificity. Utilizing a radial bias function support vector kernel, the results indicate an accuracy of 95%, sensitivity of 100%, and specificity of 92%. Shakeel et al.[13] examined biological characteristics associated with lung cancer through the application of neural and soft computing methodologies. The pertinent features were chosen based on the redundancy relevancy feature criterion and then inputted into the feature spaces and subjected to the Wolf prey searching method. The features with the highest fitness value were chosen and utilized as optimized ones to remove the weak classifiers. The testing characteristics were classified using the discrete AdaBoost optimized ensemble learning generalized neural network (DAELGNN), which effectively trained and classified the features into normal and lung cancer features. The system's efficiency was assessed by examining the testing and training characteristics, resulting in a minimum error rate of 0.0212 and a high prediction rate of 99.48%. Nanglia et al.[4] suggested technique combines a Support Vector Machine with a Feed-Forward Back Propagation Neural Network to develop a hybrid approach that effectively reduces the computational complexity of lung cancer classification. Their suggested KASC hybrid algorithm approach is implemented by combining the effective SURF feature extraction technology, the efficient optimization technique GA, and SVM with the polynomial kernel using FFBPNN. The hybridization approach yielded significantly improved performance metrics for lung cancer classification on a dataset of 500 CT images. Specifically, the accuracy, average precision, recall, and f-score achieved were 98.08%, 98.17%, 96.5%, and 97%, respectively. Nancy et al.[14] proposed machine learning may be used in the process of diagnosing lung cancer. They employed the CLAHE algorithm to enhance the overall image quality. The K Means approach is used when there is a requirement to segment an image. In this instance, feature selection is achieved through the utilization of the Particle Swarm Optimization (PSO) method. Subsequently, the images undergo categorization using the SVM, ANN, and KNN algorithms to classify them. Compared to other approaches, PSO-SVM has exceptional performance in accurately identifying instances of lung cancer, with an accuracy of 97.6% and a specificity of 99%. Table 1 displays a comparison with state-of-the-art for the most referenced studies.

Material and methods

Greylag Goose Optimization (GGO) Algorithm

The optimization approach presented in this study is referred to as the Greylag Goose Optimization (GGO) algorithm. The GGO method starts by producing a set of individuals in a random manner, as seen in Algorithm (1). Every participant presents a potential solution that might be considered as a candidate solution to the problem. The GGO population consists of individuals denoted as Y_i (i = 1, 2,..., n), where n is the total number of individuals in the population, as a gaggle. An objective function, denoted as F_n, is chosen to assess the individuals inside the group. By evaluating the objective function for every participant (agent) Y_i, the optimal

Refs	Methodology	Findings
Nanglia et al.[4]	SVM with Feed-Forward Back Propagation Neural Network and GA optimization	Accuracy: 98.08% Precision: 98.17% Recall: 96.5% F-Score: 97%
Mohamed et al.[9]	Hybrid CNN and Ebola Optimization Search Algorithm (EOSA)	Accuracy: 93.21% Specificity: - Normal: 79.41% - Malignant: 93.28% - Benign: 97.95% Sensitivity: - Normal: 90.38% - Malignant: 90.71% - Benign: 13.33%
Ren et al.[10]	Deep Convolutional GAN and VGG-DF model with regularization and transfer learning	Accuracy: 99.84% ± 0.156% Precision: 99.84% ± 0.153% Sensitivity: 99.84% ± 0.156% F1-Score: 99.84% ± 0.156%
Bhattacharjee et al.[11]	Random Forest classifier with K-means visualization	10-Fold Cross-Validation Accuracy: 92.14% Inertia Score: 16.21 (lowest) Silhouette Score: 0.815 (highest)
Vijh et al.[12]	Whale Optimization Algorithm with SVM	Accuracy: 95% Sensitivity: 100% Specificity: 92%
Shakeel et al.[13]	Discrete AdaBoost optimized ensemble learning generalized neural network	Error Rate: 0.0212 Prediction Rate: 99.48%
Nancy et al.[14]	Particle Swarm Optimization with SVM	Accuracy: 97.6% Specificity: 99%
Joshi et al.[15]	SMOCS (Cuckoo Search followed by Spider Monkey Optimization) and CSSMO (Spider Monkey Optimization followed by Cuckoo Search)	Accuracy: 100%
Saxena et al.[16]	A chaotic algorithm based on Marine Predator Algorithm (MPA) called Marine Predator Chaotic Algorithm (MPCA)	The proposed MPCA algorithm demonstrates the best graphical visualizations and statistical analysis
Yaqoob et al.[17]	Harris hawks optimization and cuckoo search algorithm (HHOCSA)	The proposed HHOCSA algorithm gives the best results when compared to another methods

Table 1. A comparison with state-of-the-art for several studies.

solution (leader) is determined and denoted as Z. The GGO algorithm utilizes the Dynamic Groups behavior to categorize individuals into two groups: an exploration group (n_1) and an exploitation group (n_2). The total number of solutions inside each group is dynamically controlled throughout each iteration based on the optimal solution. GGO commences the groups with an equal distribution of 50% exploration and 50% exploitation. Subsequently, the total number of agents in the exploration group (n_1) is reduced, while the number of agents in the exploitation group (n_2) is raised. However, if the objective function value of the best solution remains same for three consecutive iterations, the algorithm will begin to expand the number of agents in the exploration group (n1) in order to obtain a different best solution and hopefully avoid local optima[18]. Exploration has two important functions; it helps identify interesting regions inside the search space and prevents stagnation at local optima by advancing towards the optimal solution. Striving for optimal resolution by implementing this strategy, the geese explorer will actively seek out favorable new areas to investigate in close proximity to its present location. This is achieved by iteratively evaluating many potential adjacent choices in order to choose the optimal one based on its fitness. The GGO method employs the following equations to do this task, updating the B and D vectors as $B = 2b.m_1 - b$ and $D = 2.m_2$ throughout iterations with a parameter adjusted linearly from 2 to 0:

$$\mathrm{Y(t+1) = Y^{*}(t) - B.|D.Y^{*}(t) - Y(t)|} \tag{1}$$

where Y(t) presents an individual at iteration t. The $Y^{*}(t)$ denotes the optimal location of the leader (best solution). The Y(t + 1) represents the adjusted location of the individual. The values of m_1 and m_2 values are randomly changing within the range of 0 to 1.

The equation below will be employed by selecting three random search individuals, referred to as $Y_{Paddle1}$, $Y_{Paddle2}$, and $Y_{Paddle3}$, to ensure that the individuals are not influenced by a single leader position, hence promoting greater exploration. The current search agent's position will be changed as follows for |B| is greater than or equal to 1.

$$\mathrm{Y(t+1) = w_1 * Y_{Paddle1} + p * w_2 * (Y_{Paddle2} - Y_{Paddle3}) + \left(1 - p\right) * w_3 * (Y - Y_{Paddle1})} \tag{2}$$

where the values of w_1, w_2, and w_3 are adjusted within the range of 0 to 2. The values of w1, w2, and w3 are updated within the range of 0 to 2. The parameter p exhibits an exponential decrease and is determined by the following equation:

$$p = 1 - (\frac{t}{t_{max}})^{2} \tag{3}$$

where iteration number is denoted as "t", and "t_{max}" specifies the maximum number of iterations.

The second updating procedure is as follows for m_3 values greater than or equal to 0.5, where the values of b and B vectors values are reduced.

$$\mathrm{Y(t+1) = w_4 * |Y^{*}(t) - Y(t)| \cdot e^{al} \cdot \cos(2\pi l) + [2w_1(m_4 + \ m_5)] * Y^{*}(t)} \tag{4}$$

where a is a fixed value, l is a randomly selected value from range of – 1 to 1. The w4 parameter is adjusted within the range of 0 to 2, whereas m_4 and m_5 parameters are modified within the range of 0 to 1.

Exploitation operation

The exploitation group is responsible for enhancing the existing solutions. The GGO determines the individual with the best fitness at the end of each round and grants them corresponding recognition. The GGO utilizes two distinct strategies in order to accomplish its purpose of exploitation, which is elaborated below.

Moving towards the best solution

The subsequent equation is employed to advance towards the optimal solution. The three sentries (solutions), $Y_{S\,entry1}$, $Y_{S\,entry2}$, and $Y_{S\,entry3}$, direct other agents ($Y_{NonS\,entry}$) to adjust their locations towards the predicted location of the prey. The subsequent equations illustrate the updating locations procedure.

$$\begin{aligned} \mathrm{Y_1} &= \mathrm{Y_{S\,entry1} - B_1 \cdot \left|D_1 \cdot Y_{S\,entry1} - Y\right|}, \\ \mathrm{Y_2} &= \mathrm{Y_{S\,entry2} - B_2 \cdot \left|D_1 \cdot Y_{S\,entry2} - Y\right|}, \\ \mathrm{Y_3} &= \mathrm{Y_{S\,entry3} - B_3 \cdot \left|D_1 \cdot Y_{S\,entry3} - Y\right|} \end{aligned} \tag{5}$$

where the values of B_1, B_2, B_3 are determined by the equation $B = 2b.m_1 - b$, where b is a constant. Similarly, the values of D_1, D_2, D_3 are determined by the equation $D = 2m_2$. The modified population locations, Y(t + 1), can be calculated by taking the average of the three solutions: Y_1, Y_2, and Y_3 as follows.

$$\mathrm{Y(t+1)} = \overline{Yi}|_{0}^{3} \tag{6}$$

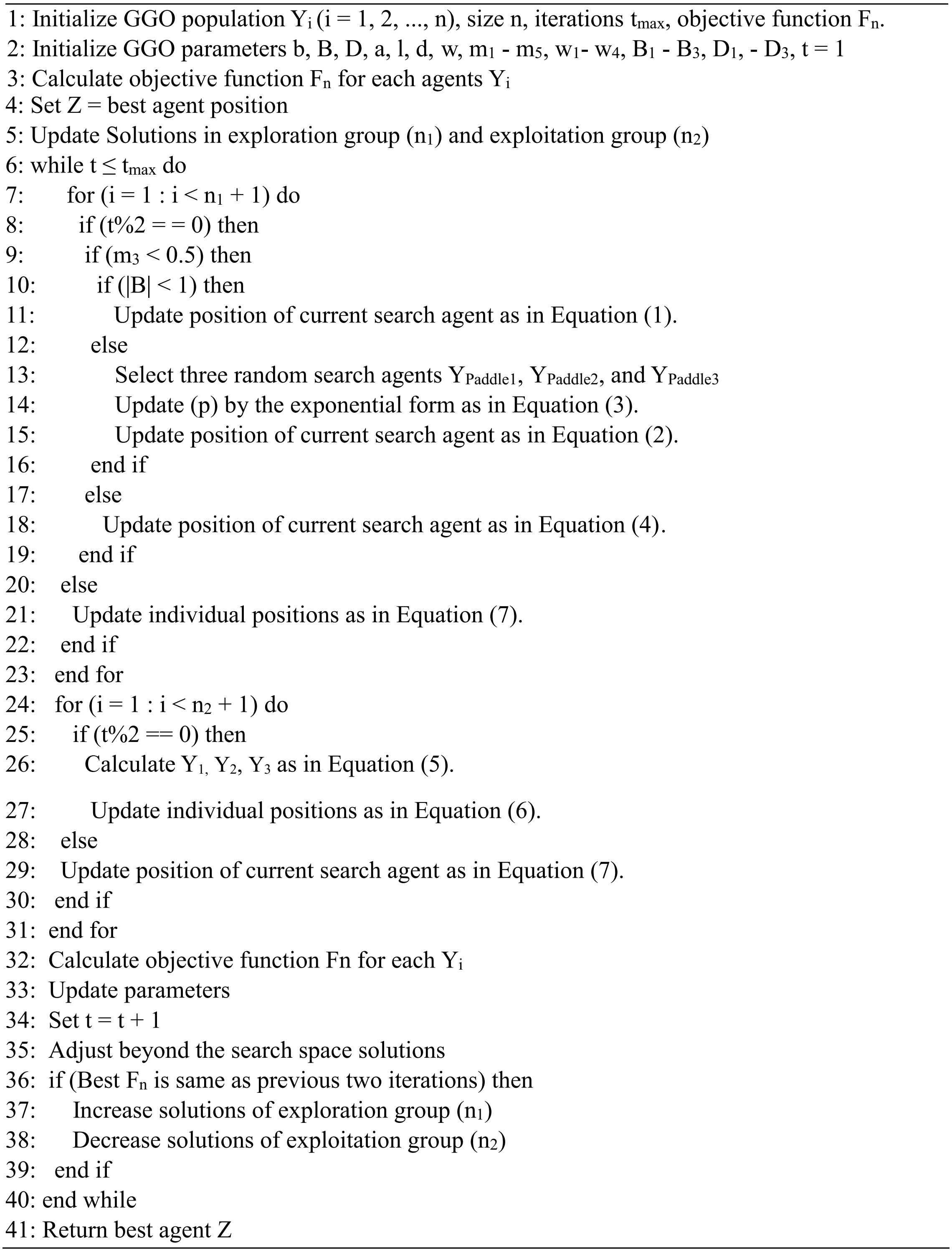
The Triangle Inequality is the second mathematical lemma that may be employed in the analysis of the GGO algorithm throughout the exploitation phase. The Triangle Inequality is a fundamental principle in metric spaces, asserting that the length of any side of a triangle is always less than or equal to the sum of the lengths of the other two sides. For GGO, this implies that the distance between any two individuals is less than or equal to the combined distances of these agents to a third individual in the search space. The most favorable option is situated in close proximity to the optimal response (leader) during flight. As a result, certain individuals are motivated to seek improvements by exploring areas adjacent to the optimal solution, referred to as YFlock1. In order to address the issue of local optima, the GGO employs the aforementioned procedure, which impacts both local and global optima, utilizing the subsequent equation.

$$\mathrm{Y}(\mathrm{t}+1)=\mathrm{Y}(\mathrm{t})+\mathrm{C}\big(1+\mathrm{p}\big)*\mathrm{w}*(\mathrm{Y}-\mathrm{Y}_{\mathrm{Flock1}}) \tag{7}$$

The Law of Large Numbers is the third mathematical concept that may be employed in the investigation of the GGO algorithm to effectively address the issue of local optima. The Law of Large Numbers is an essential concept in probability theory which asserts that as the size of the sample grows, the average of the sample approaches the average of the entire population. Within the framework of GGO, this implies that as the quantity of individuals in the swarm grows, the swarm as a collective will progressively approach the global optimum.

The GGO provides outstanding exploration capabilities by employing a mutation approach and scanning participants of the exploration group. The GGO has the ability to delay convergence due to its robust exploring abilities. Algorithm 1 includes the GGO pseudo-code, as we Initially provide GGO with essential facts, including the size of the population, rate of mutation, and number of iterations. The GGO thereafter divides the participants into two groups: those engaged in exploratory work and the other involved in exploitative labor. The GGO approach adjusts the size of each group while it iteratively searches for the optimal solution. In order to introduce diversity and conduct thorough investigation, the GGO employs a random reordering of the responses throughout each iteration.

Within a single iteration, a solution parameter from the exploration group may transfer to the exploitation group in the manner described below. The GGO's exclusive approach ensures the leader's continuous tenure in position during the process. The GGO algorithm seeks to update the locations of both the exploration group (n_1) and the exploitation group (n_2). The parameter m1 is modified iteratively according to the equation $\mathrm{m}_1 = d(1-\frac{t}{t_{max}})$, where t represents current iteration, d denotes a constant, and t_{max} specifies the number of iterations. After each iteration, GGO modifies the individuals in the search space and randomly shuffles their order to shift their roles between the exploration and exploitation groups. During the last stage, GGO retrieves the optimal solution.



```
Initialize GGO population $Y_i$ ($i = 1, 2, ..., n$), size n, iterations $t_{max}$, objective function $F_n$.
Initialize GGO parameters b, B, D, a, l, d, w, $m_1$ - $m_5$, $w_1$- $w_4$, $B_1$ - $B_3$, $D_1$, - $D_3$, $t = 1$
Calculate objective function $F_n$ for each agents $Y_i$
Set Z = best agent position
Update Solutions in exploration group ($n_1$) and exploitation group ($n_2$)
while $t \leq t_{max}$ do
    for ($i = 1 : i < n_1 + 1$) do
        if ($t\%2 = = 0$) then
            if ($m_3 < 0.5$) then
                if ($|B| < 1$) then
                    Update position of current search agent as in Equation (1).
                else
                    Select three random search agents $Y_{Paddle1}$, $Y_{Paddle2}$, and $Y_{Paddle3}$
                    Update (p) by the exponential form as in Equation (3).
                    Update position of current search agent as in Equation (2).
                end if
            else
                Update position of current search agent as in Equation (4).
            end if
        else
            Update individual positions as in Equation (7).
        end if
    end for
    for ($i = 1 : i < n_2 + 1$) do
        if ($t\%2 == 0$) then
            Calculate $Y_{1,}$ $Y_2$, $Y_3$ as in Equation (5).
            Update individual positions as in Equation (6).
        else
            Update position of current search agent as in Equation (7).
        end if
    end for
    Calculate objective function Fn for each $Y_i$
    Update parameters
    Set $t = t + 1$
    Adjust beyond the search space solutions
    if (Best $F_n$ is same as previous two iterations) then
        Increase solutions of exploration group ($n_1$)
        Decrease solutions of exploitation group ($n_2$)
    end if
end while
Return best agent Z
```

Algorithm 1: GGO algorithm.

Binary GGO Optimization Algorithm

The GGO optimization algorithm is a more effective approach for improving the feature selection of MLP parameters. The GGO utilizes the binary format for feature selection.

Feature selection problems revolve around a restricted search space that just encompasses binary values of 0 and 1. The objective is to determine the significance of a particular feature. Therefore, the ongoing GGO values are converted to a binary [0, 1] format in the binary GGO technique suggested in this section, in order to align with the feature selection process. The fundamental purpose of this procedure, as outlined in Eqs. (8, 9), is to convert the continuous data into binary data using the following Sigmoid function.

$$Bi_t^* = \begin{cases} 1 & \textit{if sigmoid } \left(Bi_t^*\right) \geq 0.5 \\ 0 & \textit{otherwise} \end{cases} \tag{8}$$

$$Sigmoid\left(Bi_t^*\right) = \frac{1}{1 + e^{-10\left(Bi_*^i - 0.5\right)}} \tag{9}$$

where Bi_t^* represents the best solution at specific iteration t. Algorithm 2 presents the proposed bGGO stages utilized to choose the optimal feature set, enhancing the Caries' case classification accuracy.

```
Initialize GGO population, objective function, and GGO parameters
Convert solution to binary [0 or 1]
Calculate objective function for each agent and get best agent position
Update Solutions in exploration group and exploitation group
while $t \leq t_{max}$ do
    for ($i = 1 : i < n_1 + 1$) do
        if (t%2 == 0) then
            if ($m_3 < 0.5$) then
                if ($|B| < 1$) then
                    Update position of current search agent in exploration group
                else
                    Update position of current search agent based on three random search agents
                end if
            else
                Update position of current search agent
            end if
        else
            Update individual positions
        end if
    end for
    for ($i = 1 : i < n2 + 1$) do
        if (t%2 == 0) then
            Update position of current search agent in exploitation group
        else
            Update position of current search agent
        end if
    end for
    Convert updated solution to binary
    Calculate objective function
    Update parameters
    Adjust beyond the search space solutions
    Update Solutions in exploration group and exploitation group
end while
Return best agent
```

Algorithm 2: bGGO algorithm.

Multilayer Perceptron (MLP) Model

Multilayer perceptron models, derived from the neural architecture of the human brain, have significant aptitude in representing the nonlinear behavior of intricate systems. Moreover, these models provide the capability to tackle prediction problems with nonlinear structure. This model functions by acquiring knowledge about the problem-solving procedure in order to get the desired outcome by identifying the underlying relationship inside the process. In order to achieve this objective, a substantial amount of data is utilized during the training phase, leveraging the discovered relationship in that phase to compute the desired output. Among several neural network models, the back-propagation network is the most commonly utilized. This network is composed of layers, each containing parallel-acting units known as neurons. Each layer is fully interconnected with the preceding and subsequent layers[19].

The proposed framework

In this paper, the initial phase involves data processing, which includes removing null values, normalizing, and scaling data. The primary focus in this stage is to prepare and expand the input data. This study employed feature selection approaches to carry out 7 optimization techniques in binary format- namely, Greylag Goose Optimization (GGO)[18], Sine Cosine Algorithm (SC)[20], Mean variance optimization (MVO)[21], Particle Swarm Optimizer (PSO)[22], Whale Optimization Algorithm (WOA)[23], Gray Wolf Optimizer (GWO)[24], and Falcon Optimization Algorithm (FOA)[25]. The second phase involves the process of selecting features, using the proposed feature selection approach. The pertinent features are subsequently extracted using bGGO. The objective of this step is to determine the optimal characteristics that will enable precise classification of the input data. This step offers the advantage of diminishing the overall quantity of features by deleting irrelevant ones. The input data was classified using ML classifiers, with features picked based on the bGGO. The ML models proposed in this study include the Support Vector Classifier (SVC)[26], Decision Tree (DT)[27], Random Forest Classifier (RFC)[28], 1 K-Nearest Neighbors (KNN)[29], and multilayer perceptron (MLP)[30]. The parameters of the MLP are optimized to effectively utilize the proposed optimization technique. The objective of this stage is to choose the optimal set of classification parameters. The algorithm begins by producing a collection of solutions for different parameter setups. Each key, or individual Greylag Goose, is allocated a fitness score depending on its performance on a validation set. In order to identify the optimal solutions, individuals within the population are systematically traversed over the search space. GGO utilizes weighted vectors to direct individuals to suitable destinations. These vectors are calculated using the fitness ratings of population. The algorithm consistently adjusts the placements of the participants to accurately reflect the optimal solution. GGO employs an iterative process to progressively improve the solutions with the goal of reaching the ideal configuration of MLP parameters. Once the convergence threshold is attained, the algorithm terminates after a specific number of iterations. In this case, the optimal solution is chosen based on the parameter configuration with the highest fitness value. This research explores how GGO can enhance the tuning parameters of MLP (Multilayer Perceptron). The sequential process of the proposed framework is visually depicted in Fig. 2.

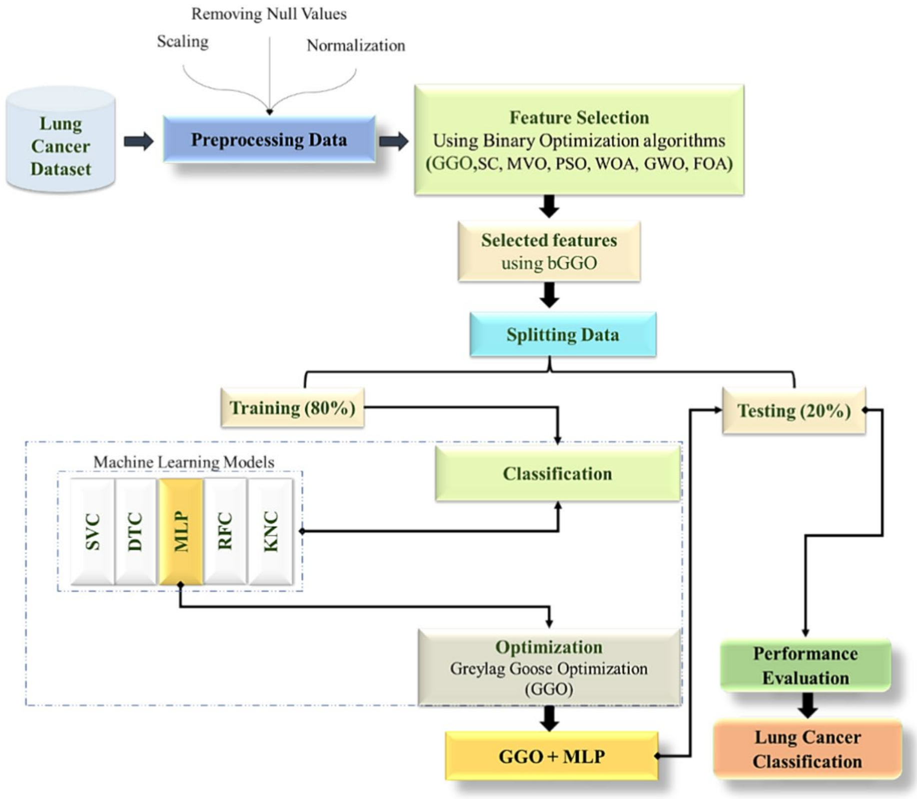


Fig. 2. The proposed lung cancer classification framework.

Experimental results

This section describes the assessment of the proposed algorithm in several experimental conditions. The tests employed traditional mathematical functions as reference points to ascertain their minimum values within a designated search area. These functions are frequently utilized in the previous research to evaluate the effectiveness of optimization strategies, and there are several optimization methods available in the literature. This study conducted a comparison analysis to demonstrate the higher performance and effectiveness of the proposed algorithm, Greylag Goose Optimization (GGO), compared to six known optimization techniques. The algorithms GGO, SC, MVO, PSO, WOA, GWO, and FOA were chosen because to their extensive acknowledgment and practical significance. The study platform employed the following technical specifications: a primary memory of 16 GB, an Intel Core i7 CPU, anda graphics processing unit (GPU) utilizing a GeForce RTX2070 Super with 8 GB of RAM. On the other hand, the software specifications were made up of Ubuntu 20.04 as the operating system, TensorFlow 1.15, CUDA9.0, Cudnn7.1, and Python 3.7 for the Spider Integrated Development Environment (IDE).

Dataset description

Machine learning and data science specialists may utilize this dataset to construct prediction models for diagnosing lung cancer, examine the influence of different cancer-related characteristics, and devise algorithms to enhance cancer therapy and prevention. This research uses a data set termed "Lung Cancer Dataset", which was gathered and submitted to Kaggle. The efficacy of cancer classification and prediction systems enables individuals to ascertain their susceptibility to cancer at a minimal expense. Moreover, it aids specialists in making informed decisions in accordance with their cancer risk profile. The data is gathered from the online lung cancer prediction system on the website, available at: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

The input features of the used dataset consist of Gender, Smoking, Age, Yellow fingers, Peer_pressure, Anxiety, Chronic Disease, Swallowing Difficulty, Allergy, Fatigue, Wheezing, Alcohol, Shortness of Breath, Coughing, and Chest pain. The attributes count is 16, and there are 284 occurrences. These variables are used to classify the output variable, which is Lung Cancer. Figure 3 depicts a scatter plot that showcases the relationship between the input and output variables of the Lung Cancer dataset.

Figure 4 exhibits the correlation matrix, which is a helpful statistical instrument for analyzing the relationship between variables in the dataset. It often generates a matrix that shows the pairwise correlations between all variables. The correlation coefficients, which span from − 1 to + 1, signify the relative significance and direction of the interactions. In order to examine the relationships, patterns, and potential predictors within the data, we may employ a correlation matrix to discover variables that have positive or negative correlations. This information is crucial for predictive modeling as it aids in the selection of pertinent characteristics, the reduction of dimensionality, and the identification of multicollinearity difficulties[31].

Feature selection results

This study employed feature selection techniques to carry out seven optimization algorithms in binary format, specifically: Greylag Goose Optimization (GGO), Gray Wolf Optimizer (GWO), Mean variance optimization (MVO), Whale Optimization Algorithm (WOA), Sine Cosine Algorithm (SC), Particle Swarm Optimizer (PSO),

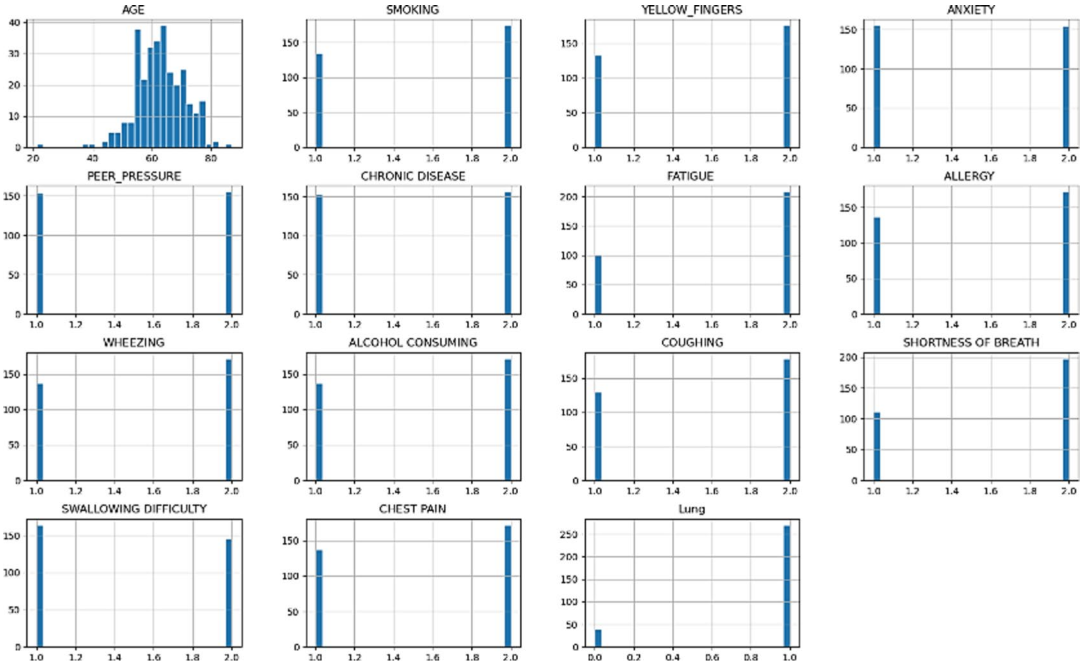


Fig. 3. Scatter plot for each feature in the dataset.

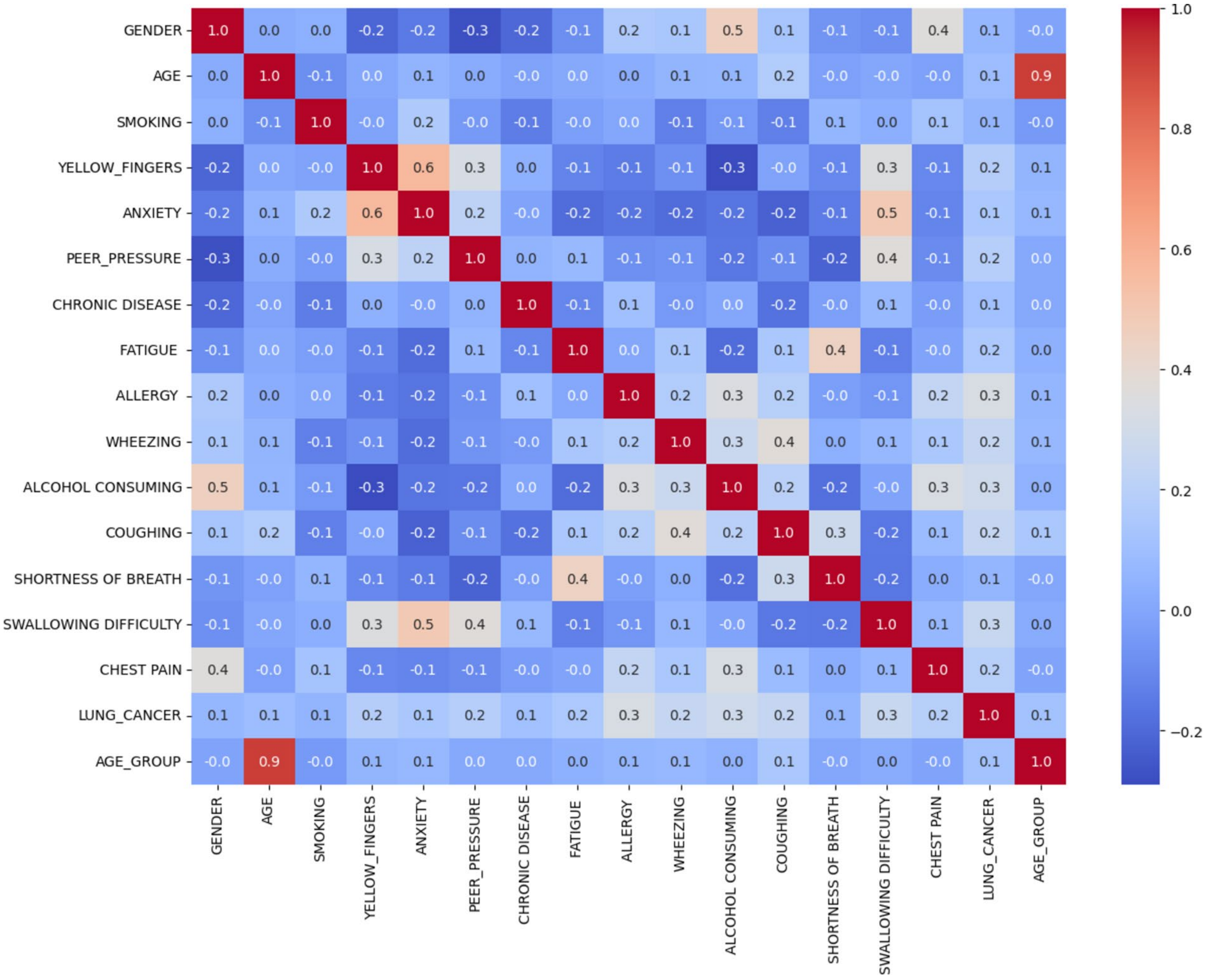


Fig. 4. A correlation matrix between features in the dataset.

	bGGO	bSC	bMVO	bPSO	bWOA	bGWO	bFOA
Average error	0.7743031	0.791503	0.805103	0.825303	0.825103	0.811603	0.823703
Average Select size	0.7271031	0.927103	0.869503	0.927103	1.090503	0.849903	0.961603
Average Fitness	0.8375031	0.853703	0.865103	0.852103	0.859903	0.859803	0.904003
Best Fitness	0.7393031	0.774003	0.768403	0.832403	0.824003	0.837603	0.822703
Worst Fitness	0.8378031	0.840903	0.883503	0.900103	0.900103	0.913803	0.920303
Standard deviation Fitness	0.6598031	0.664503	0.666103	0.663903	0.666103	0.665103	0.700703

Table 2. Evaluation of the suggested feature selection technique (bGGO) in comparison to other competitive techniques.

and Falcon Optimization Algorithm (FOA). Table 2 presents an evaluation of the outcomes attained by different feature selection techniques. The table clearly illustrates that the proposed bGGO achieved the best results against the other binary feature selection algorithms in the term of average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness. Lower values for average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness indicate the best results. The average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness values are 0.7743031, 0.7271031, 0.8375031, 0.7393031, 0.8378031 and 0.6598031, respectively.

The Fig. 5 displays a graph that compares the average error of the proposed feature selection technique with six other feature selection strategies. The bGGO technique has the lowest average error, so showcasing its robustness, as shown in this figure.

Table 3 illustrates the superior performance of the proposed bGGO technique compared to previous feature selection strategies across many measures. The p-values were computed by comparing the outcomes of each

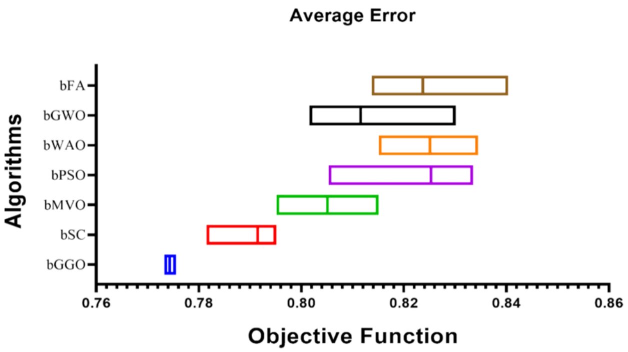


Fig. 5. The Average Error of the Results Acquired using bGGO, the Proposed Feature Selection Technique.

	bGGO	bSC	bMVO	bPSO	bWOA	bGWO	bFOA
Theoretical median	0	0	0	0	0	0	0
Actual median	0.7743	0.7915	0.8051	0.8253	0.8251	0.8116	0.8237
Number of values	10	10	10	10	10	10	10
Wilcoxon Signed Rank Test							
Sum of signed ranks (W)	55	55	55	55	55	55	55
Sum of positive ranks	55	55	55	55	55	55	55
Sum of negative ranks	0	0	0	0	0	0	0
P value (two tailed)	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Exact or estimate?	Exact	Exact	Exact	Exact	Exact	Exact	Exact
P value summary	**	**	**	**	**	**	**
Significant (alpha = 0.05)?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
How big is the discrepancy?							
Discrepancy	0.7743	0.7915	0.8051	0.8253	0.8251	0.8116	0.8237

Table 3. The Wilcoxon signed-rank test for evaluating the effectiveness of the proposed feature selection technique (bGGO) in comparison to existing binary optimization techniques.

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Treatment (between columns)	0.02274	6	0.00379	F (6, 63) = 135.3	P < 0.0001
Residual (within columns)	0.001764	63	2.8E−05		
Total	0.02451	69			

Table 4. The analysis-of-variance (ANOVA) test for assessing the proposed technique, bGGO.

algorithm pair, revealing that the proposed feature selection technique (bGGO) exhibits statistically significant superiority. The null hypothesis and the alternative hypothesis are the primary hypotheses under consideration in this study. The null hypothesis, shown as H0, assumes that the mean values (m) of bGGO are equal to the mean values of bGWO, bPSO, bWOA, bSC, bMVO, and bFOA. However, the H1 hypothesis does not consider the averages of the algorithms. To conduct this inquiry, the Wilcoxon rank-sum test was employed. The results of the Wilcoxon rank-sum test are presented in Table 3. The proposed technique has statistical superiority over previous techniques, as seen by its lower p-value ($p < 0.005$). A one-way analysis of variance (ANOVA) test was performed to see if there were statistically significant differences between the proposed bGGO technique and the other binary optimization techniques. The results of the ANOVA test are presented in Table 4. The results shown in these tables confirm the superiority, significance, and effectiveness of the proposed feature selection technique.

The plots in Fig. 6 illustrate the results achieved by the proposed feature selection technique. The figure employs residual plots, quartile–quartile (QQ), homoscedasticity, and heatmap to showcase the effectiveness and reliability of the proposed technique. The results shown in the QQ plot demonstrate a strong correlation with a linear trend, revealing that the selected features are reliable in correctly determining the presence of lung cancer. Moreover, the focus on results is further underscored by the reported outcomes in the homoscedasticity and residual plots. Additionally, the superiority of the proposed technique is further evidenced by the heatmap depicted in Fig. 5, which clearly illustrates that the proposed technique surpassed the other six binary feature

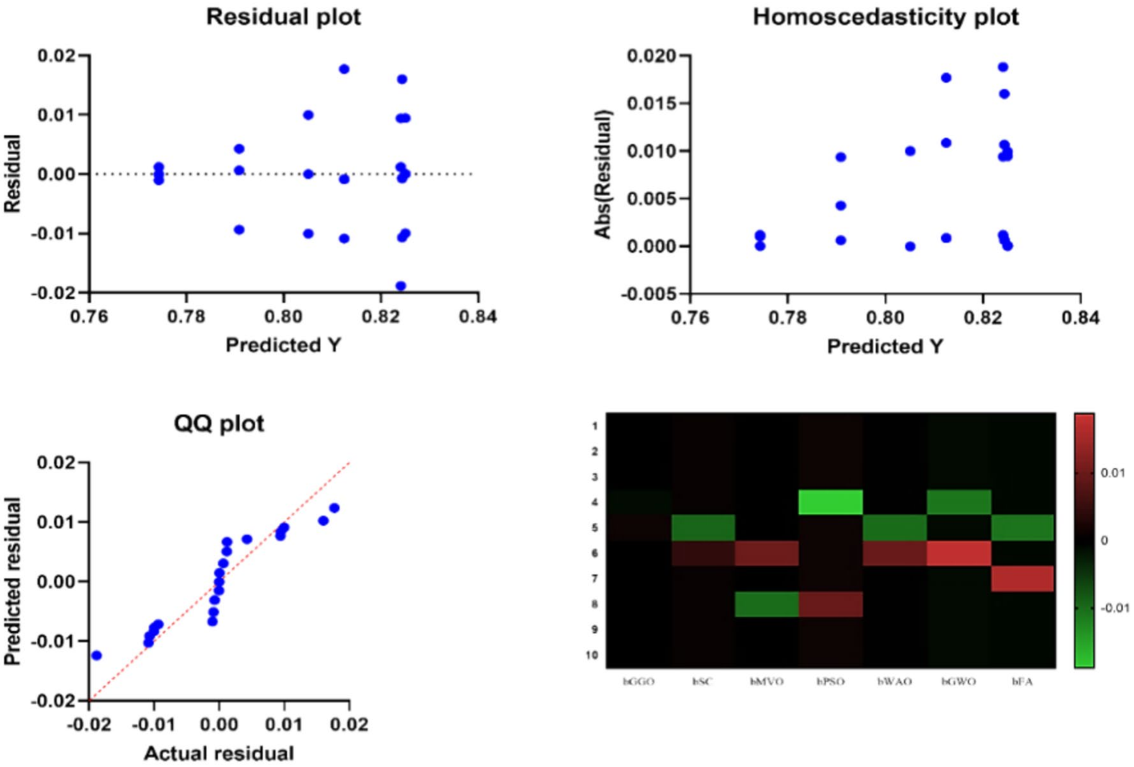


Fig. 6. Analysis plots of the obtained outcomes based on bGGO, the proposed feature selection technique.

Models	Accuracy	Sensitivity (TRP)	Specificity (TNP)	Pvalue (PPV)	Nvalue (NPV)	F-Score	AUC
SVC	0.83389544	0.837171	0.83045	0.83855	0.829016	0.83786	0.874
DT	0.83808724	0.842715	0.833333	0.83855	0.837607	0.84062	0.876
RFC	0.86343993	0.877586	0.84922	0.854027	0.87344	0.86564	0.893
KNN	0.8872	0.877586	0.895522	0.879102	0.894188	0.87834	0.882
MLP	0.91803278	0.926686	0.909091	0.913295	0.923077	0.91994	0.935

Table 5. Various classifiers for the categorization of lung cancer.

selection strategies. The effectiveness of the bGGO technique is confirmed by the heatmap, as it achieved the most optimal results compared to other feature selection strategies.

Classification results

A supplementary experiment was conducted to showcase the influence of the feature selection technique on the classification findings. Machine learning classifiers were employed to categorize the input data based on characteristics chosen according to the bGGO. The bGGO technique was employed to enhance the network's characteristics with the aim of optimizing performance. The classification outcomes for several machine learning

Models	Accuracy	Sensitivity (TRP)	Specificity (TNP)	Pvalue (PPV)	Nvalue (NPV)	F-Score	AUC
GGO + MLP	0.983837	0.977337	0.990237	0.989957	0.977961	0.983607	0.993
SC + MLP	0.966184	0.966387	0.965986	0.965035	0.967302	0.96571	0.975
GWO + MLP	0.960879	0.961003	0.960758	0.959666	0.96206	0.960334	0.972
PSO + MLP	0.951087	0.949106	0.95302	0.951724	0.950469	0.950413	0.964
WOA + MLP	0.94086	0.949106	0.932983	0.931174	0.950469	0.940054	0.957
FOA + MLP	0.937287	0.943912	0.930707	0.931174	0.943526	0.9375	0.953
MVO + MLP	0.927978	0.933694	0.921986	0.926174	0.9299	0.929919	0.937

Table 6. Findings of optimization methods MLP model for the classifying lung cancer.

models following feature selection are displayed in Table 5. The ML models listed in this table include the Support Vector Classifier (SVC), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), K Neighbors Classifier (KNC), and multilayer perceptron (MLP). The MLP model achieved the greatest values of 0.9180327, 0.92668, 0.9091, 0.9133, 0.9231, 0.9199 and 0.935 for accuracy, sensitivity, specificity, p-value, n-value, F-score and Area Under the ROC Curve (AUC) respectively. The MLP model acts as a fitness function and is optimized using the GGO algorithm and six additional optimization techniques. The outcomes for classification of the seven optimization algorithms, using the fitness function of the MLP model, are presented in Table 6. The findings of GGO combined with MLP are compared to those of SC, GWO, PSO, WOA, FOA, MVO with MLP to demonstrate the superiority of the proposed approach (GGO + MLP). The GGO + MLP approach demonstrated superior performance with accuracy, sensitivity, specificity, p-value, n-value F-score and Area Under the ROC Curve (AUC) with values of 0.983837, 0.977337, 0.990237, 0.989957, 0.977961, 0.983607 and 0.993, respectively. Various optimizers were employed to refine the parameters of the MLP, and the outcomes were investigated and assessed. The data shown in this table clearly demonstrates that the proposed approach surpasses the previous optimization approaches. These findings clearly show the necessity of feature selection in order to enhance the accuracy of the classification results.

Table 7 demonstrates the parameter settings for Greylag Goose Optimization (GGO), Gray Wolf Optimizer (GWO), Mean variance optimization (MVO), Whale Optimization Algorithm (WOA), Sine Cosine Algorithm (SC), Particle Swarm Optimizer (PSO), and Falcon Optimization Algorithm (FOA).

Algorithm	Parameter	Values
GGO	Population size	50
	Inertia weight	0.9
	Iterations	100
WOA	Population size	200
	Iterations	100
GWO	Iterations	100
	Wolves	20
FOA	Iterations	100
	Population size	150
SC	Initial amplitude	2
	Iterations	100
PSO	Particles	80
	Iterations	100
MVO	Aversion parameter (λ)	0.6
	Iterations	100

Table 7. Parameter settings for the algorithms used in this study.

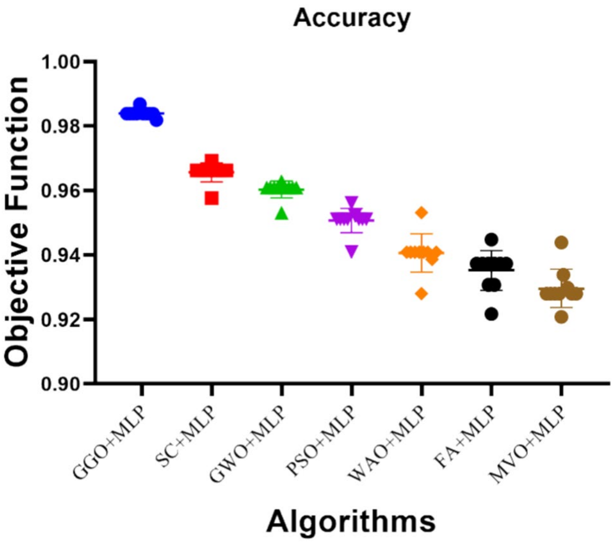


Fig. 7. Assessing the accuracy of the GGO + MLP approach and optimization algorithms using the MLP model, considering the objective function.

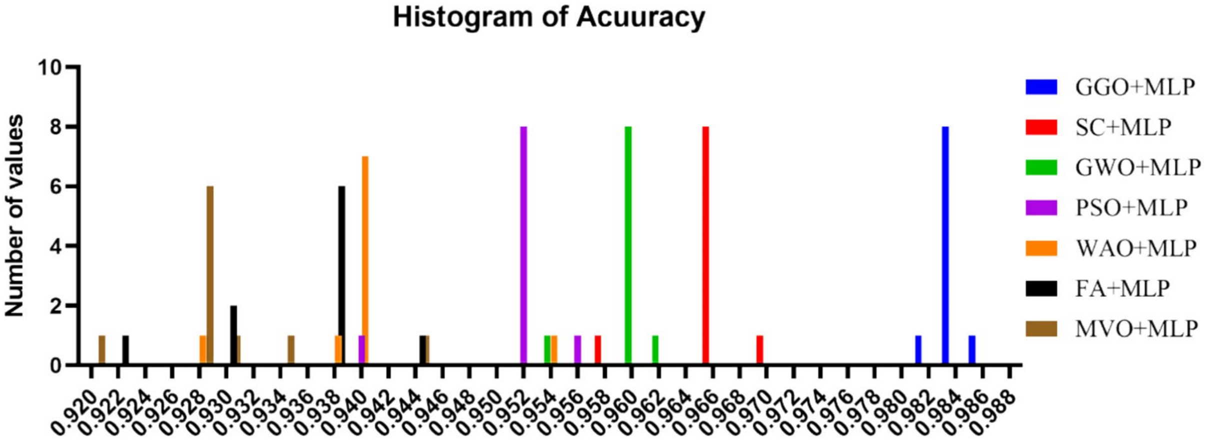


Fig. 8. Histograms of the accuracy results achieved by GGO + MLP approach as well as alternative combinations of optimization techniques with MLP models.

ANOVA table	SS	DF	MS	F (DFn, DFd)	P value
Treatment (between columns)	0.02191	6	0.003652	F (6, 63) = 184.1	P < 0.0001
Residual (within columns)	0.00125	63	1.98E−05		
Total	0.02316	69			

Table 8. The outcomes of the ANOVA of the proposed GGO algorithm with MLP model for lung cancer classification.

	GGO + MLP	SC + MLP	GWO + MLP	PSO + MLP	WOA + MLP	FOA + MLP	MVO + MLP
Theoretical median	0	0	0	0	0	0	0
Actual median	0.9838	0.9662	0.9609	0.9511	0.9409	0.9373	0.928
Number of values	10	10	10	10	10	10	10
Wilcoxon Signed Rank Test							
Sum of signed ranks (W)	55	55	55	55	55	55	55
Sum of positive ranks	55	55	55	55	55	55	55
Sum of negative ranks	0	0	0	0	0	0	0
P value (two tailed)	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Exact or estimate?	Exact	Exact	Exact	Exact	Exact	Exact	Exact
P value summary	**	**	**	**	**	**	**
Significant (alpha = 0.05)?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
How big is the discrepancy?							
Discrepancy	0.9838	0.9662	0.9609	0.9511	0.9409	0.9373	0.928

Table 9. The Wilcoxon signed-rank test findings of the proposed approach, (GGO + MLP), with various configurations of other optimization algorithms with the MLP model for lung cancer classification.

The efficacy of the proposed GGO + MLP approach for optimizing the objective function is confirmed by a comparative analysis with other optimization algorithms utilizing the MLP model. The accuracy and histogram of the results obtained using GGO + MLP are presented in the accuracy plots depicted in Figs. 7 and 8, correspondingly. The accuracy plot demonstrates that the proposed approach surpassed the other six optimization algorithms while using the MLP model. In addition, the accuracy histogram plot demonstrates the effectiveness of the proposed GGO + MLP approach in accurately classifying lung cancer cases in the input dataset.

The statistical differences between the proposed algorithm and other competing algorithms are evaluated using ANOVA and Wilcoxon's rank-sum tests. The ANOVA results for the proposed GGO + MLP approach are presented in Table 8. The Wilcoxon's rank-sum test, as presented in Table 9, is employed to evaluate the presence of a statistically significant difference in the outcomes produced by the algorithms. A p-value below 0.05 indicates statistically significant supremacy. The outcomes reveal that the GGO + MLP approach is superior and shows the statistical significance of the approach.

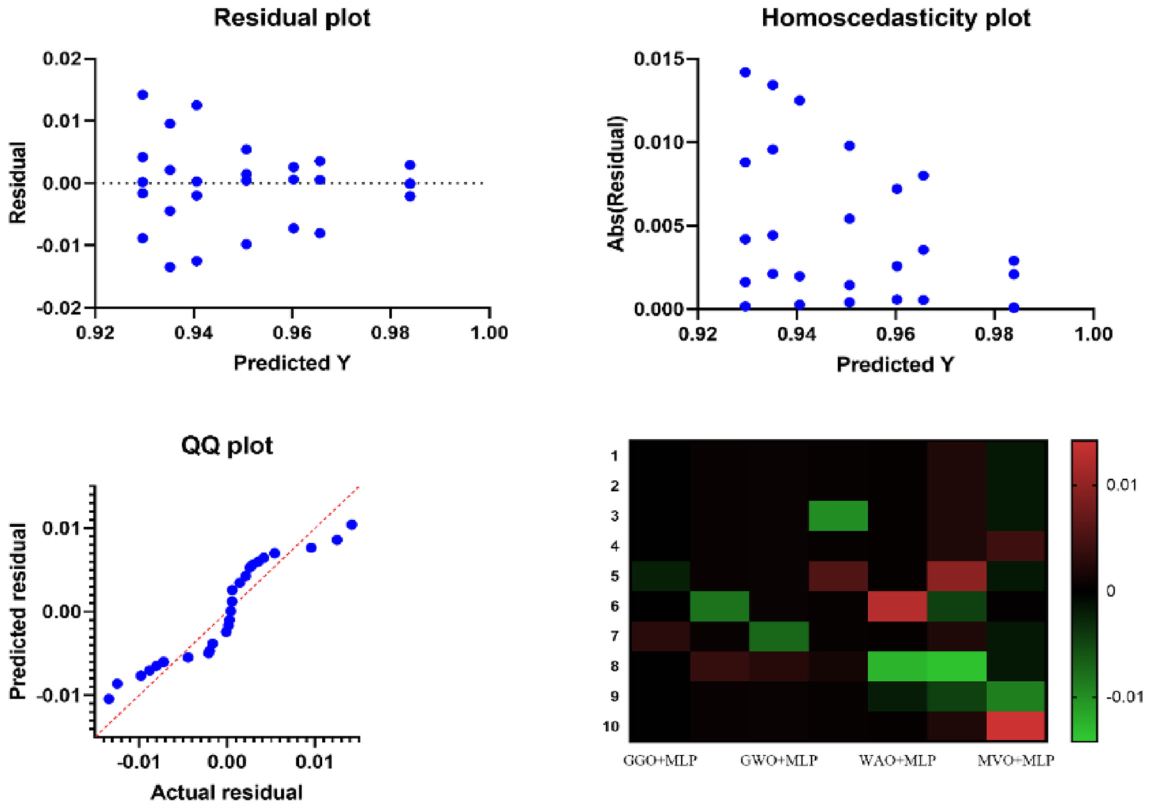


Fig. 9. Analysis plots of the obtained results using the proposed GGO + MLP approach.

The statistical findings in Table 8 demonstrate the comparative efficiency of the GGO + MLP approach compared to other five optimizers (SC, GWO, PSO, WOA, FOA, and MVO) with MLP model on benchmark functions. Table 9 illustrates that the GGO + MLP approach outperformed the other six optimizers using the MLP model due to its employment of two separate exploitation procedures in each cycle. The initial procedure entails progressing towards the most optimal solution identified at a specific point, whereas the second approach involves aggressively pursuing superior solutions in close vicinity. The GGO + MLP approach can get exceptional results by implementing these procedures to effectively use the search space. In order to achieve optimal use, it is imperative to maintain a harmonious equilibrium between exploration and exploitation within the search domain. Furthermore, it is crucial to initiate the exploitation process at an early stage in every iteration and gradually augment the total number of the participants in exploitation group. In general, the GGO algorithm demonstrated superior performance compared to other optimizers on most of the unimodal benchmark functions.

Figure 9 exhibits the residual plot, QQ plot, heteroscedasticity plot, and heat map for this scenario. The figure employs residual plots, quartile-quartile (QQ) plots, and homoscedasticity to showcase the effectiveness and robustness of the proposed GGO + MLP approach. The efficacy of the selected features in case categorization is demonstrated by the values depicted in the QQ plot, which approximately fit to a linear trend. The data shown in the residual and homoscedasticity plots provide more evidence to support these results. Moreover, the superiority of the proposed approach is further evidenced by the heatmap depicted in Fig. 5, which clearly illustrates that the proposed approach outperformed the other six binary feature selection techniques. The effectiveness of the GGO + MLP approach is confirmed by the heatmap, as it achieved the most optimal results compared to other feature selection strategies. The analysis plots shown in Fig. 8 demonstrate the effectiveness of the proposed

Metric	Rastrigin (bGGO)	Rastrigin (bSC)	Rastrigin (bMVO)	Rastrigin (bPSO)	Rastrigin (bWOA)	Rastrigin (bGWO)	Rastrigin (bFA)
Average error	0.473	0.485	0.496	0.516	0.545	0.553	0.571
Average Select size	0.454	0.472	0.481	0.497	0.522	0.536	0.554
Average Fitness	0.518	0.538	0.568	0.583	0.598	0.624	0.646
Best Fitness	0.485	0.513	0.546	0.563	0.581	0.616	0.636
Worst Fitness	0.532	0.562	0.574	0.592	0.615	0.628	0.643
Standard deviation Fitness	0.324	0.347	0.361	0.378	0.393	0.416	0.435

Table 10. Results of GGO algorithm compared to others optimization algorithms using Rastrigin function.

Metric	Ackley (bGGO)	Ackley (bSC)	Ackley (bMVO)	Ackley (bPSO)	Ackley (bWOA)	Ackley (bGWO)	Ackley (bFA)
Average error	0.327	0.363	0.371	0.378	0.415	0.443	0.451
Average Select size	0.314	0.345	0.353	0.358	0.394	0.416	0.424
Average Fitness	0.482	0.516	0.535	0.541	0.552	0.559	0.563
Best Fitness	0.471	0.502	0.515	0.519	0.526	0.534	0.542
Worst Fitness	0.517	0.553	0.573	0.578	0.587	0.594	0.605
Standard deviation Fitness	0.263	0.292	0.302	0.307	0.318	0.326	0.332

Table 11. Results of GGO algorithm compared to others optimization algorithms using Ackley function.

Refs	Methodology	Accuracy
Ref[4]	SVM with Feed-Forward Back Propagation Neural Network and GA optimization	Accuracy: 98.08%
Ref[9]	Hybrid CNN and Ebola Optimization Search Algorithm (EOSA)	Accuracy: 93.21%
Ref[11]	Random Forest classifier with K-means visualization	Accuracy: 92.14%
Ref[12]	Whale Optimization Algorithm with SVM	Accuracy: 95%
Proposed work	GGO + MLP	Accuracy: 98.4%

Table 12. Comparison of proposed work with a recent existing system.

GGO + MLP approach in addressing the optimization concerns for categorizing lung cancer patients discussed in this article.

To verify the performance of the proposed model, we used two benchmark functions, namely, Rastrigin function and Ackley function. Table 10 demonstrates the results of GGO algorithm compared to others optimization algorithms using Rastrigin function using the metrics, average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness. As demonstrated in Table 10, GGO algorithm gives the best results in terms of average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness.

Table 11 demonstrates the results of GGO algorithm compared to others optimization algorithms using Ackley function using the metrics, average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness. As demonstrated in Table 11, GGO algorithm gives the best results in terms of average error, average select size, average fitness, best fitness, worst fitness and standard deviation fitness.

Table 12 demonstrates a performance comparison of proposed work with a recent existing system.

Statistical analysis of reference models results

Among the visualization methods applied in statistics of the results of this prototypical approach, the description of the distribution of the results is performed. A KDE Plot is given in Fig. 10, which is comprehensive in analyzing models' accuracy. This graph output measures differences in both norms and fluctuations between various models

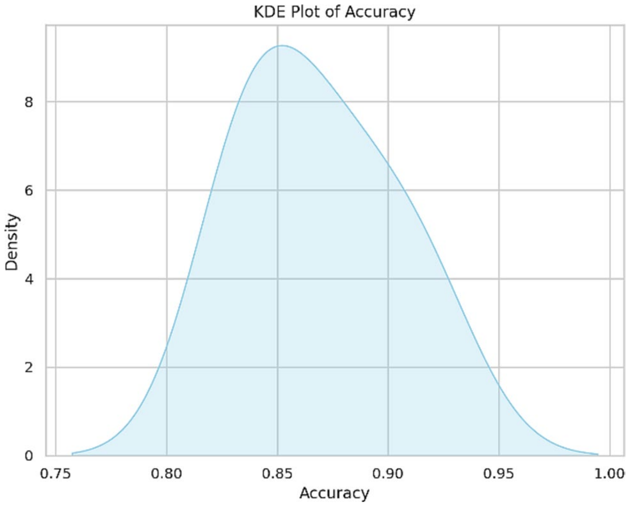


Fig. 10. KDE plot of accuracy of reference models results.

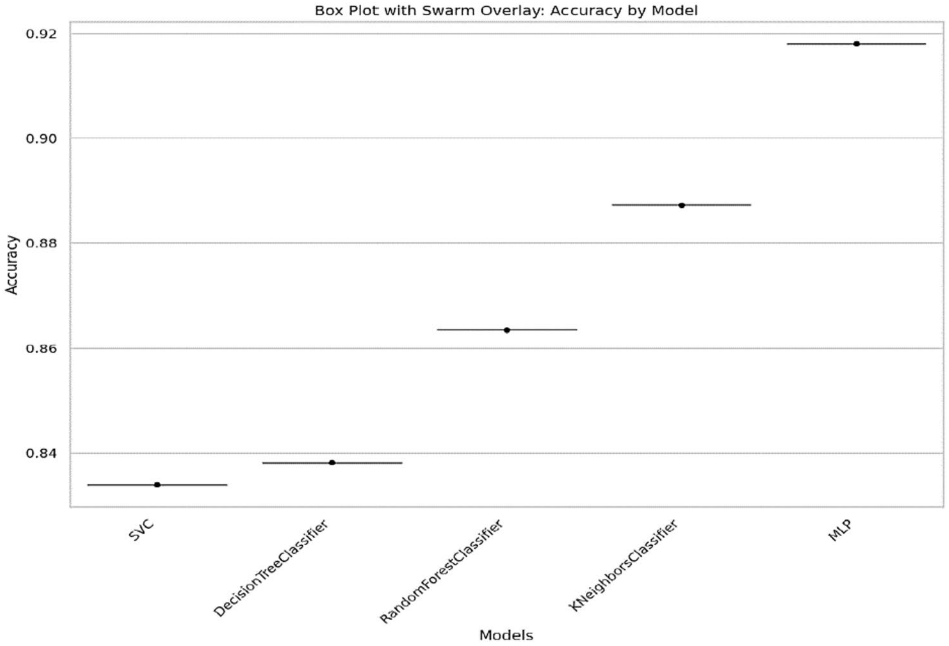


Fig. 11. Box plot with swarm overlay: accuracy by reference models.

of the algorithm in question. The role of observation is vital in enhancing the efficacy of such models since it helps to identify the particular models that are performing well.

The KDE (Kernel Density Estimation) plots help to estimate the probability density function of a continuous random variable. Here, the graph shows the error levels of the reference models in the x-axis and then the number of times an error level happens in the y-axis. The KDE plot assists researchers in understanding the distribution behavior of the accuracy by identifying frequent routines, examples of modes and often some clusters of high-performant models.

Therefore, a Box Plot with Swarm Overlay illustrates the distribution of precision in addition to the reference models in Fig. 11. This way, it is not only a short description of the accuracy values but also a comparison of the models, depicting the median, quartiles, and possibly the outliers. Annotations on individual parts of data clouds are added, including united data. This detail lifts the level of resolution and provides information about the accuracy and deviations from the most common form of an example.

Figure 12 shows a Pairplot with Regression Lines for the Reference Models Results. The network diagram technique thus helps examine the interrelation of all the variables considered to be of particular interest, such as accuracy, sensitivity, specificity, and so forth, for all the considered reference models. Each relation between the variables is sketched as a scatterplot, with the lines representing regression (showing the trend between these variables) as well.

Figure 13 is the representation of a Regression Plot; this illustration reveals the relationship between Accuracy and F-Score when using the Reference Models Results. These two observations graphically demonstrate the relationship between these two most important metrics, thus helping researchers quantify the matter of a tradeoff between accuracy and F-score. The regression line is fitted to the inter-connected data plots, which advertises the direction and the strength of Accuracy, and F-Score link relative to different reference models.

Statistical analysis of optimization results

The KDE chart in Fig. 14 highlights the accuracy distribution from the specimen of the optimization process. The visualization will show the levels of dispersion and measure a central tendency for the reference models, just like a KDE plot. This will help with understanding the variability and accuracy. The curve shape and density analyses provide a way to see the trends and patterns in the optimization outcomes so that different strategies may be compared.

Figure 15, which KDE Plot powers, is supplemented by another Box Plot demonstrating the distribution of error levels of optimization models. The visualization technique very well captures the composite values like the median, the percentiles and, could be, the outlier of the accuracy values in order to compare them to different optimization schemes.

Rounding off the metrics evaluation, Fig. 16 ranks optimization models using different metrics such as Accuracy, Sensitivity, Specificity, PPV, NPV, and F1 Score. The performance of all the optimization models is evaluated comprehensively using multiple performance metrics, which help in selecting the most appropriate model, tuning its parameters, and finally in process improvement.

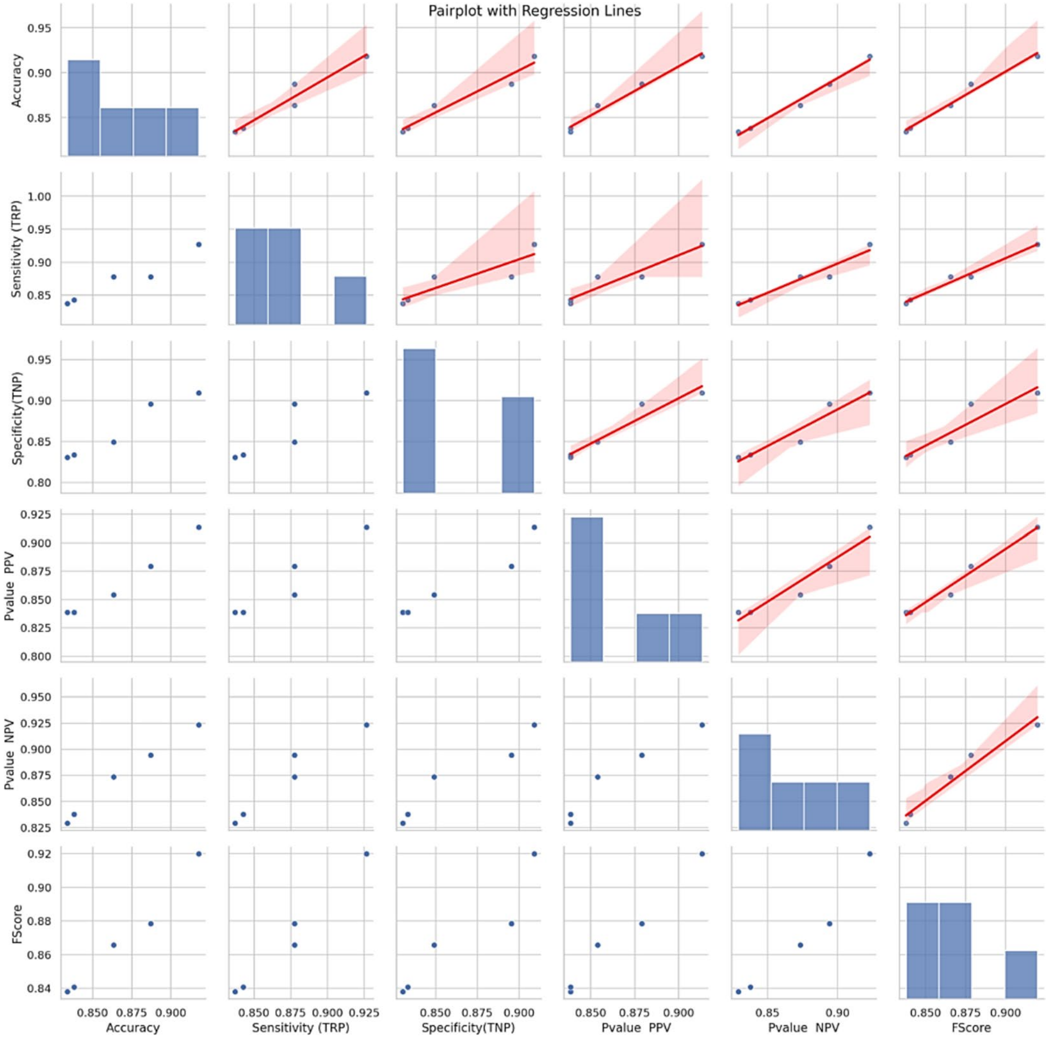


Fig. 12. Pairplot with Regression Lines for Reference Models Results.

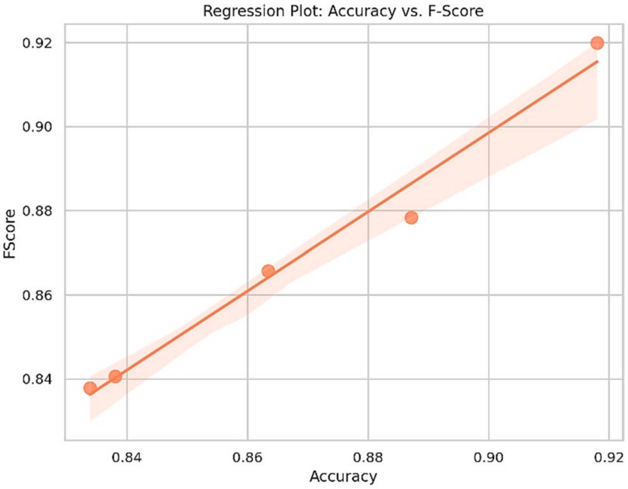


Fig. 13. Regression Plot: Accuracy vs. F-Score for Reference Models Results.

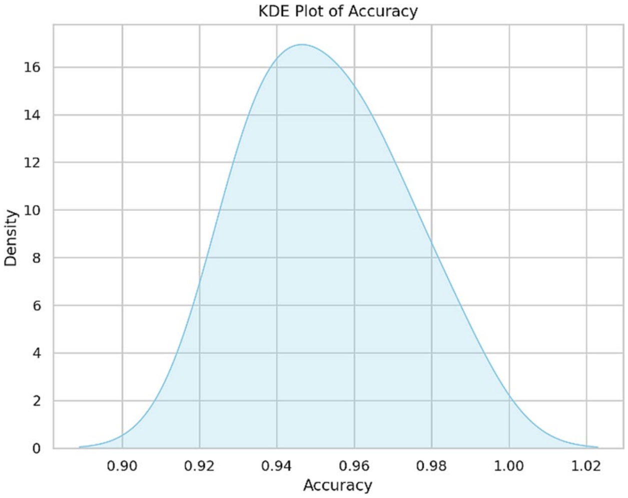


Fig. 14. KDE plot of accuracy of optimization results.

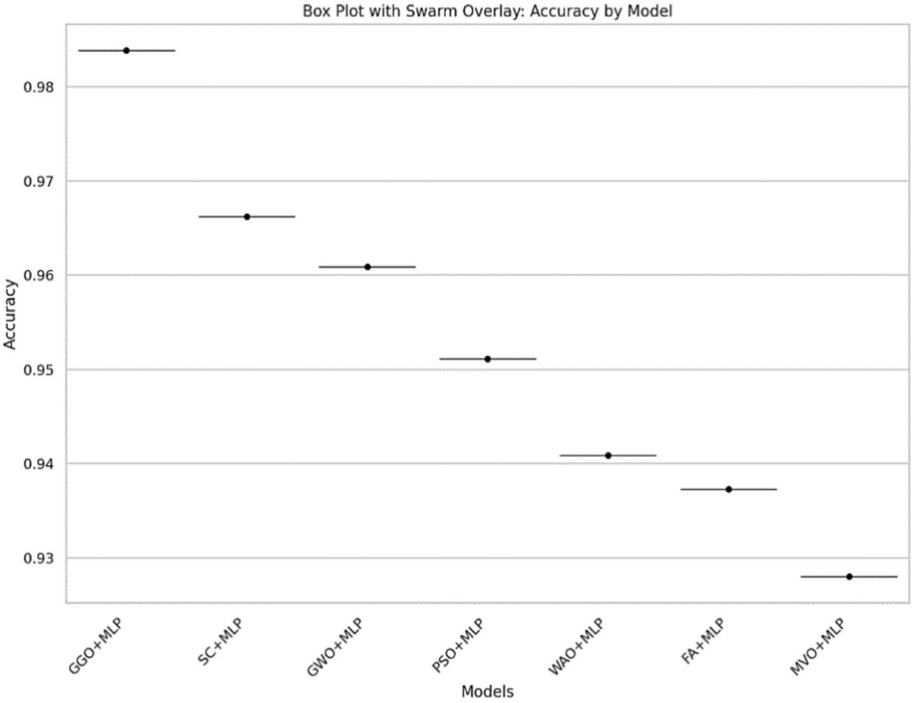


Fig. 15. Box plot with swarm overlay: accuracy by optimization models.

Limitations

GGO algorithm study in improving lung cancer classification shows good results, especially in improving feature selection and classification accuracy. However, there are several limitations and directions for future work that can be addressed to further improve the effectiveness of GGO in this field:

- The performance of the GGO algorithm was tested on one dataset only that may not represent the variety of real-world scenarios. Results may vary across different datasets, which may affect the generalizability of the results.
- Classification accuracy depends largely on the quality and completeness of the input data. Any inconsistencies or missing values in the data can affect the results.

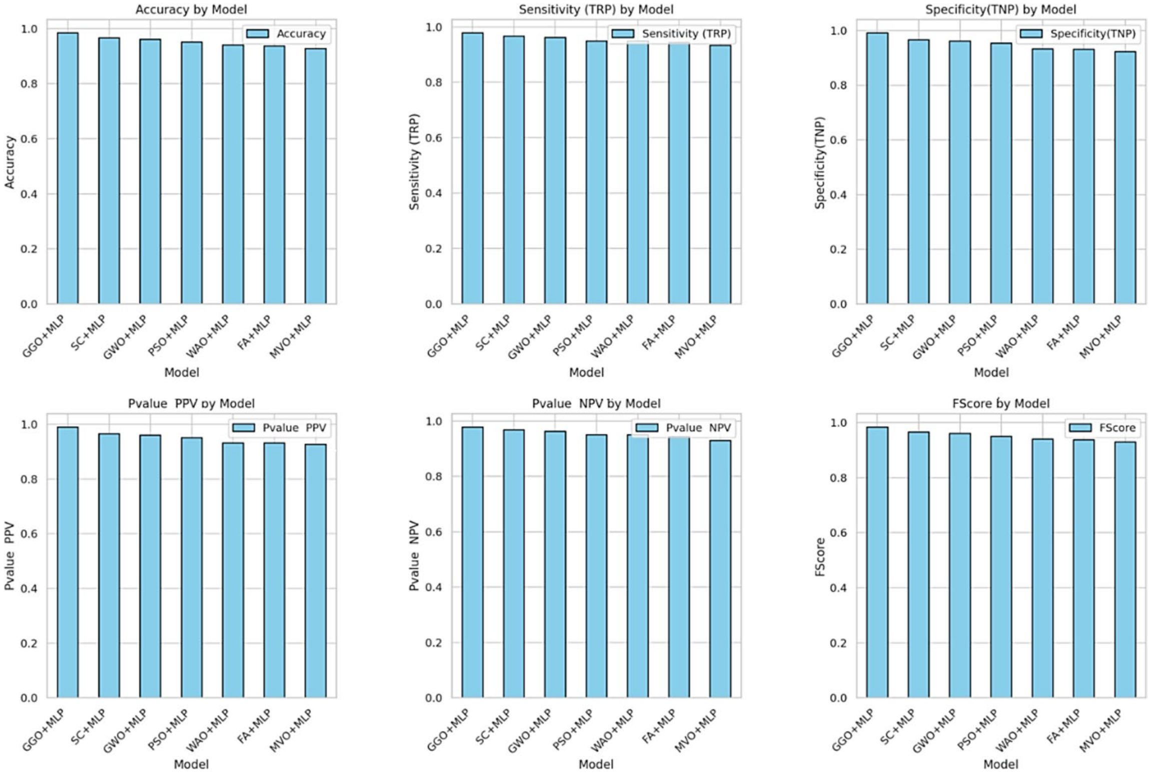


Fig. 16. Performance of the Reference models using Accuracy, Sensitivity (TRP), Specificity(TNP) by Model, Pvalue (PPV), Pvalue (NPV), FScore by Optimization Models.

Conclusion and future work

The GGO algorithm was introduced in this work with the objective of improving the accuracy of lung cancer case classification. In the beginning, data preparation techniques such as scaling, normalization, and removal of null values are carried out. In addition, the feature selection is performed utilizing the binary format of GGO (bGGO). The GGO algorithm's binary format is specifically intended to choose the most optimum combination of features that can improve classification accuracy when compared to six other binary optimization algorithms (SC, MVO, PSO, WOA, GWO, and FOA). The classification phase utilizes many classifiers, such as SVC, DTC, RFC, KNC, and MLP. The findings indicated that the Multilayer Perceptron (MLP) emerged as the most effective classifier, achieving an accuracy rate of 91.8%. The hyperparameter of the MLP model is tuned using GGO, and the outcome is compared to six alternative optimizers. The GGO with MLP model generated the highest result, with an accuracy of 98.4%. The statistical analysis employed the Wilcoxon signed-rank test and ANOVA for the feature selection and classification outcomes. Furthermore, a set of visual representations of the results was created to confirm the robustness and effectiveness of the proposed approach. Overall, the experimental and statistical results clearly indicate that the proposed approach outperforms other competing approaches for classifying lung cancer. The effective selection of cancer features, and the reduction of feature dimensionality enhanced the overall prediction accuracy and successfully mitigated the problem of overfitting in cancer features analysis. Enhancing the early prediction rates of lung cancer may be achieved by acquiring and analyzing sensor data, then employing an optimum approach in the future. The performance of the GGO algorithm was tested on specific dataset for lung cancer classification, so in the future, several large datasets can be used for generalizing the results. Also, in the future several optimization models and deep learning models[32–34] can be used for lung cancer classification.

Received: 14 May 2024; Accepted: 2 September 2024
Published online: 10 October 2024

References

1. Asuntha, A. & Srinivasan, A. Deep learning for lung Cancer detection and classification. *Multimed. Tools Appl.* **79**, 7731–7762 (2020).
2. Chaturvedi, P., Jhamb, A., Vanani, M. & Nemade, V. Prediction and classification of lung cancer using machine learning techniques. In *IOP Conference Series: Materials Science and Engineering* 12059 (IOP Publishing, 2021).
3. Zhu, Y. *et al.* Feature selection and performance evaluation of support vector machine (SVM)-based classifier for differentiating benign and malignant pulmonary nodules by computed tomography. *J. Digit. Imaging* **23**, 51–65 (2010).
4. Nanglia, P., Kumar, S., Mahajan, A. N., Singh, P. & Rathee, D. A hybrid algorithm for lung cancer classification using SVM and Neural Networks. *ICT Express* **7**(3), 335–341 (2021).
5. Elshewey, A. M. *et al.* A Novel WD-SARIMAX model for temperature forecasting using daily Delhi climate dataset. *Sustainability* **15**(1), 757 (2022).

6. Elhoseny, M., Tarek, Z. & El-Hasnony, I. M. Advanced cognitive algorithm for biomedical data processing: COVID-19 pattern recognition as a case study. *J. Healthc. Eng.* **2022**, 1–11 (2022).
7. Deserno, T. M., Antani, S. & Long, R. Ontology of gaps in content-based image retrieval. *J. Digit. Imaging* **22**, 202–215 (2009).
8. Ezugwu, A. E. *et al.* Metaheuristics: A comprehensive overview and classification along with bibliometric analysis. *Artif. Intell. Rev.* **54**, 4237–4316 (2021).
9. Mohamed, T. I. A., Oyelade, O. N. & Ezugwu, A. E. Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm. *PLoS ONE* **18**(8), e0285796 (2023).
10. Ren, Z., Zhang, Y. & Wang, S. A hybrid framework for lung cancer classification. *Electronics* **11**(10), 1614 (2022).
11. Bhattacharjee, A., Murugan, R. & Goel, T. A hybrid approach for lung cancer diagnosis using optimized random forest classification and K-means visualization algorithm. *Health Technol. (Berl).* **12**(4), 787–800 (2022).
12. Vijh, S., Gaur, D. & Kumar, S. An intelligent lung tumor diagnosis system using whale optimization algorithm and support vector machine. *Int. J. Syst. Assur. Eng. Manag.* **11**, 374–384 (2020).
13. 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).
14. Nancy, D. P. *et al.* Optimized feature selection and image processing based machine learning technique for lung cancer detection. *IJEER* **10**(4), 888–894 (2022).
15. Joshi, A. A. & Aziz, R. M. A two-phase cuckoo search based approach for gene selection and deep learning classification of cancer disease using gene expression data with a novel fitness function. *Multimed. Tools Appl.* **83**, 71721–71752 (2024).
16. Saxena, A., Chouhan, S. S., Aziz, R. M. & Agarwal, V. A comprehensive evaluation of Marine predator chaotic algorithm for feature selection of COVID-19. *Evol. Syst.* **15**, 1235–1248 (2024).
17. Yaqoob, A., Verma, N. K., Aziz, R. M. & Saxena, A. Enhancing feature selection through metaheuristic hybrid cuckoo search and Harris Hawks optimization for cancer classification. In *Metaheuristics for Machine Learning: Algorithms and Applications* 95–134 (Springer, 2024).
18. El-kenawy, E. S. M. *et al.* Greylag Goose Optimization: Nature-inspired optimization algorithm. *Expert Syst. Appl.* **238**, 122147 (2024).
19. Samadianfard, S. *et al.* Wind speed prediction using a hybrid model of the multi-layer perceptron and whale optimization algorithm. *Energy Rep.* **6**, 1147–1159 (2020).
20. Mirjalili, S. SCA: A sine cosine algorithm for solving optimization problems. *Knowl.-Based Syst.* **96**, 120–133 (2016).
21. Rigamonti, A. Mean-variance optimization is a good choice, but for other reasons than you might think. *Risks* **8**(1), 29 (2020).
22. Piotrowski, A. P., Napiorkowski, J. J. & Piotrowska, A. E. Particle swarm optimization or differential evolution: A comparison. *Eng. Appl. Artif. Intell.* **121**, 106008 (2023).
23. Mirjalili, S. & Lewis, A. The whale optimization algorithm. *Adv. Eng. Softw.* **95**, 51–67 (2016).
24. Al-Tashi, Q., Md Rais, H., Abdulkadir, S. J., Mirjalili, S. & Alhussian, H. A review of grey wolf optimizer-based feature selection methods for classification. In *Evolutionary Machine Learning Techniques: Algorithms and Applications* 273–86 (Springer, 2020).
25. de Vasconcelos Segundo, E. H., Mariani, V. C. & dos Santos, C. L. Design of heat exchangers using falcon optimization algorithm. *Appl. Therm. Eng.* **156**, 119–144 (2019).
26. Saigal, P. & Khanna, V. Multi-category news classification using support vector machine based classifiers. *SN Appl. Sci.* **2**(3), 458 (2020).
27. Shams, M. Y. *et al.* A machine learning-based model for predicting temperature under the effects of climate change. In *The Power of Data: Driving Climate Change with Data Science and Artificial Intelligence Innovations* 61–81 (Springer, 2023).
28. Tarek, Z. *et al.* Soil erosion status prediction using a novel random forest model optimized by random search method. *Sustainability* **15**(9), 7114 (2023).
29. Elshewey, A. *et al.* Weight prediction using the hybrid stacked-LSTM food selection model. *Comput. Syst. Sci. Eng.* **46**(1), 765–81 (2023).
30. Al Bataineh, A., Kaur, D. & Jalali, S. M. J. Multi-layer perceptron training optimization using nature inspired computing. *IEEE Access* **10**, 36963–36977 (2022).
31. Saeed, M. *et al.* Electrical power output prediction of combined cycle power plants using a recurrent neural network optimized by waterwheel plant algorithm. *Front. Energy Res.* **11**, 1234624 (2023).
32. Shams, M. Y., El-kenawy, E. S., Ibrahim, A. & Elshewey, A. M. A hybrid dipper throated optimization algorithm and particle swarm optimization (DTPSO) model for hepatocellular carcinoma (HCC) prediction. *Biomed. Signal Process. Control* **85**, 104908 (2023).
33. Elshewey, A. M., Tawfeek, S. M., Alhussan, A. A., Radwan, M. & Abed, A. H. Optimized deep learning for potato blight detection using the waterwheel plant algorithm and sine cosine algorithm. *Potato Res.* https://doi.org/10.1007/s11540-024-09735-y (2024).
34. Shams, M. Y., Tarek, Z., El-kenawy, E. S., Eid, M. M. & Elshewey, A. M. Predicting Gross Domestic Product (GDP) using a PC-LSTM-RNN model in urban profiling areas. *Comput. Urban Sci.* **4**(1), 3 (2024).

Acknowledgements

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R 308), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

All authors have contributed equally.

Funding

Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.M.E.

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.