Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-024-01770-9

ORIGINAL RESEARCH

Enhancing lung cancer prediction through crow search, artificial bee colony algorithms, and support vector machine

Samira Tared[1] · Latifa Khaouane[1] · Salah Hanini[1] · Affaf Khaouane[1] · Mohamed Roubehie Fissa[1]

Received: 28 October 2023 / Accepted: 30 January 2024
© The Author(s), under exclusive licence to Bharati Vidyapeeth's Institute of Computer Applications and Management 2024

Abstract Lung cancer is one of the leading causes of death worldwide. According to statistics from the World Health Organization (WHO) for the year 2020, lung cancer has caused the death of about 1.80 million people. It is the most common form of cancer, and it is often diagnosed in later stages when metastatic spread to other areas of the body has occurred. Early detection and treatment can help reduce the high death rate. In this research, we present a novel approach for predicting lung cancer using artificial intelligence (AI) techniques. We evaluated and compared several machine learning (ML) models, including artificial neural networks (ANN), K-nearest neighbors (KNN), and support vector machine (SVM), to identify the best-performing model. However, to enhance accuracy and efficiency, we introduced a hybridization optimization approach known as Crow Search Algorithm -Artificial Bee Colony-Support Vector Machine (CSABC-SVM). The CSABC-SVM algorithm combines the strengths of the CSABC Optimizer and SVM to create a powerful predictive model. Our study yielded a remarkable accuracy of 98.36%, demonstrating the effectiveness of the CSABC-SVM hybrid model in lung cancer prediction. These results underscore the potential of AI in transforming medical diagnostics, helping to improve healthcare outcomes.

Keywords Lung cancer · Prediction · Artificial intelligence · Optimization · CSABC-SVM

✉ Samira Tared
tared.samira@gmail.com; samirstard@gmail.com

Latifa Khaouane
latifa_khaouane@yahoo.fr

Salah Hanini
s_hanini2002@yahoo.fr

Affaf Khaouane
affoufa80@gmail.com

Mohamed Roubehie Fissa
roubehiefissa.mohamed2021@gmail.com

[1] Biomaterials and Transport Phenomena Laboratory (LBMPT), Department of Process and Environmental Engineering, Faculty of Technology, University of Yahia Fares, 26000 Medea, Algeria

1 Introduction

Lung cancer is one of the leading causes of death worldwide [1–3]. According to the WHO study, lung cancer was the deadliest cancer, with about 1.80 million people dying from it in 2020 [4]. It is the second most prevalent kind of cancer and is often diagnosed at later stages when metastatic dissemination to other regions of the body has occurred [2]. As it originates in the lining of the bronchi and bronchioles, lung cancer obstructs gas exchange in the lungs and causes breathing difficulties [4]. It begins as an abnormal growth of cells in the lungs and can spread to other parts of the body [4]. Difficult to evaluate and discover in the early stages since it does not create pain or symptoms in certain situations [5]. A patient diagnosed with Lung cancer may experience symptoms such as chest pain, cough, shortness of breath, wheezing, hemoptysis i.e. coughing up blood, Pancoast syndrome (shoulder pain), hoarseness (paralysis of vocal cords), weakness, weight loss, fatigue [5]. Symptoms are classified according to the location and size of the tumor [5]. Early detection and treatment can help to lower such a high death rate [6].

In recent years, there has been continuous development in cancer research [7]. Multitudes of procedures are used to diagnose cancer and predict the outcomes of cancer treatment [4]. Researchers used a variety of methods,

Published online: 04 March 2024

including early-stage screening, to identify cancer types before they caused specified degrees of symptoms [7].

AI is the ability of robots to mimic human behavior and is particularly effective at managing large volumes of data [8]. AI also possesses the ability to analyze complex medical data [9]. ML is a branch of AI [10, 11]. That combines several statistics and finds difficult-to-detect patterns in big or complicated data sets [11]. Due to the availability of clinical data, ML has achieved an important role in medical decision-making [12]. Creating an ML model could greatly assist in identifying illnesses and making appropriate healthcare choices in real time [12]. Additionally, it would enable better utilization of hospital resources by identifying individuals with major risk factors at an early stage [12].

Singh and Gupta [13] presented a methodology for detecting and classifying Computed Tomography (CT) images associated with lung cancer into two categories: benign and malignant. They first applied image processing techniques to pre-process the images. Then, they used supervised learning algorithms to classify them. They extracted texture features along with statistical features and provided many of the extracted features to the classifiers. They used a dataset of 15,750 clinical images consisting of 6910 benign images and 8840 malignant lung cancer images, and seven different classifiers known as the KNN, SVM, Decision Tree (DT), Multinomial Naive Bayes, Stochastic Gradient Descent (SGD), Random Forest (RF), and Multi-Layer Perceptual (MLP). They obtained simulation results using three parameters, namely accuracy, F1 score, precision, and recall. The outcomes revealed that the MLP exhibited the highest accuracy, achieving 88.55%, surpassing other classifiers. The MLP model comprised a total of 3 hidden layers, each with 100 hidden units. The Rectified Linear Unit (ReLU) served as the activation function for this model. The aim of the research conducted by Faisal et al. [14] is to evaluate the discriminatory power of several predictors including an SVM, C4.5 decision tree, MLP, neural network, and Naive Bayes (NB), and Gradient-boosted Tree to increase the efficiency of lung cancer detection by its symptoms. The researchers extracted datasets from the University of California at Irvine (UCI) repository and performed pre-processing to clean the data. Subsequently, they trained and tested the selected classifiers using standard tenfold cross-validation. Results for each classifier were calculated, and the top three classifiers, based on performance metrics such as accuracy, were chosen to form an ensemble using Majority Voting. This ensemble, combined with RF, was utilized to compute the results, enabling a comprehensive performance comparison of both individual and ensemble classifiers. Notably, the Gradient-boosted Tree achieved an accuracy of 90%, surpassing all other individual and ensemble classifiers in terms of performance. Wu et al. [15] introduced an RF model designed for the identification of lung cancer using routine cancer indices. The validation of the model was conducted through a tenfold cross-validation method. The dataset utilized in this study was gathered from the Second Hospital of Lanzhou University, consisting of 277 patients. The results demonstrated an accuracy of 95.7%, sensitivity of 96.3, and specificity of 94.97. In 2022, Mamun et al. [4] conducted a comprehensive review of previous studies related to lung cancer prediction models. They developed and introduced ensemble learning techniques, including XGBoost, LightGBM, Bagging, and AdaBoost for the prediction of lung cancer based on a survey dataset of 309 individuals. The attributes used in lung cancer prediction models included age, smoking, yellow fingers, anxiety, peer pressure, chronic diseases, fatigue, allergies, wheezing, alcohol, cough, shortness of breath, and swallowing. The dataset was processed using the Synthetic Minority Over-Sampling Technique (SMOTE) oversampling method to balance it and make it suitable for analysis. The proposed models were trained and tested on the dataset using a standard tenfold cross-validation approach. XGBoost outperformed the other techniques, achieving an accuracy of 94.42%, precision of 95.66%, recall of 94.46%, and an Area Under the Curve (AUC) of 98.14%. Lakshmanaprabu SK et al. [16] proposed an innovative method for automated lung cancer diagnosis using CT images. The approach involves employing an Optimal Deep Neural Network (ODNN) for deep feature extraction from CT lung images. Linear Discriminate Analysis (LDA) is then applied to reduce the dimensionality of these features, followed by the optimization of the ODNN using the Modified Gravitational Search Algorithm (MGSA). The classifier achieved promising results, with a sensitivity of 96.2%, a specificity of 94.2%, and an overall accuracy of 94.56%. This method aims to enhance early detection of lung cancer, crucial for improving patient survival rates. Dritsas and Trigka [17] utilized supervised learning to create models for detecting lung cancer in individuals based on various symptoms. Several models, including NB, Bayesian Network (BayesNet), SGD, SVM, Logistic Regression (LR), ANN, KNN, Logistic Model Tree (LMT), Random Tree (RT), Rotation Forest (RotF), Reduced Error Pruning Tree (RepTree), RF, and AdaBoostM1, were evaluated for accuracy, precision, recall, F-Measure, and AUC. After applying SMOTE with tenfold cross-validation, RotF emerged as the top-performing model, achieving an accuracy, precision, recall, and F-Measure of 97.1% and an AUC of 99.3% in identifying lung cancer. Maulidina F et al. [18] proposed an ML approach for lung cancer diagnosis using the Particle Swarm Optimization-Genetic Algorithm-Support Vector Machine (PSO-GA-SVM) method. The research utilized a CT Scan dataset from Cipto Mangunkusumo Hospital, Jakarta, Indonesia. Swarm Optimization-Genetic Algorithm (PSO-GA) was employed to optimize SVM parameters, and the method's performance was assessed based on accuracy, precision, recall, and f1-score. In comparison to the SVM

without parameter optimization, PSO-GA-SVM achieves an accuracy value of 97.69%. K. Vijila Rani et al. [19] presented an intellectual classification system to differentiate between normal and abnormal lung CT images to address the task of predicting cancer cells in lung images. They used Gray Wolf Optimization and Whale Optimization-Support Vector Machine (GWO & WOA-SVM) algorithms for lung tumor detection. The investigations focused on the segmentation and classification part to find the lung lesion area. The process involved four stages: image acquisition, preprocessing, and optimization using advanced clustering, segmentation using the Advanced Surface Normalized Overlapping (ASNO) lung segmentation algorithm, and classification using a hybrid classifier with optimization techniques different. The hybrid classifier (KNN-SVM) achieved a remarkable classification accuracy rate of 97.6% when applied to 700 images, outperforming other methods. Nanglia P et al. [20] explored the application of a combinational Feed-Forward Back Propagation Neural Network (FFBPNN) for making judgments in lung cancer diagnosis. To improve classification efficiency, the study integrates SVM with the FFBPNN, creating a hybrid algorithm. The research employs a dataset of 500 images, with 75% used for training and the remaining 25% for classification. The proposed three-block mechanism involves dataset preprocessing, feature extraction using the Speeded Up Robust Features (SURF) technique optimized by Genetic Algorithm (GA), and classification via FFBPNN. The hybrid classification algorithm, named Kernel Attribute Selected Classifier (KASC), achieves an impressive overall classification accuracy of 98.08%. Vijh S et al. [21] presented a comprehensive approach for the automated detection of lung cancer tumors in CT images. The proposed intelligent system involves various image processing steps, including enhancement, segmentation, post-processing, feature extraction, feature selection, and classification using the SVM kernel. The Gray Level Co-occurrence Matrix method is utilized to extract 19 texture and statistical features from lung CT images. The Whale Optimization Algorithm (WOA) is employed for selecting the best prominent feature subset. WOA_SVM was developed to automate an ancillary diagnostic system, determining whether a CT lung image is normal or abnormal. The proposed methodology achieved a high accuracy of 95% using a Radial Bias Function (RBF) support vector kernel.

After a comprehensive examination of the existing literature, we now provide a summarized overview of key findings and characteristics of the reviewed studies, the best performing models of the studies containing multiple models are in bold in Table 1

In this study, we conducted a comprehensive evaluation and comparison of various machine learning models, including ANN, KNN, and SVM for prediction of lung cancer prediction. To enhance accuracy and efficiency in the prediction of lung cancer, and to make the model clinically valuable for healthcare decision-making, we adopted an optimization approach by hybridizing the best-performing model with two nature-inspired optimization algorithms: the Crow Search Algorithm (CSA) and the Artificial Bee Colony (ABC)**.** CSA and ABC excel in efficiently exploring complex search spaces, each employing different techniques for exploration and exploitation. This hybridization approach aims to leverage the strengths of these algorithms and overcome their limitations, making it particularly advantageous for challenging problems like prediction of lung cancer prediction.

ABC is distinguished by its ability to search effectively but lacks exploration capability, while the Crow Search Algorithm utilizes the position and memory of each crow for objective function evaluation. This combination of strengths motivated us to propose an alternative method of prediction based on hybridization, avoiding the limitations of traditional technologies. The hybridization technique aims to provide a robust and clinically relevant predictive model, addressing the shortcomings of individual models. Through this research effort, we aim to advance the field of lung cancer prediction, enabling early diagnosis and proactive patient care, ultimately benefiting healthcare professionals and patients alike.

Table 1 Summary of literature review

Authors	Year	Models	Accuracy (%)
[4]	2022	**XGBoost,** LightGBM, Bagging, AdaBoost	94.42
[13]	2019	KNN, SVM, DT, Multinomial Naive Bayes, SGD, RF, **MLP**	88.55
[14]	2019	SVM, C4.5 DT, MLP, neural network, NB, **Gradient Boosted Tree**	90
[15]	2019	RF	95.7
[17]	2022	NB, BayesNet, SGD, SVM, LR, ANN, KNN, LMT, RT, **RotF,** RepTree, RF AdaBoostM1	97.1
[18]	2021	PSO-GA-SVM	97.69
[19]	2022	GWO & WOA-SVM **KNN-SVM**	97.6
[21]	2020	WOA_SVM	95

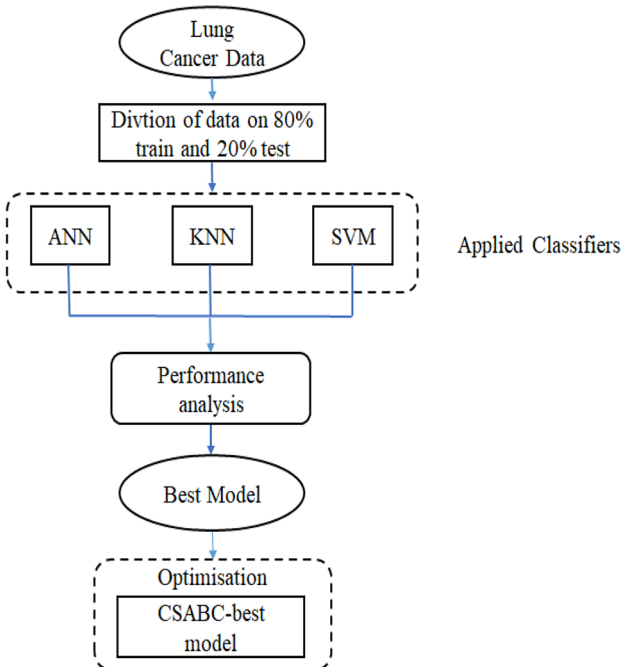


Fig. 1 The organizational structure of the applied procedure

Table 2 Description of database attributes

Attribute	Description
Gender	Female = 0, Male = 1
Age	Age In Years (21–87)
Smoking	Yes = 2, No = 1
Yellow fingers	Yes = 2, No = 1
Anxiety	Yes = 2, No = 1
Peer pressure	Yes = 2, No = 1
Chronic disease	Yes = 2, No = 1
Fatigue	Yes = 2, No = 1
Allergy	Yes = 2, No = 1
Wheezing	Yes = 2, No = 1
Alcohol consuming	Yes = 2, No = 1
Coughing	Yes = 2, No = 1
Shortness of breath	Yes = 2, No = 1
Swallowing difficulty	Yes = 2, No = 1
Chest pain	Yes = 2, No = 1
Lung cancer	Yes = 1, No = 0

2 Materials and method

In this study, the methodology was structured around a series of steps outlined in the flow diagram depicted in Fig. 1.

2.1 Data collection

In this study, we processed a lung cancer dataset to develop our proposed model. The dataset was sourced from Kaggle [22], encompassing a total of 16 attributes. Within this dataset, we found records for 309 patients, comprising 162 males and 147 females. Among these patients, 39 were categorized as normal, while 270 were diagnosed with lung cancer (see Table 2).

2.2 Data preprocessing

In this study, Python software was employed for our data analysis. The dataset we utilized was characterized by the absence of any missing values or outliers, rendering further data preprocessing unnecessary. To discern correlations among the features, we employed a heatmap visualization technique. Moreover, we employed Kernel Density Estimation (KDE) plots to compare the distribution of patients by age, distinguishing those with lung cancer from those without.

2.3 Model building

In this study, we performed an exhaustive analysis of a lung cancer dataset, employing multiple classification algorithms, including ANN, KNN, and SVM. To evaluate their efficacy, we partitioned the dataset into a training set (comprising 80% of the data) and a testing set (20%), subsequently computing performance metrics for each model. We conducted a comprehensive comparison of the performance of the three models to determine the most promising model.

2.3.1 Artificial neural network

ANN is an ML technique that imitates the function and workings of the human brain [23]. A neural network's artificial neuron utilizes specific input information to apply appropriate mathematical weights, which can ultimately predict an output object [8, 24]. In various healthcare domains, ANN approaches are valuable, such as in diagnosis and indicating breast cancer, lung cancer, and other ontology prediction, diagnostic systems, and drug analysis [5].

2.3.2 K-Nearest Neighbors

The KNN algorithm used for classification and regression is a non-parametric method [25]. It is a distance-based method (either Euclidean or Manhattan) to calculate the similarity or dissimilarity between two instances in a data set [26]. KNN is often used when there is little prior knowledge of the data [8]. It is sensitive to outliers in the data set [24]. In the classification, the number of k neighbors, where k is equal to

the square root of the instances' number and distance [24]. It is possible to determine the perfect collection of k-values for best predictions [24, 27].

2.3.3 Support vector machine

SVM is a supervised machine learning algorithm utilized to classify text into different categories [28]. It is used for both classification and regression tasks [5, 29]. It uses boundaries known as hyperplanes to divide the incoming data set in half and classify it [5]. This differentiation is magnified by increasing the margin on either side of the hyperplane [24]. In the end, the hyperplane-bounded region with the largest possible margin is used for analysis [24]. Finding nonlinear correlations by using a kernel function is one of the SVM method's features [24, 30].

2.4 Optimization

Following the identification of the most promising model, our focus shifted to optimization to enhance accuracy and other performance metrics. This optimization strategy was distinctive, as we leveraged a novel approach that combined the strengths of the CSA and ABC algorithm in conjunction with the best-performing model. This hybridization was designed to further enhance the model's capabilities and deliver superior results. Our unwavering dedication to achieving heightened accuracy underscores our commitment to providing reliable predictions with precision in the domain of lung cancer prediction.

2.4.1 Artificial bee colony

ABC algorithm is a stochastic optimization algorithm that inspiration by honeybee behavior, used in many optimization problems [31]. Karaboga proposed it in 2005 [32]. Individual bees are categorized into three sorts in the algorithm (employed, observer, and scout), with each agent having a particular function to perform in the process [31, 33].

2.4.2 Crow search algorithm

The CSA an algorithm proposed by Askarzadeh in 2016 is a meta-heuristic algorithm [34]. That mimics the strategies of crows for hiding and stealing food [35]. It is said that the most intelligent bird is the crow [34]. Crows can recognize faces and can communicate with one another to warn when an unfamiliar one is around [34]. In addition, they can communicate in sophisticated ways and remember the hiding place of their food several months later [34, 36]. Crows watch where other birds hide their food and steal it [36]. They use their own experience with theft to predict the behavior of thieves, to avoid becoming a victim in the future, and they can determine the safe path to protect their caches [36]. The primary source for this algorithm was the way crows hide their food [37].

The following definitions outline the four major CSA tenets [36]:

- Crows exist in the shape of a flock.
- Crows remember where to hide their food.
- When the robbery is committed, each of the members tracks down the others.
- Crows guard their caches by being extremely vigilant against theft.

2.5 Performance evaluation metrics

Once our models were trained, we proceeded to assess their performance through the application of relevant metrics. We gauged the performance of each algorithm by calculating performance measures utilizing a confusion matrix, a valuable tool for delineating the actual and expected outcomes of a classification system [37]. This matrix includes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) elements. Various statistical metrics, such as accuracy, sensitivity, specificity, precision, recall, F-measure, Matthew's correlation coefficient (MCC), and Receiver Operating Characteristic (ROC), were utilized to evaluate and compare the effectiveness and efficiency of the diverse algorithms. Our study encompassed a thorough comparative analysis of these algorithms based on the performance evaluation metrics employed.

The accuracy (Ac) describes the classification task's performance and counts the number of accurately predicted examples from all data instances [17]. It is calculated using Eq. 1.

$$Ac = TP + TN \,/\, (TP + TN + FP + FN) \tag{1}$$

The sensitivity (Se) is the ability of the model to correctly identify positive instances [38, 39]. It is calculated using Eq. 2.

$$Se = TP / (TP + FN) \tag{2}$$

The specificity (Sp) is the classifier's ability to identify negative outcomes [38, 39]. It is calculated using Eq. 3.

$$\mathrm{Sp} = \mathrm{TN} / (\mathrm{TN} + \mathrm{FP}) \tag{3}$$

The precision (Pr) is an evaluation measure that calculates the percentage of expected positives that are actual positives [38, 39]. It is calculated using Eq. 4.

$$\mathrm{Pr} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{4}$$

The recall is the percentage of successfully classified positives [38]. It is calculated using Eq. 5.

$$\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{5}$$

F-Measure is a statistically significant measure of accuracy and recall that maintains a balance between the accuracy and recall of a classifier [38, 39]. F-Measure is calculated using Eq. 6.

$$\mathrm{F-Measure} = 2 * \mathrm{Precision} * \mathrm{Recall}/(\mathrm{Precision} + \mathrm{Recall}) \tag{6}$$

MCC is a correlation coefficient value between -1 and + 1 [38]. MCC is calculated using Eq. 7.

$$\mathrm{MCC} = (\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN})/\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})} \tag{7}$$

ROC is used to gauge how well a model can discriminate across classes [38]. The area under the ROC curve is a combined space of the true positive rate and the false positive rate [38].

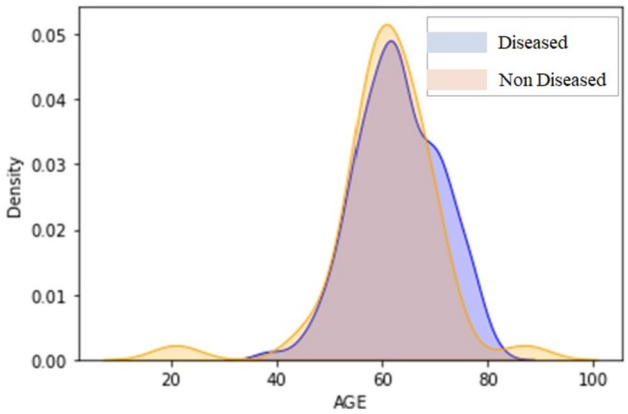


Fig. 3 KDE plots for patients with and without lung cancer based on the age distribution

3 Results and discussion

3.1 Result from exploratory data analysis

In Fig. 2, we present a heatmap illustrating the correlations between various features. Each cell within the heatmap is populated with values and colors, serving as indicators of the strength of the associations between these features. In this representation, warmer colors denote a higher correlation,

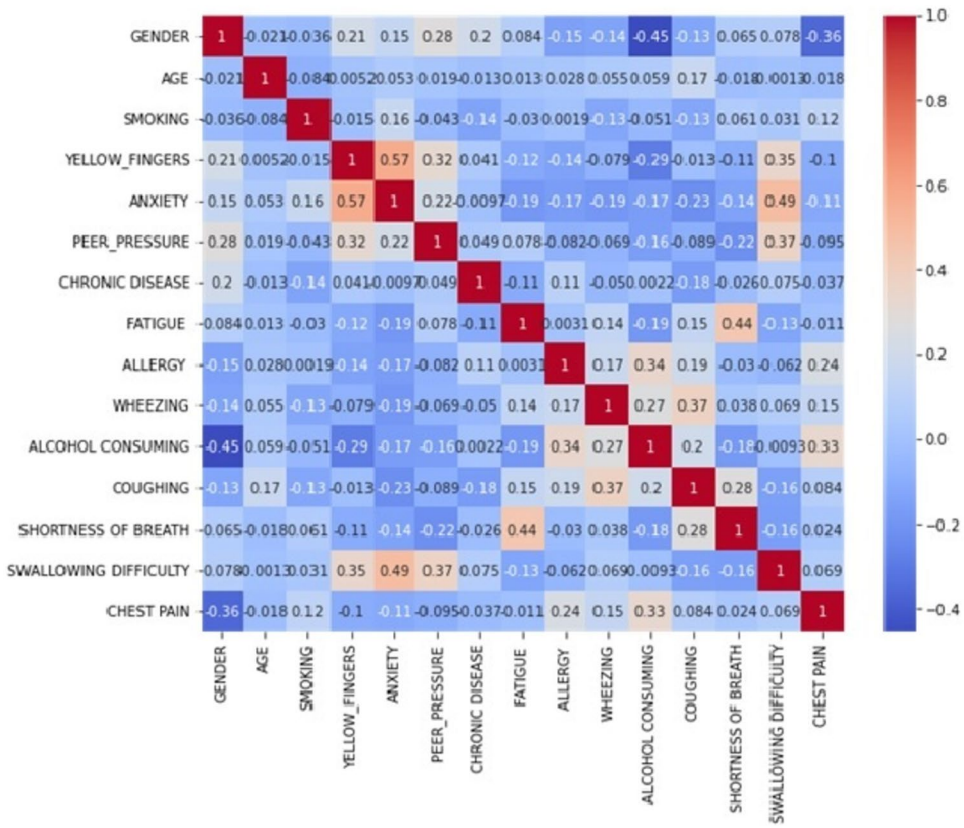


Fig. 2 Heatmap displaying correlations between all of the features

while cooler colors signify a lower degree of correlation among the features. Turning our attention to Fig. 3, it depicts the density distribution of a lung cancer dataset, differentiating between patients with and without the disease. Notably, the age group of 50–70 years emerges as the most significantly affected, with the peak incidence observed at 60 years of age.

3.2 Results of the proposed models

In this study, we processed a lung cancer dataset and applied various classification algorithms, including (SVM, ANN, and KNN). We conducted a comprehensive assessment of their performance on a dataset divided into 80% for training and 20% for testing.

Table 3 presents the performance parameters of the classification algorithms employed, specifically accuracy, sensitivity, specificity, Recall, precision, and F-Measure. Figure 4 illustrates the ROC curve for the applied classification algorithms.

Among the algorithms tested, SVM emerged as the top-performing individual algorithm, demonstrating an exceptional balance between accuracy, sensitivity, and specificity. It achieved the highest accuracy, reaching up to 96.80%, surpassing all other algorithms across various performance metrics, including sensitivity, recall, and F-Measure. This well-rounded performance positions SVM as the optimal choice for predicting lung cancer.

3.3 Optimization Results

Following the evaluation of the algorithms using performance metrics, our analysis revealed that the SVM outperformed the other algorithms, underscoring its superior performance. Nevertheless, it is important to acknowledge that while SVM displayed remarkable performance compared to other models in our study, its accuracy may not yet meet our desired threshold for Lung cancer prediction. As we continually aspire to achieve even greater accuracy in our predictive models, we remain committed to the pursuit of ongoing improvement and exploration of alternative approaches. In response to this challenge, we have introduced the CSABC-SVM hybrid model, designed to further enhance accuracy by capitalizing on the synergies of its constituent components. Specifically, the CSA and ABC algorithms are applied to optimize the hyperparameters of the SVM model. Our unwavering dedication to achieving heightened accuracy underscores our commitment to providing reliable predictions with precision in the domain of lung cancer prediction.

Table 3 Prediction performance of the classification algorithms

	Ac %	Se %	Sp %	Recall %	Pr %	F-Mesure %	MCC %
ANN	95.18	93.33	100	93.33	100	96.55	55.77
KNN	95.21	96.66	50	96.66	98.30	97.47	38.42
SVM	96.80	98.33	50	98.33	98.33	98,33	48.33

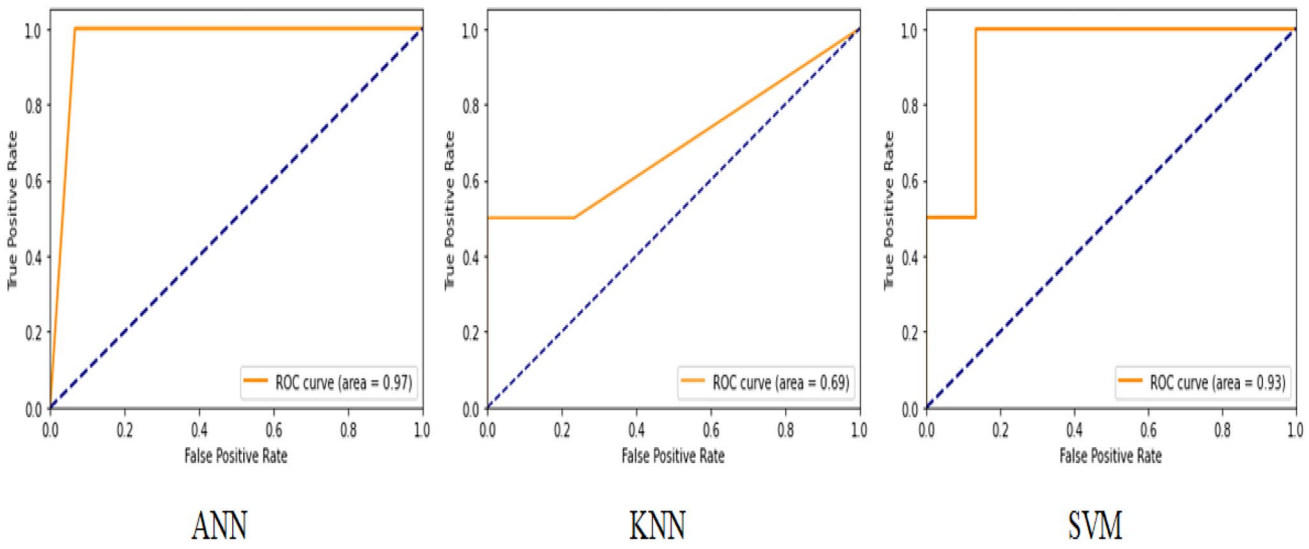


Fig. 4 ROC curve for the applied classification algorithms

The CSABC algorithm aims to find the best hyperparameters of the SVM classifier by developing a set of solutions through iterations, where each solution represents a pair of C and gamma values. The goal is to find the solution that achieves the highest accuracy on the training data.

The execution steps of the CSABC-SVM are as follows:

1. Splitting the Data

Divide the data set into training (80%) and test (20%).

2. Defining the CSABC Class:

We define the CSABC class responsible for optimizing SVM hyperparameters using the hybrid CSA and ABC algorithm. The constructor takes two parameters: population size (number of solutions), and maximum iterations.

3. Optimizing Hyperparameters:

Inside the optimized method of the CSABC class, a population of potential solutions (combinations of hyperparameters C and gamma) is randomly generated.

The population is iteratively evaluated for fitness. The fitness of each solution is determined by training an SVM with the given hyperparameters and computing its accuracy on the training data.

Updates the best solution (the hyperparameters that result in the highest accuracy) found during each iteration.

The population is evolved by selecting solutions based on their fitness and introducing mutations to explore new hyperparameter combinations.

The optimization process continues for 100 iterations.

4. Selecting the Best Solution:

- After the optimization process is completed, the best solution (best hyperparameters) is obtained from the CSA and ABC hybrid algorithm.

5. Extracting Hyperparameters:

The optimal values of the SVM hyperparameters C and gamma are extracted from the best solutions, which are used to construct the SVM classifier.

The performance of the new model was evaluated using metrics such as accuracy, sensitivity, specificity, precision, recall, F-measure, and MCC, as shown in Table 4.

The CSABC-SVM model delivered a remarkable accuracy rate of 98.36%, representing a substantial improvement of 1.58% when compared to the standalone SVM model. This noteworthy enhancement in prediction performance underscores the model's effectiveness, achieved through the synergistic combination of the CSA and ABC algorithms. Furthermore, the CSABC-SVM model exhibited a marked increase in both the MCC at 70.12% and the F-Measure at 99.17%. The hybridization of these algorithms played a pivotal role in elevating sensitivity and recall, both reaching 100%, a feat accomplished through the intelligent fusion of distinct optimization techniques. This increase bears significant implications for the model's capability to accurately identify positive lung cancer cases, a critical aspect in clinical decision-making and proactive patient care. The hybridization approach, as employed in the CSABC-SVM model, has effectively enhanced the overall performance of our predictive model, solidifying it as a robust and dependable tool for precise lung cancer prediction. This advancement ultimately contributes to improved healthcare outcomes for patients.

3.4 Feature importance

In the realm of ML, analyzing feature importance and visualizing it are significant and popular techniques [38]. Due to the ease of use and interpretability of feature ranking or risk analysis, these techniques are especially employed in fields like biology and social sciences [38]. Feature importance and ranking are determined based on the coefficient value of each characteristic [38, 40]. Figure 5 provides a graphical representation of feature importance for CSABC-SVM offering a clear illustration of feature order and importance. The figure effectively showcases the most influential features responsible for lung cancer. From Fig. 5 we can identify the seven most significant features based on their importance. According to Fig. 5, it is evident that age is a significant feature or factor for lung cancer identification and prediction. Followed by coughing, alcohol consuming, anxiety, wheezing, chest pain, and allergy.

Table 4 Prediction performance of the SVM and CSABC-SVM

	Ac %	Se %	Sp %	Recall %	Pr %	F-mesure %	MCC %
SVM	96.80	98.33	50	98.33	98.33	98,33	48.33
CSABC-SVM	98,38	100	50	100	98.36	99.17	70.12

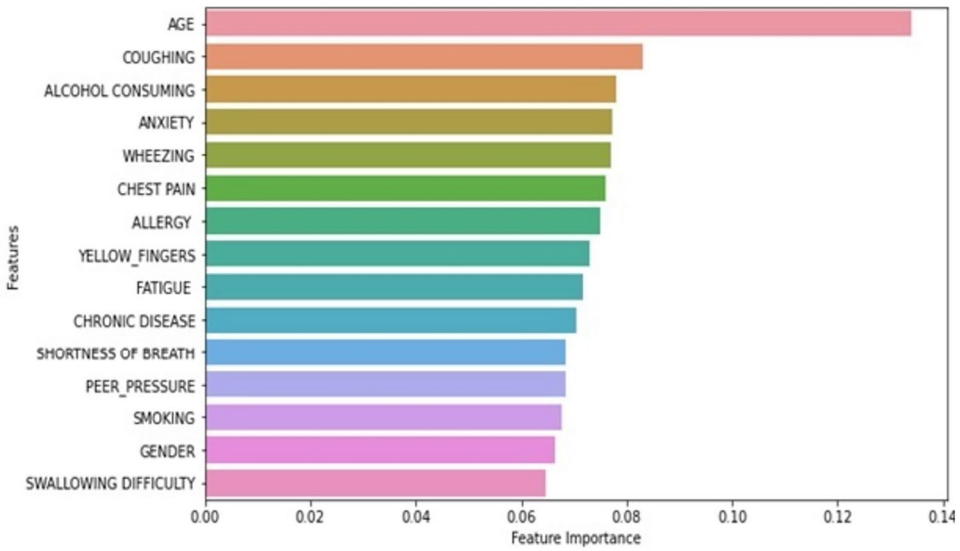


Fig. 5 Illustration of feature importance

Table 5 Comparison with other studies in the prediction of lung cancer

Author	Model	Ac (%)
[17]	RotF	97.1
[14]	Gradient Boosted Tree	90
[4]	XGBoost	94.42
[18]	PSO-GA-SVM	97.69
[19]	KNN-SVM	97.6
[21]	WOA_SVM	95
Proposed model	CSABC-SVM	98.38

4 Comparison with previous studies

Numerous high-quality studies have been conducted to detect and predict lung cancer. In comparison with other studies in the field of lung cancer prediction using AI, Table 5, our research stands out due to its comprehensive approach and novel hybridization technique. Our proposed model CSABC-SVM achieved a remarkable accuracy of 98.36%. While previous studies have explored various machine learning algorithms, including SVM, CNN, ANN, XGBoost, and RotF, our work not only conducted a thorough comparison of these algorithms but also introduced a new hybrid algorithm that combines the power of SVM and CSABC-SVM.

We underscore the significant benefits and superior predictive power that hybridization brings to the challenges of complex data analysis and predictive modelling. In this comparison, we used non-hybrid models such as SVM, CNN, and ANN because they help us to evaluate the performance of new and innovative techniques. Demonstrating superior performance compared to these models indicates that our approach has practical significance and real-world applicability. It allows us to attribute the improved accuracy to the hybridization itself, highlighting that the combination of different algorithms or techniques is the reason for our superior results. This emphasis on non-hybrid models as benchmarks is crucial because it helps us measure the true advancements and contributions made by hybrid models in addressing complex problems like lung cancer prediction, making them valuable tools for medical research and clinical applications.

5 Conclusion and future scope

Our research aims to enhance cancer prediction through an innovative hybridization optimization approach, presenting an intelligent framework for early-stage lung cancer diagnosis. Recognizing the crucial role of early detection in improving treatment success and reducing lung cancer-related mortality, our study establishes the foundation for an effective diagnostic approach. The evaluation of various machine learning models revealed that the CSABC-SVM hybrid model outperformed them all, achieving an exceptional accuracy of 98.36%. This high accuracy underscores the effectiveness and practicality of our approach for real-world clinical applications. However, limited lung cancer data in our dataset presented challenges.

In the future, our primary focus is refining the hybrid CSABC-SVM model by combining it with another optimization algorithm to explore novel ways of enhancing predictive performance. This iterative process will include a comprehensive analysis to identify and address potential limitations, further improving the model's accuracy in predicting lung cancer. Beyond lung cancer, we aim to

extend the application of this hybrid model to diagnose various cancers and predict other diseases, such as kidney disease and stroke. This extension will not only evaluate the model's robustness, reliability, and generalizability in different clinical contexts but will also incorporate advanced optimization techniques to advance predictive medicine.

Acknowledgements The authors appreciate the efforts of the (LBMPT) laboratory team for their encouragement throughout this project. The authors are also thankful to the anonymous reviewers of this manuscript for constructive observations and suggestions.

Author contributions All authors contributed to the study conception and design and material preparation, data collection, and analysis. All authors read and approved the final manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The dataset used in this article can be found at: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Anil Kumar C, Harish S, Ravi P et al (2022) Lung cancer prediction from text datasets using machine learning. Biomed Res Int. https://doi.org/10.1155/2022/6254177
2. Qureshi R, Basit SA, Shamsi JA et al (2022) Machine learning based personalized drug response prediction for lung cancer patients. Sci Rep 12:1–15. https://doi.org/10.1038/s41598-022-23649-0
3. Bhat AR, Hashmy R (2023) Hierarchical autoencoder-based multi-omics subtyping and prognosis prediction framework for lung adenocarcinoma. Int J Inf Technol 15:2541–2549. https://doi.org/10.1007/s41870-023-01310-x
4. Mamun M, Farjana A, Al Mamun M (2022) Ahammed MS (2022) Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. IEEE World AI IoT Congr AIIoT 2022:187–193. https://doi.org/10.1109/AIIoT54504.2022.9817326
5. Raoof SS, Jabbar MA, Fathima SA (2020) Lung cancer prediction using machine learning: a comprehensive approach. In: 2nd Int Conf Innov Mech Ind Appl ICIMIA 2020 - Conf Proc 108–115. https://doi.org/10.1109/ICIMIA48430.2020.9074947
6. Binczyk F, Prazuch W, Bozek P, Polanska J (2021) Radiomics and artificial intelligence in lung cancer screening. Transl Lung Cancer Res 10:1186–1199. https://doi.org/10.21037/tlcr-20-708
7. Obulesu O, Kallam S, Dhiman G et al (2021) Adaptive diagnosis of lung cancer by deep learning classification using wilcoxon gain and generator. J Healthc Eng. https://doi.org/10.1155/2021/5912051
8. Xie Y, Meng WY, Li RZ et al (2021) Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl Oncol. https://doi.org/10.1016/j.tranon.2020.100907
9. Patil S, Patil KR, Patil CR, Patil SS (2020) Performance overview of an artificial intelligence in biomedics: a systematic approach. Int J Inf Technol 12:963–973. https://doi.org/10.1007/s41870-018-0243-8
10. Yadav RS (2020) Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. Int J Inf Technol 12:1321–1330. https://doi.org/10.1007/s41870-020-00484-y
11. Gunasinghe AD, Aponso AC, Thirimanna H (2019) Early prediction of lung diseases. In: 2019 IEEE 5th Int Conf Converg Technol I2CT 2019. https://doi.org/10.1109/I2CT45611.2019.9033668
12. Wu CC, Yeh WC, Hsu WD et al (2019) Prediction of fatty liver disease using machine learning algorithms. Comput Methods Programs Biomed 170:23–29. https://doi.org/10.1016/j.cmpb.2018.12.032
13. Singh GAP, Gupta PK (2019) Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. Neural Comput Appl 31:6863–6877. https://doi.org/10.1007/s00521-018-3518-x
14. Faisal MI, Bashir S, Khan ZS, Hassan Khan F (2019) An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In: 2018 3rd Int Conf Emerg Trends Eng Sci Technol ICEEST 2018 1–4. https://doi.org/10.1109/ICEEST.2018.8643311
15. Wu J, Zan X, Gao L et al (2019) A machine learning method for identifying lung cancer based on routine blood indices: Qualitative feasibility study. JMIR Med Inform 7:1–12. https://doi.org/10.2196/13476
16. Lakshmanaprabu SK, Mohanty SN, Shankar K et al (2019) Optimal deep learning model for classification of lung cancer on CT images. Futur Gener Comput Syst 92:374–382. https://doi.org/10.1016/j.future.2018.10.009
17. Dritsas E, Trigka M (2022) Lung Cancer Risk Prediction with Machine Learning Models. Big Data Cogn Comput. https://doi.org/10.3390/bdcc6040139
18. Maulidina F, Rustam Z, Pandelaki J (2021) Lung Cancer classification using support vector machine and hybrid particle swarm optimization-genetic algorithm. In: 2021 Int Conf Decis Aid Sci Appl DASA 2021 751–755. https://doi.org/10.1109/DASA53625.2021.9682259
19. Vijila Rani K, Joseph Jawhar S (2022) Lung lesion classification scheme using optimization techniques and hybrid (KNN-SVM) Classifier. IETE J Res 68:1485–1499. https://doi.org/10.1080/03772063.2019.1654935
20. Nanglia P, Kumar S, Mahajan AN et al (2021) A hybrid algorithm for lung cancer classification using SVM and Neural Networks. ICT Express 7:335–341. https://doi.org/10.1016/j.icte.2020.06.007
21. Vijh S, Gaur D, Kumar S (2020) An intelligent lung tumor diagnosis system using whale optimization algorithm and support vector machine. Int J Syst Assur Eng Manag 11:374–384. https://doi.org/10.1007/s13198-019-00866-x
22. Lung Cancer Prediction Dataset. https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. Accessed 22 Jan 2023
23. Sarwar A, Ali M, Manhas J, Sharma V (2020) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. Int J Inf Technol 12:419–428. https://doi.org/10.1007/s41870-018-0270-5
24. Rashidi HH, Tran NK, Betts EV et al (2019) Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Acad Pathol 6:2374289519873088. https://doi.org/10.1177/2374289519873088

25. Laidi M, El Hadj AA, Si-Moussa C et al (2021) CMC of diverse gemini surfactants modeling using a hybrid approach combining SVR-DA. Chem Ind Chem Eng Q 27:299–312. https://doi.org/10.2298/CICEQ200907048L
26. Mishra I, Mohapatra S (2023) An enhanced approach for analyzing the performance of heart stroke prediction with machine learning techniques. Int J Inf Technol 15:3257–3270. https://doi.org/10.1007/s41870-023-01321-8
27. Tran NK, Sen S, Palmieri TL et al (2019) Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. Burns 45:1350–1358. https://doi.org/10.1016/j.burns.2019.03.021
28. Khanday AMUD, Rabani ST, Khan QR et al (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 12:731–739. https://doi.org/10.1007/s41870-020-00495-9
29. Madani A, Benkortbi O, Laidi M (2023) In silico prediction of the inhibition of new molecules on SARS-CoV-2 3CL protease by using QSAR: PSOSVR approach. Brazil J Chem Eng. https://doi.org/10.1007/s43153-023-00332-z
30. Rahi P, Sood SP, Bajaj R, Kumar Y (2021) Air quality monitoring for Smart eHealth system using firefly optimization and support vector machine. Int J Inf Technol 13:1847–1859. https://doi.org/10.1007/s41870-021-00778-9
31. Yasen M, Al-Madi N, Obeid N (2018) Optimizing neural networks using dragonfly algorithm for medical prediction. In: 2018 8th Int Conf Comput Sci Inf Technol CSIT 2018 71–76. https://doi.org/10.1109/CSIT.2018.8486178
32. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Erciyes university, engineering faculty, computer …
33. Euldji I, Belghait A, Si-Moussa C et al (2023) A new hybrid quantitative structure property relationships-support vector regression (QSPR-SVR) approach for predicting the solubility of drug compounds in supercritical carbon dioxide. AIChE J. https://doi.org/10.1002/aic.18115
34. Askarzadeh A (2016) A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. Comput Struct 169:1–12. https://doi.org/10.1016/j.compstruc.2016.03.001
35. Dubey AK, Sinhal AK, Sharma R (2023) Heart disease classification through crow intelligence optimization-based deep learning approach. Int J Inf Technol. https://doi.org/10.1007/s41870-023-01445-x
36. Sayed GI, Hassanien AE, Azar AT (2019) Feature selection via a novel chaotic crow search algorithm. Neural Comput Appl 31:171–188. https://doi.org/10.1007/s00521-017-2988-6
37. Polat K, Güneş S (2007) Breast cancer diagnosis using least square support vector machine. Digit Signal Process A Rev J 17:694–701. https://doi.org/10.1016/j.dsp.2006.10.008
38. Ali MM, Paul BK, Ahmed K et al (2021) Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Comput Biol Med 136:104672. https://doi.org/10.1016/j.compbiomed.2021.104672
39. Asaduzzaman S, Ahmed MR, Rehana H et al (2021) Machine learning to reveal an astute risk predictive framework for Gynecologic Cancer and its impact on women psychology: Bangladeshi perspective. BMC Bioinformatics 22:1–17. https://doi.org/10.1186/s12859-021-04131-6
40. Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P (2012) Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. Bioinformatics 28:1766–1774. https://doi.org/10.1093/bioinformatics/bts238

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.