Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 10: 1–15
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241249661
journals.sagepub.com/home/dhj

Sage

Optimizing lung cancer classification through hyperparameter tuning

Syed Muhammad Nabeel[1], Sibghat Ullah Bazai[1], Nada Alasbali[2], Yifan Liu[3], Muhammad Imran Ghafoor[4], Rozi Khan[5], Chin Soon Ku[6], Jing Yang[7], Sana Shahab[8] and Lip Yee Por[7]

Abstract

Artificial intelligence is steadily permeating various sectors, including healthcare. This research specifically addresses lung cancer, the world's deadliest disease with the highest mortality rate. Two primary factors contribute to its onset: genetic predisposition and environmental factors, such as smoking and exposure to pollutants. Recognizing the need for more effective diagnosis techniques, our study embarked on devising a machine learning strategy tailored to boost precision in lung cancer detection. Our aim was to devise a diagnostic method that is both less invasive and cost-effective. To this end, we proposed four methods, benchmarking them against prevalent techniques using a universally recognized dataset from Kaggle. Among our methods, one emerged as particularly promising, outperforming the competition in accuracy, precision and sensitivity. This method utilized hyperparameter tuning, focusing on the Gamma and C parameters, which were set at a value of 10. These parameters influence kernel width and regularization strength, respectively. As a result, we achieved an accuracy of 99.16%, a precision of 98% and a sensitivity rate of 100%. In conclusion, our enhanced prediction mechanism has proven to surpass traditional and contemporary strategies in lung cancer detection.

Keywords

Lung cancer, hyperparameter tuning, machine learning, XGBoost, support vector machines, decision trees, and logistic regression

Submission date: 4 October 2023; Acceptance date: 4 April 2024

[1] Department of Computer Engineering, Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS), Quetta, Balochistan, Pakistan

[2] Department of Informatics and Computing Systems, College of Computer Science, King Khalid University, Abha, Saudi Arabia

[3] Department of Electronic Science, Binhai College of Nankai University, Tianjing, China

[4] Department of Engineering, Pakistan Television Corporation, Lahore, Pakistan

[5] Department of Computer Science, National University of Sciences and Technology (NUST) Balochistan Campus Quetta, Quetta, Balochistan, Pakistan

[6] Department of Computer Science, Universiti Tunku Abdul Rahman, Kampar, Malaysia

[7] Department of Computer System and Technology, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia

[8] Department of Business Administration, College of Business Administration, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

Corresponding author:

Syed Muhammad Nabeel, Department of Computer Engineering, Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS). Quetta, Balochistan, Pakistan.
Email: nabeelm71@yahoo.com

Sibghat Ullah Bazai, Department of Computer Engineering, Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS). Quetta, Balochistan, Pakistan.
Email: sibghat.ullah@buitms.edu.pk

Chin Soon Ku, Department of Computer Science, Universiti Tunku Abdul Rahman, Kampar 31900, Malaysia.
Email: kucs@utar.edu.my

Lip Yee Por, Department of Computer System and Technology, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia.
Email: porlip@um.edu.my

Introduction

The human body is made up of trillions of cells, which typically grow and transform themselves into new cells through cell division over time. During this process, cells become old and damaged, with damaged cells replaced by new ones. However, this process can become ineffective, leading to the unchecked growth of damaged or abnormal cells, which can be either cancerous or non-cancerous (benign).[1] Damaged cells may affect nearby body tissues and have the potential to spread to other parts of the body, forming additional tumors through a process called metastasis.[2]

As previously explained, lung cancer is primarily a genetic disease.[3] However, external factors such as smoking tobacco, exposure to passive smoking and toxic gases, as well as polluted air, can also contribute to its development.[4,5] These factors disrupt the cell division process and promote the exponential growth of damaged cells that would typically die or stop growing, ultimately leading to the formation of cancerous tumors.

Lung cancer is among the most prevalent and deadly malignancies globally. According to statistics from the World Health Organization, out of the 10 million cancer cases registered in 2020, almost 2.21 million were lung cancer cases.[6] Approximately 1.8 million people die from lung cancer annually, making it the leading cause of cancer-related deaths. Lung cancer has four stages; while detection at stages 1 and 2 is curable, these stages typically show no symptoms. Symptoms usually manifest in stages 3 and 4, which are extremely dangerous. Malignant cancer in these stages spreads rapidly and damages nearby tissues and organs, while benign tumors grow slowly and can be removed through surgery, therapy or other techniques.

The challenge lies in the early detection of lung cancer, as it often goes undetected during its initial stages, resulting in low survival rates. Currently, paramedics and doctors employ various diagnostic methods, including mammography, CT scans and MRI images. Anti-cancer treatments utilize a range of techniques, including laser surgery, radiation, chemical alterations and transplantation. However, these methods are costly, time-consuming and cumbersome.[7]

Early identification is crucial for improving the chances of survival among lung cancer patients. Machine learning (ML) techniques hold the key to maximizing survival rates in this and other medical research areas. Various ML algorithms are applied in this study for the early diagnosis of lung cancer.[8] Furthermore, this study improves the performance of ML algorithms over previous studies by incorporating additional characteristics and parameters into the dataset. The four ML algorithms used in this study are extreme gradient boosting (XGB), support vector machine (SVM), decision tree (DT) and linear regression (LR). These models were selected based on a comprehensive literature review, demonstrating strong performance in terms of accuracy and precision.

The subsequent section of this paper examines several studies aimed at early stage lung cancer detection using different ML models on diverse datasets collected and uploaded to public repositories. The remaining blank area represents the prediction rate, which can be true positive (TP), true negative (TN), false positive (FP) or false negative (FN). The accuracy of a model's predictions, distinguishing between true and false positives and negatives, largely depends on the prediction rate.

The focus of this study is to detect lung cancer using ML techniques, employing SVM, XGB, DT and LR models. While relevant literature was reviewed for this research, the primary focus was on improving precision, accuracy and the F1-score. The results of this study show greater promise compared to previous studies discussed in the subsequent section.

The following sections of this article cover a variety of topics. The second section 2 provides a comprehensive literature review. The third section briefly introduces the dataset, preprocessing methods, techniques used and evaluation metrics. The fourth section analyzes the results and compares them with current publications. Finally, the fifth section concludes the article and outlines potential future directions.

Related works

The existing literature on lung cancer detection and classification offers a range of methodologies and findings. Several factors contribute to tumor development, including genetic inheritance, smoking, and alcohol consumption. Various techniques, such as narrow-band imaging (NBI) and broncho-over-white light (WL), have been utilized for the analysis of lung cancer, with NBI being recognized for its higher accuracy in early-stage detection.[9]

Mamun et al.[10] explored ensemble learning methods, including XGB and light gradient boosting machines (LightGBM), for lung cancer classification, employing an oversampling technique, SMOTE, for enhanced results. This study introduces new ensemble learning models developed based on a survey dataset of 309 individuals with or without lung cancer. The results indicate that the XGBoost (XGB) ensemble learning technique outperformed other methods, achieving an accuracy of 94.42%, precision of 95.66%, recall of 94.46%, F1-score of 94.74% and an area under the curve (AUC) of 98.14%.

Similarly, Aljabar et al.[11] used the eXtreme Gradient Boost machine learning model to process a compiled dataset of diverse lung cancer types, showing improved performance in comparison to other models. This study utilizes gene expression and transcriptome data, specifically from real new generation RNA_seq (NGS) and microarray datasets. The results of the study demonstrate the effectiveness of the XGB model in improving the prediction of lung cancer diagnosis, detection, and relapse.

Multiple studies have utilized various ML models, such as SVM, K-nearest neighbors (KNN) and convolutional neural networks (CNN), for analyzing CT scans and other datasets for efficient and accurate lung cancer classification, each offering unique insights into model accuracy and performance.[12–19]

Significant efforts have been invested in enhancing lung cancer detection through innovative methods. A study employing a device named e-nose used exhaled breath samples for classifying lung cancer, asthma and COPD, offering promising yet improvable results.[16] Deep learning applied to chest X-ray films, as conducted by Ausawalaithong et al.,[17] demonstrated the potential of ML algorithms in cancer detection. Various ML models, including gradient boosting machines (GBM) and custom ensembles, were analyzed for their efficacy in lung cancer detection based on diverse features, underscoring the supremacy of GBM.[18]

Studies have also focused on early-stage lung tumor detection, using models such as SVM and CNN to achieve high accuracy.[19] El Guabassi et al. have explored and discussed the employment of artificial intelligence in lung cancer detection, showcasing a comparative analysis of various ML models for enhanced cancer detection.[20] This study integrates AI techniques, specifically artificial neural networks (ANN), Naive Bayes (NB), KNN, SVM, DT and LR, for the early diagnosis and prediction of lung cancer. This result indicated that SVM achieves a high prediction accuracy of 94.6% and has the potential to automatically predict lung cancer with a high degree of accuracy, highlighting its robustness in this particular medical application.

Early detection of lung cancer has been addressed by Abdullah et al.,[21] who investigate and identify the most effective classifier among SVM, KNN and CNN for early detection, ultimately aiming to contribute to the improvement of prognosis and outcomes for patients with lung cancer. The proposed solution involves applying SVM, KNN and CNN to datasets obtained from the UCI ML repository, which contain information about patients affected by lung cancer. The use of these classifiers is facilitated by the WEKA tool. The experimental results indicate that SVM outperforms the other classifiers with an accuracy of 95.56%. CNN follows with an accuracy of 92.11%, while KNN achieves an accuracy of 88.40%.

Pradhan et al.'s goal is to improve lung cancer classification accuracy using an ensemble learning model, ultimately aiding in early diagnosis and treatment.[22] The author of this study combined ensemble learning with recurrent neural networks (RNN) and the best fitness-based squirrel search algorithm (BF-SSA) feature extraction method to address time and memory complexities and a commitment to achieving higher accuracy in practical applications. The datasets used were gathered from Kaggle and the UCI repository with 7 attributes and 56 attributes, respectively. The accuracy, precision, F1-score and recall of the proposed model are above 92% for both datasets, which shows better than CNN, neural network (NN), RNN, 5LEVEL-RNN and Self-Adaptive-Sea Lion Optimization-Recurrent Neural Networks (SA-SLnO-RNN) algorithms on lung disease diagnosis.

Yamini et al. investigate and test ML approaches for early detection and intervention, potentially saving lives in the face of the lung cancer epidemic.[23] This study proposed the LR, DT, RF, SVM, XGB classifier, gradient boosting and KNN to analyze lung cancer data. It offers insights into the performance of different ML models in the context of lung cancer prediction. The dataset contains a total of 16 variables, 15 of which are input variables and one of which is a class label. The results of the research indicate that the XGB classifier outperforms other ML models such as LR, DT, RF, SVM, gradient boosting and KNN in terms of accuracy when predicting lung cancer. In addition, the accuracy, precision, F1-score and recall of the proposed XGB classifier are 99.1%, 100.0%, 99.0% and 98.0%, respectively.

Singh et al. enhance the diagnosis and early detection of lung cancer, a highly fatal disease with challenging characteristics, through the application of ML techniques such as RF, XGB, DT, AdaB, SVM, GBM, LightGBM and Cat Boost.[24] This study focuses on improving and evaluating the classifier learning systems' accuracy using specific models, with AdaBoost and XGB identified as top performers. Two datasets were used and gathered from Kaggle: the first dataset contains 309 records and 16 attributes and the second dataset contains 1000 records and 24 attributes. The results highlight the superior performance of the AdaB and XGB models, with accuracy rates of 96.77% and 96.75%, respectively.

Addressing the critical issue of early lung cancer diagnosis, the literature reveals extensive use of ML models optimized with diverse hyperparameters and datasets for enhanced prediction and classification results.[25–31] Despite achieving commendable accuracy in lung cancer detection, continuous efforts are being made to improve these figures to reduce the chances of false negative or false positive detections, as highlighted in various studies.[32,33] The exploration and application of numerous ML models, from ensemble learning methods to deep learning, underscore the ongoing pursuit of enhancing lung cancer detection and classification.

To enhance the effectiveness of lung cancer diagnosis, this study, inspired by prior research, employed similar ML models, including SVM, XGB, LR and DT, with a focus on optimizing the accuracy of each model individually. The goal of convergence toward 100% accuracy aims to minimize the possibility of erroneous lung cancer detection in patients. This endeavor has resulted in achieving more accurate and reliable results based on essential metrics such as precision, accuracy and F1-score.

Recent studies further expand on this pursuit. Alsinglai et al. introduced a framework for managing lung cancer patients with imbalanced datasets, employing models like XGB, LR and random forest (RF) and achieving an impressive accuracy of 98% with RF.[34] Other studies have explored the usage of AdaBoost and CNN on CT scan images, reporting significant improvements in accuracy,[35] and diverse ML models with Chi-square for feature selection, also reporting enhanced results.[36] The evaluation and comparison of various ML models on extensive datasets have consistently highlighted the potential of AdaBoost to outperform other models with high accuracy and AUC.[37]

In the quest for optimizing lung cancer prognosis, studies like that of Wu et al. have adopted cost-effective and efficient methods, such as blood tests, to employ ML models like RF for accurate lung cancer detection, achieving remarkable results in terms of accuracy, recall and AUC.[38]

The extensive datasets used in the research mentioned above, ranging from the Kaggle dataset to the Lanzhou University dataset, highlight the diverse sources for obtaining reliable and comprehensive data for enhancing lung cancer prediction and classification.[10,12,19,20,31,34,36–38]

In conclusion, the comprehensive literature review underscores the extensive and varied research conducted in the field of lung cancer detection and classification. Despite the impressive results achieved by numerous studies, the perpetual quest for enhancing accuracy and reducing errors remains paramount. This study aims to significantly contribute to this ongoing pursuit, providing valuable insights and findings that potentially augment the body of knowledge in this critical healthcare domain. A concise overview of the datasets, models, and results from selected related research studies is shown in Table 1.

Methods

Dataset

The dataset used for this research was taken from a public repository named Kaggle.[39] The data available in the dataset contained 309 entries and had a total of 16 attributes: 15 independent and 1 dependent, which are described in Table 2.

Preprocessing

Appropriately preprocessing the data before feeding it into the proposed methods can significantly improve their performance. Figure 1 illustrates the distribution of cancerous and non-cancerous cases across different age groups. The total number of cases is represented in blue, with cancerous cases in orange and non-cancerous cases in gray. It is notable that the majority of lung cancer cases occur within the age bracket of 50–70 years, with significantly fewer cases below 50 and above 70.

In this research, we proposed four methods: SVM, DT, LR and XGB. Our proposed SVM and LR methods require normalization (converting categorical to numeric data) or scaling of input data. SVM, in particular, is sensitive to the scale of the input features as it seeks to maximize the margin between classes, leading to biased models with upscaled data. Similarly, LR is sensitive to input feature scaling as it aims to fit a linear decision boundary between classes. Conversely, our proposed DT and XGB methods are not sensitive to input feature scaling and do not require normalization or scaling. Consequently, several preprocessing operations were carried out, including converting strings to numeric values, removing duplicates, implementing random oversampling for data balancing and scaling values between 0 and 1 before model training, to ensure compatibility with our proposed methods.

Following the preprocessing step, the dataset was divided into two sets: a training dataset and a testing dataset. The split was determined using the “train_test_split” method, with a parameter for “random_state” to ensure reproducibility, resulting in a 75% and 25% split.

Proposed methods

Figure 2 illustrates the block diagram of our proposed methodology. Drawing from thorough research, we designed and tested four proposed methods using a standard lung cancer dataset. The methods we selected include SVM, DT, XGB, and logistic regression (LR). Each of these models was improvised and trained using both default and tuned parameters.

ML models serve as invaluable assets in addressing complex challenges across various domains. Each model is characterized by a unique suite of hyperparameters, and these parameters play a pivotal role in shaping the model’s performance and accuracy. To optimize these parameters, we employed the ParameterGrid method, allowing us to explore a range of parameter values beyond the defaults. Initially, we trialed single values to discern the optimal range, establishing these ranges by considering the extreme values for each parameter. From this analysis, the ParameterGrid was crafted—a grid designed to methodically assess each model autonomously, ensuring superior results. This fine-tuning phase was integral to our research, as it markedly enhanced the outcomes of our models. After this optimization process, we identified the fine-tuned parameters for each method, which are detailed as follows:

The proposed SVM method. Figure 3 illustrates the structure of the proposed SVM method. Our method comprises 16 variables, with 15 independent attributes as input data and 1 dependent attribute as classified output (see Figure 2(b)). These variables connect to a hidden layer, either from the input or output side. Notably, the method self-determines the number of hidden layers based on the

Table 1. Synthesis and summary of the selected related work.

Research Works	Dataset	Method(s) Used	Result for Proposed Method
El Guabassi et al.[20]	Kaggle dataset (309)	SVM (Proposed by authors), NN, Naive Bayes, k-NN, DT, and LR	Accuracy: 94.6%
Mamun et al.[10]	Kaggle dataset (309)	XGB (Proposed by authors), LightGBM, Bagging, and AdaBoost	Accuracy: 94.42% Recall: 94.46% Precision: 95.6% AUC: 98.1%
Alsinglawi et al.[34]	MIMIC-III data (53,423)	RF(Proposed by authors), XGB, and LR.	AUC: 98% Recall: 98%
Anil Kumar et al.[12]	University of California, Irvine.	SVM	98.80%
Venkatesh et al.[35]	SEER dataset (1000)	Ada-boost (Proposed by authors), Bagging, K-NN, DT, NN	Accuracy 98.2%
Prabhpreet et al.[14]	Data World dataset (1000), 25 attributes	SVM (Proposed by authors), and RF	Accuracy: 97.9% Precision: 99.9% F1-score: 99.9% Recall: 99.9%
Puneet et al.[30]	Lanzhou University (277)	XGB (Proposed by authors), GridSearchCV, LR, SVM, Gaussian NB, DT, and K-NN	Accuracy: 92.2% Recall: 96.7% AUC: 95%
Sim et al.[37]	HRQOL data (809)	AdaBoost (Proposed by authors), LR, DT, RF, and Bagging	Accuracy: 94.8% AUC: 94.9%
Patra[31]	UCI repository (32)	RBFN (Proposed by authors), K-NN, J48, SVM, LR, A-NN, NB, and RF	Accuracy: 81.3% Precision: 81.29% Recall: 81.29% F1-score: 81.29% AUC: 75%
Benusiglio et al.[3]	Data World dataset (1000)	SVM (Proposed by authors), Naïve Bayes, DT, and LR	Accuracy: 99%
Wu et al.[38]	Lanzhou University (277)	Random Forest (Proposed by authors)	Accuracy: 96% Recall: 96.29% AUC: 99%
Faisal et al.[19]	UCI repository (32)	Gradient BT(Proposed by authors), MLP, NN, NB, SVM, Majority Voting, and RF	Accuracy: 90.1%, Precision: 88% F1-score: 85.7% Recall: 83.7%
Dritsas et al.[32]	UCI repository (15750)	RF (proposed by authors), DT, SVM, NB, K-NN, SG, and multilayer perceptron.	Accuracy: 97.1% Precision: 97.1% F1-score: 97.1%
Vieira et al.[33]		SVM (Proposed by authors), K-NN, NB, DT, and ANN	Accuracy: 88.5%

(continued)

Table 1. Continued.

Research Works	Dataset	Method(s) Used	Result for Proposed Method
			Precision: 91.3% F1-score: 95.8%
Pradhan et al.[22]	Kaggle dataset (59)	BF-SSA-HR-DEL (Proposed by authors), CNN, NN, RNN, 5LEVEL-RNN, and SA-SLnO-RNN	Accuracy: 93.0% Precision: 93.4% F1-score: 93.1% Recall: 92.9%
Pradhan et al.[22]	UCI repository	BF-SSA-HR-DEL (Proposed by authors), CNN, NN, RNN, 5LEVEL-RNN, and SA-SLnO-RNN	Accuracy: 93.2% Precision: 93.4% F1-score: 93.3% Recall: 93.1%
Yamini et al.[23]		XGB (Proposed by authors), LR, DT, RF, SVM, Gradient Boosting, and KNN	Accuracy: 99.1% Precision: 100.0% F1-score: 99.0% Recall: 98.0%
Singh et al.[24]	Kaggle (309) + Kaggle (1000)	AdaBoost and XGB (Proposed by authors), RF, DT, SVM, GBM, LightGBM, and Cat Boost	AdaBoost Accuracy: 96.77% Precision: 96.62% F1-score: 98.28% Recall: 100.00% XGB Accuracy: 96.77% Precision: 97.70% F1-score: 98.26% Recall: 98.83%

input, with hyperparameters provided as function arguments. To achieve the objective of classifying a tumor as malignant (cancerous) or benign (non-cancerous), we adjust the values of the hyperparameters gamma and C. Figure 2(b) illustrates the bidirectional relationship between the inputs and outputs of each layer. Every layer communicates with both its preceding and succeeding layers, ensuring the accuracy of predictions. If inaccuracies are detected in the predictions of subsequent layers, parameters are adjusted iteratively to achieve a finely-tuned and precise result. This iterative process involves generating an output, applying hyperparameters, storing the modified output, and comparing the two. The superior output, in terms of accuracy, is recognized as the "tuned" output.

The proposed SVM method in this case involves a minor change in terms of optimization: hyperparameter tuning. The parameters were selected from a pre-defined range of values, and the mathematics behind the hyperparameter tuning and optimization are as follows:

Kernel selection:

$$K\ (x_i,\ y_j) = \exp(-\ \gamma\ ||x_i,\ y_j||^2) \qquad (1)$$

where γ is the hyperparameter controlling the influence and the objective function is:

$$minimize\ \frac{1}{2}\ ||W||^2 + C\ \sum_{i=1}^{m} max\ (0,\ 1 - \ y_i\ (w.x_i\ + b) \qquad (2)$$

where $minimize\ \frac{1}{2}\ ||W||^2$ is the regularization term that balances the maximization margin and minimization of training error and C is the trade-off between allowing training errors and forcing rigid margins. In optimization, the results were checked between $[10^{-3},\ 10^{-2},\ \ldots 10^3]$. Existing parameters may vary depending on the required methodology; however, gamma and C are the most common among them. The difference lies in the selection of parameter values in a pre-defined grid.

Our proposed SVM method employs the Gamma and C parameters to dictate the kernel width and regularization strength, respectively. Initially, both parameters are set at a default value of 1 and are of the float data type. While the C parameter can take any value, gamma is restricted to non-negative numbers. Our fine-tuned SVM parameters

Table 2. Dataset attributes and their possible outputs.

Sr. #	Attribute	Value 1	Value 2	Description
1	Gender	M(male)	F(female)	Defines that patient was either male or female[40]
2	Age	Patient age		Defines the age of the patient
3	Wheezing	Y = 2	N = 1	Defines whether the patient has suffered wheezing[14]
4	Swallowing difficulty	Y = 2	N = 1	Defines whether the patient has suffered difficulty swallowing[41]
5	Yellow fingers	Y = 2	N = 1	Defines whether the patient has the symptom of a yellow finger[42]
6	Chronic Disease	Y = 2	N = 1	Defines whether the patient has a chronic disease symptom[4]
7	Anxiety	Y = 2	N = 1	Defines whether the patient has anxiety[14]
8	Coughing	Y = 2	N = 1	Defines whether the patient has a coughing issue[14]
9	Alcohol	Y = 2	N = 1	Defines whether the patient is an alcoholic[9]
10	Chest pain	Y = 2	N = 1	Defines whether the patient has a chest pain symptom[43]
11	Lung cancer	Y = 2	N = 1	Defines whether the patient has lung cancer[44]
12	Allergy	Y = 2	N = 1	Defines whether the patient has any allergies[45]
13	Smoking	Y = 2	N = 1	Defines whether the patient is a smoker[46]
14	Peer pressure	Y = 2	N = 1	Defines whether the patient has peer pressure symptom[47]
15	Shortness of breath	Y = 2	N = 1	Defines whether the patient feels short of breath[48]
16	Fatigue	Y = 2	N = 1	Defines whether the patient feels fatigue[49]

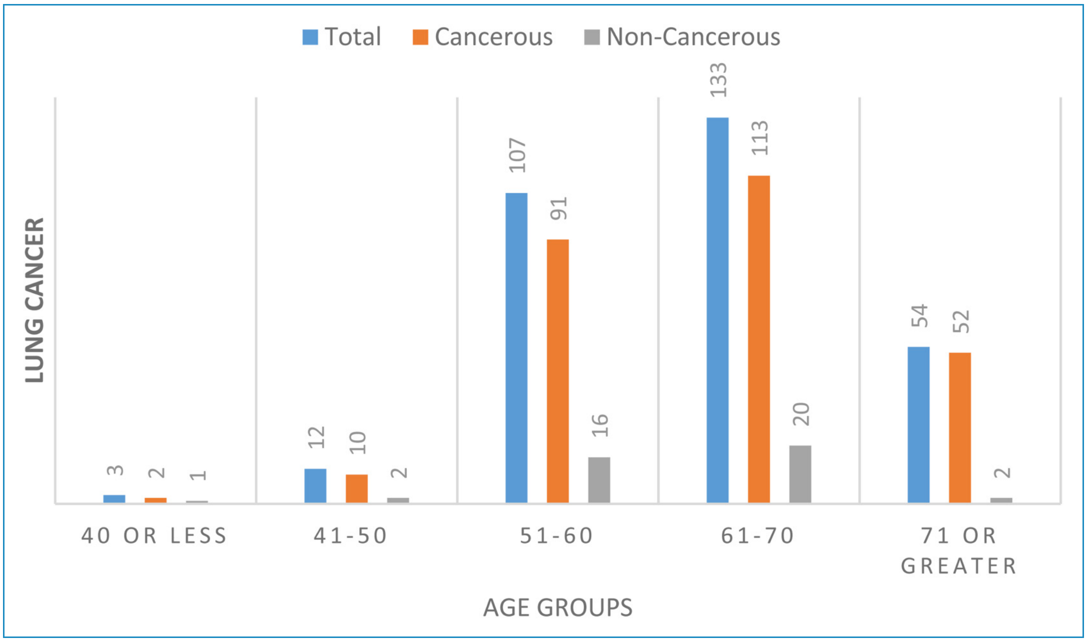

Figure 1. Number of lung cancers versus different age groups.

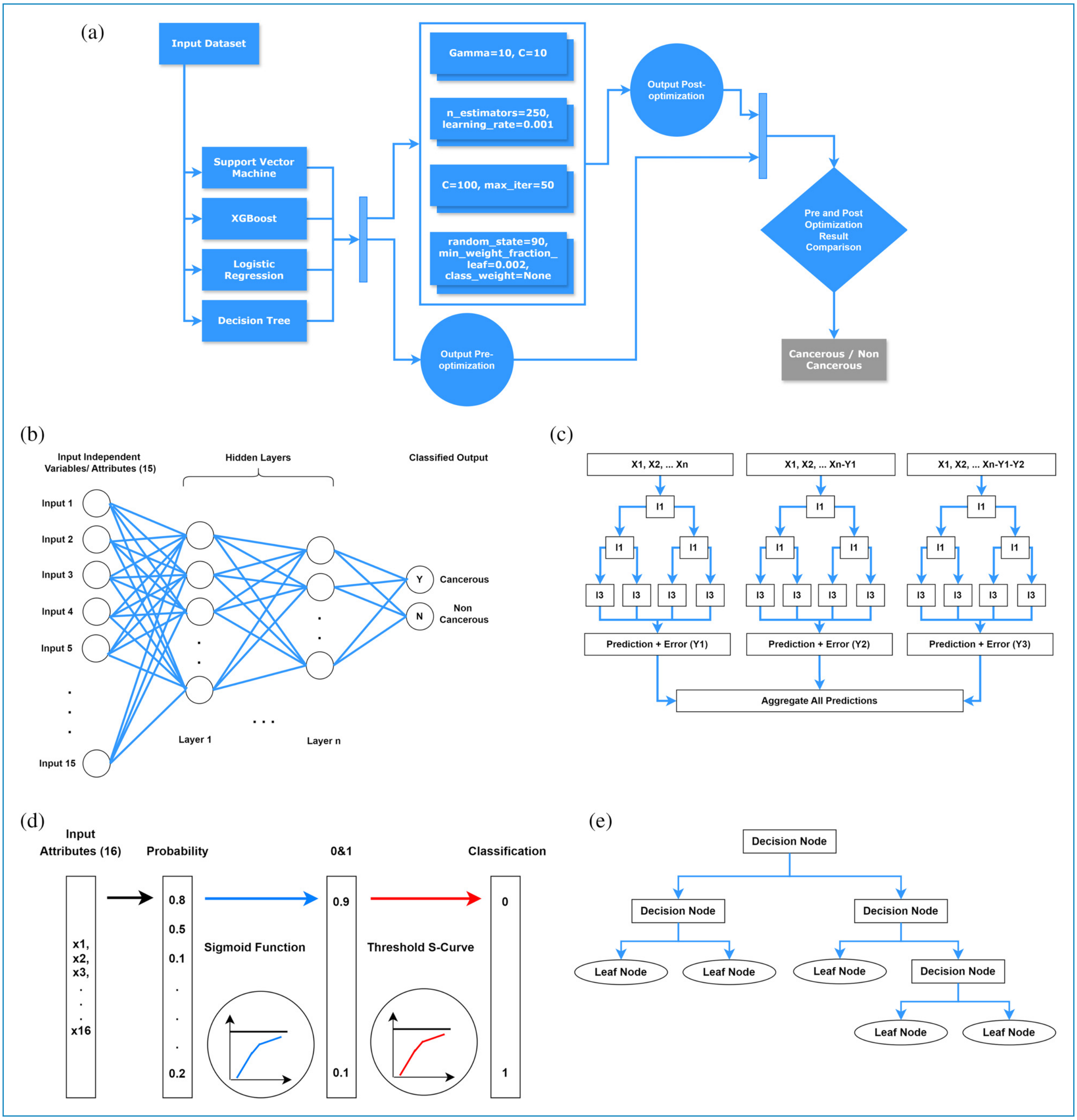

Figure 2. Block diagram of our proposed methodology (a) overview; (b) SVM; (c) XGB; (d) Logistic regression; (e) DT.

allow for improved adaptability to the dataset and problem at hand. In some cases, default SVM parameters may not yield optimal results, and our approach ensures better generalization and performance.

The proposed XGB method. The proposed XGB method employed the "n_estimators" and "learning_rate" parameters to regulate the number of DT and the optimization step size.[50] Additionally, default parameter values were set to 100 and 0.1, with data types of integer and float, respectively. "N_estimators" can range between 1 and positive infinity; however, "learning_rate" may vary between 0.0 and positive infinity. As XGB operates on multiple trees, depicted in Figure 2(c), the depth of these trees, denoted as levels such as "l1" and "l2," is solely determined by the model's complexity required to attain maximum accuracy. The predicted outcomes from each tree are then aggregated with the associated detection error Y (see Figure 2(c), where Y is a small error generated by each tree and X is a variable to identify the error produced by each tree while predicting). Collectively, these results

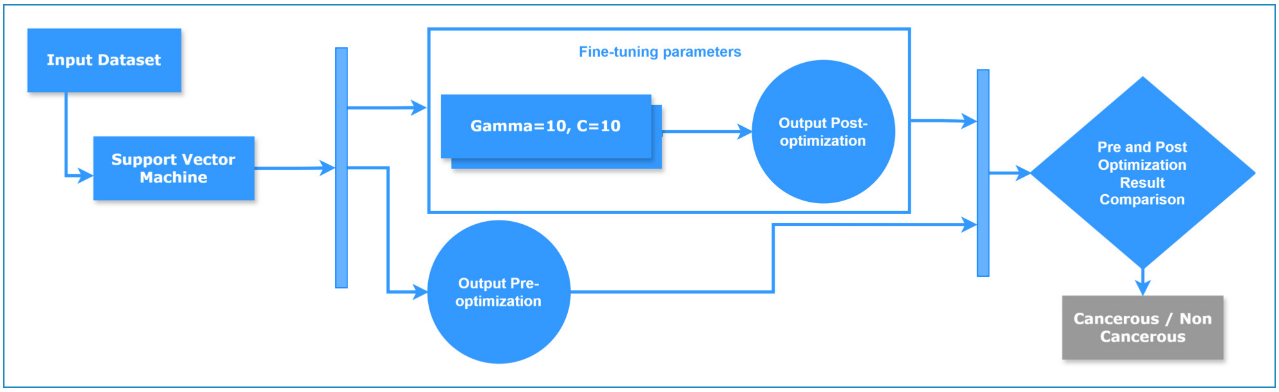

Figure 3. The proposed SVM method.

are utilized to classify whether the provided data pertains to a cancerous or non-cancerous nature. Consequently, the aforementioned parameters were adjusted to seek the optimal solution for detection. Tuning these parameters involves striking the right balance between model complexity (controlled by the n_estimator) and regularization (controlled by the learning rate). The parameters were selected from a pre-defined range of values, and the mathematics running behind the hyperparameter tuning and optimization are as follows:

$$\mathrm{L}(\theta) = \sum_{i=1}^{n} \ell(yi, \hat{y}i) + \sum_{k=1}^{K} \Omega(fk) \tag{3}$$

where θ represents the model parameter, n is the number of training samples, ℓ is the loss function, fk is the individual tree and Ω is the regularization function. For optimization, grid search is used to find the best combination in the grid.

$$Best\ Hyperparameter = \arg\min_{\theta} \{ \frac{1}{k} \sum_{i=1}^{k} Matric\ (\theta\ |data_{train_i},\ data_{val_i})\} \tag{4}$$

where $argmin_\theta$ denotes the combination of hyperparameters that minimizes the objective function. The optimization is performed over all combinations of hyperparameter values in the specific grid. Existing parameters may vary depending on the required methodology; however, "n_estimators" and "learning_rate" are the most common among them. The difference lies in the selection of parameter values in a pre-defined grid. By fine-tuning the "n_estimators" and "learning_rate" parameters, our XGB model can achieve better convergence and potentially avoid overfitting or underfitting issues that may occur with default settings compared to existing techniques.

The proposed dt method. The proposed DT method utilized "random_state," "min_weight_fraction_leaf" and "class_weight" parameters for controlling reproducibility, the minimum fraction of input samples required at a leaf node, and the weight of classes. In addition to this, the default parameter values are "None," 0.0, and 1 with the data types of integer, float and float, respectively. It operates based on the number of input parameters needed to make decisions under given conditions. Beginning with the decision node (refer to Figure 2(e)), which lacks a definitive output classifiable as positive or negative, the process continues to delve deeper into subsequent decision nodes until a decision is reached at the leaf node. Parameters were tuned to ensure consistent and reproducible results; the splits are determined by impurity or entropy, along with class weights for dataset balancing.

"Min_weight_fraction_leaf" can be any float number; however, "class_weight" can be any fractional number, for example, [{0: 1, 1: 1}, {0: 1, 1: 5}, {0: 1, 1: 1}, {0: 1, 1: 1}]. The parameters were selected from a pre-defined range of values, and the mathematics running behind the hyperparameter tuning and optimization are as follows:

The objective is to find the optimal parameters θ that minimize the impurity or mean square error.

$$\text{Minimize} \ = \ \mathcal{L}(\theta) \tag{5}$$

where $\mathscr{L}$ is the Gini impurity, and for regression, it could be the mean square error. For the best hyperparameter:

$$argmin_\theta\ (\mathcal{L}(\theta)) \tag{6}$$

For minimizing the cross-validated loss, the following equation is used:

$$CV(\theta) = \frac{1}{k} \sum_{i=1}^{k} (\theta\ |data_{train_i},\ data_{val_i}) \tag{7}$$

Existing parameters may vary depending on the required methodology; however, "random_state," "min_weight_fraction_leaf" and "class_weight" are the most common among them. The difference lies in the selection of parameter values in a pre-defined grid. Our fine-tuned DT parameters enhance the model's stability ("random_state") and control its growth ("min_weight_fraction_leaf"), leading to improved decision boundaries and potentially better accuracy compared to using default settings with existing techniques.

The proposed LR method. The proposed LR method utilized “C” and “max_iter” parameters for controlling regularization strength and the maximum number of iterations for the solver to converge. In addition to this, the default parameter values are 1 and 50 with the data types of float and integer, respectively. “C” must be any positive float number, and “max_iter” can be any positive number. By adjusting the “C” and “max_iter” parameters, our LR model can better handle the trade-off between regularization and fitting the data, potentially enhancing its predictive performance compared to using default parameter values and existing techniques. Figure 2(d) illustrates the input and assigns probabilities to each input. Subsequently, all inputs undergo a sigmoid function, yielding values ranging from 0.1 to 0.9 as probabilities. The process then assesses whether the probability surpasses the threshold value of 0.5; if so, it is classified into one class, whereas if it falls below 0.5, it is assigned to another class.

Evaluation

In this study, five metrics were used to evaluate our proposed methods: accuracy, precision, sensitivity, F1-score and area under the curve (AUC). Each metric was computed using the elements in a confusion matrix: TP, TN, FP and FN.

Accuracy measures the performance of the classification activity and the number of correctly predicted cases. It is calculated as follows:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{8}$$

Precision, also known as positive predictive value, measures the quality of the result that an algorithm returns. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Sensitivity, also known as recall or true positive rate, measures the accuracy of a test that reports the presence or absence of a condition. It is calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

The F1-score is the harmonic mean of precision and recall, used to measure a test’s accuracy. It is calculated as follows:

$$F1 - score = \frac{(2 \times TP)}{(2 \times TP) + FP + FN} \tag{11}$$

AUC is used to measure separability, clarifying the model’s position in terms of performance measures and aiding in the classification of different classes. This curve compares the true positive and true negative rates, ranging between 0 and 1. A value closer to 1 indicates more accurate performance. A value of 0.5 shows no discrimination, while values between 0.7 and 0.8 are considered acceptable, those between 0.8 and 0.9 as good and those beyond 0.9 as outstanding for distinguishing between two class distributions.[33]

Experiments

To evaluate the effectiveness of the proposed methods, a simulation was conducted to predict lung cancer cases using 309 entries from the Kaggle dataset. The evaluation of the proposed models was based on four performance metrics: accuracy, precision, sensitivity and F1-score (as discussed in the Evaluation section). The simulation process also compared the performance before and after fine-tuning the hyperparameters for each method, utilizing 15 dependent attributes and 1 independent attribute (see Table 2) for training. The parameters were set for each proposed method as follows:

- The proposed SVM method: A range of values was provided to the “ParameterGrid” method, with the optimal values for Gamma and C found to be 1.
- The proposed XGB method: An array of different values was passed as arguments to the “ParameterGrid” method, and the most feasible parameter values for “n_estimator” and “learning_rate” were determined to be 250 and 0.001, respectively.
- The proposed DT method: An array of different values was passed as arguments to the “ParameterGrid” method, and the most feasible parameter values for “random_state,” “min_weight_fraction_leaf” and “class_weight” were determined to be 90, 0.002 and “None,” respectively.
- The proposed LR method: An array of different values was passed as arguments to the “ParameterGrid” method, and the most feasible parameter values for “C” and “max_iter” were determined to be 100 and 50, respectively.

Results and discussion

Results were gathered in two phases: first, after training with default parameters, and second, after hyperparameter tuning. Both sets of results were compared, and the best outcomes were considered to conclude this research at this stage. Figure 4 displays the results for both pre- and post-hyperparameter optimization, showcasing the accuracy, precision, sensitivity and F1-score for the proposed methods.

The accuracy, precision, sensitivity and F1-score for each of the proposed models improved after fine-tuning the hyperparameters, with the exception of the precision of the proposed DT method, which decreased by 4%, and the sensitivity and F1-score of the proposed XGB method, which remained unchanged. Additionally, the F1-score of the proposed DT and LR decreased by 1%. Among these models, the proposed SVM method demonstrated the most significant improvement after hyperparameter fine-tuning, while the proposed XGB, DT and LR methods showed lesser improvement. Specifically, the accuracy, precision, sensitivity and F1-score results of the proposed SVM method improved by 13.45%, 13.0%, 10.0% and 7.0%, respectively. These results indicate that the proposed SVM method exhibits greater capability in detecting lung cancer cases after fine-tuning the hyperparameters.

Table 3 presents the results of accuracy, precision, sensitivity, and area under the curve (AUC) for the four proposed methods for detecting lung cancer. The proposed SVM method achieved the highest accuracy of 99.16%, precision of 98%, sensitivity of 100%, F1-score of 99% and AUC of 0.992. For XGB, DT and LR, the accuracy results are 94.16%, 94.12% and 91.06%, respectively. The precision results for XGB, DT and LR are 91%, 88% and 88%, respectively. The sensitivity results for XGB, DT and LR are 100%, 100% and 97%, respectively. The F1-score results for XGB, DT and LR are 95%, 94% and 92%, respectively. Additionally, the AUC results for XGB, DT and LR are 0.949, 0.941 and 0.916, respectively. These findings suggest that the proposed SVM method outperforms the other methods in detecting lung cancer.

Table 4 displays the confusion matrix results for the proposed methods. The proposed SVM method exhibits the highest numbers of TP and TN, with the lowest counts of FN and FP compared to the other three methods. Conversely, the proposed LR method has the lowest TP and TN numbers, with the highest FN and FP counts among the proposed methods. This indicates that SVM

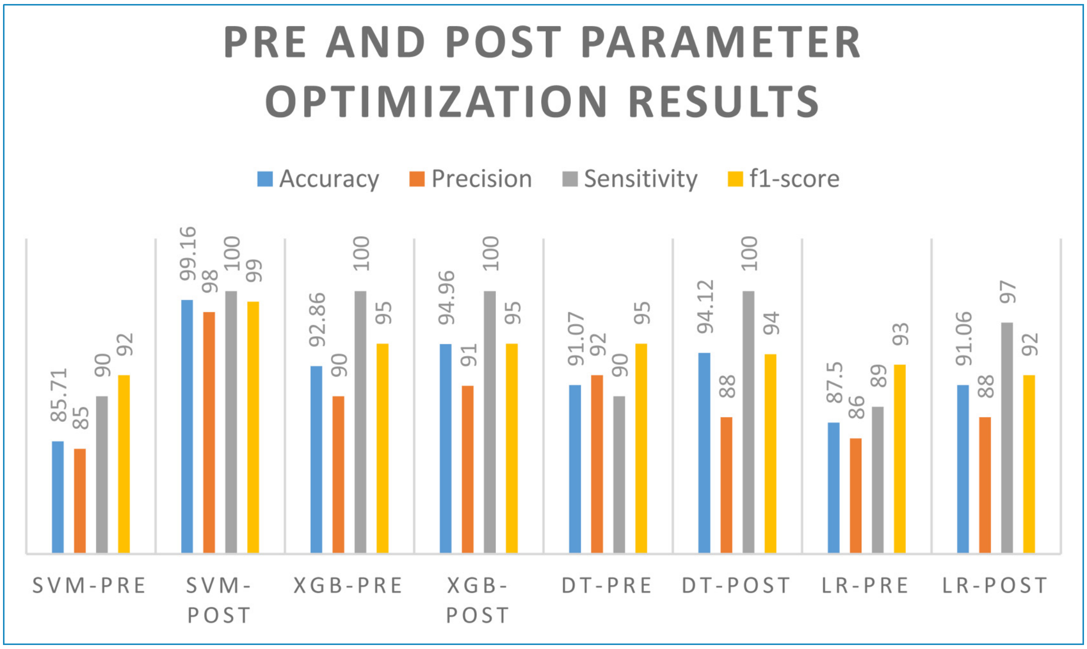

Figure 4. Pre- and post-hyperparameter optimization results comparison.

Table 3. Results for proposed methods for detecting lung cancer.

Proposed Method	Accuracy	Precisions	Sensitivity	F1-Score	Area Under the Curve
SVM	99.16	98.00	100.00	99.00	99.20
XGB	94.96	91.00	100.00	95.00	94.90
DT	94.12	88.00	100.00	94.00	94.10
LR	91.06	88.00	97.00	92.00	91.60

demonstrates a higher number of correct predictions and a lower number of erroneous predictions compared to the other methods.

The results of the AUC for the proposed methods are depicted in Figure 5. Generally, all the proposed methods exhibit AUC values exceeding 0.9. Specifically, our proposed SVM method achieved the highest AUC value (0.992), followed by our proposed XGB method (0.949), our proposed DT method (0.941) and our proposed LR method (0.916), respectively. These findings indicate that the higher the AUC value, the better the performance of our proposed method in distinguishing between the positive and negative classes.

Table 4. Confusion matrix results for the proposed methods.

Proposed Method			Actual value	
			Positive (1)	Negative (0)
SVM	Predicted value	Positive (1)	58	1
		Negative (0)	0	60
XGB		Positive (1)	53	6
		Negative (0)	0	60
DT		Positive (1)	52	7
		Negative (0)	0	60
LR		Positive (1)	51	8
		Negative (0)	2	58

Comparison with previous work

Mamun et al.,[10] Dritsas et al.[32] and Vieira et al.[33] were selected as the related works for comparison with our proposed methods. We chose them to ensure a fair comparison, as they employ similar ML models and the same dataset as ours. Table 5 presents the comparison results among our proposed methods and the selected related works. From the table, it is evident that the accuracy of our proposed SVM, XGB and DT methods has improved by 3.76%, 0.54% and 0.82%, respectively, compared to the other related works. The sensitivity/recall results of our proposed SVM, XGB and DT methods all achieved 100%, indicating improvements of 4.6%, 5.54% and 5.0%, respectively. Regarding precision, our proposed SVM method achieved 98%, outperforming the other selected related works. As for our proposed LR method, it achieved accuracy, sensitivity and precision results of 91.6%, 97% and 88%, respectively. Based on these results, we can conclude that our proposed SVM method demonstrates superior performance in detecting lung cancer compared to the other selected related works.

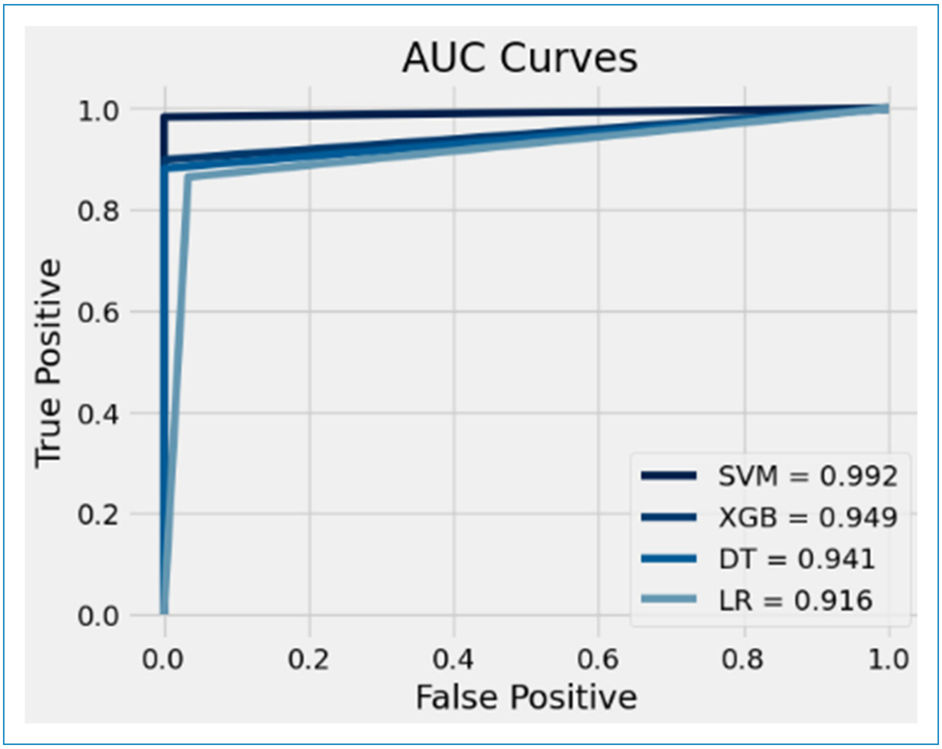

Figure 5. AUC results for the proposed methods.

Table 5. Comparison table of the proposed methods with the selected related works.

Parameters	Research	SVM	XGB	DT	LR
Accuracy in %	Proposed method	99.16	94.96	94.12	91.6
	2022[32]	95.40	–	93.73	–
	2021[33]	88.50	–	86.00	–
	2022[10]	–	94.42	–	–
	2023[23]	60.18	99.07	98.14	85.18
	2023[24]	90.32	96.77	93.54	94.62
Sensitivity/ recall in %	Proposed method	100.00	100.00	100.00	97.00
	2022[32]	95.40	–	93.70	–
	2021[33]	95.80	–	93.70	–
	2022[10]	–	94.46	–	–
	2023[23]	48.00	98.00	96.00	94.00
	2023[24]	91.22	98.26	96.47	97.14
Precision in %	Proposed method	98.00	91.00	90.00	88.00
	2022[32]	95.40	–	93.70	–
	2021[33]	91.30	–	90.00	–
	2022[10]	–	95.60	–	–
	2023[23]	63.00	100.00	100.00	100.00
	2023[24]	91.22	97.70	97.61	95.50

Conclusion and future work

Lung cancer, recognized by the World Health Organization (WHO) as the leading cause of cancer-related deaths, remains a significant global health challenge.[41–50] Early detection can critically enhance survival rates and improve patient outcomes. To address this, our study investigated the potential of ML for the early diagnosis of lung cancer.

Drawing from extensive research, we crafted and evaluated four proposed methods using a dedicated lung cancer dataset. Given the dataset's unbalanced nature, we applied random oversampling to ensure appropriate weighting.

Among our approaches, one stood out, showcasing an impressive accuracy of 99.16%, precision of 98%, recall of 100%, an F1-Score of 99% and an AUC of 99.2%. One of the main reasons our proposed method outperformed others is that we utilized hyperparameter tuning, focusing on the Gamma and C parameters, which were set at a value of 10. These parameters influence kernel width and regularization strength, respectively.

While these results are promising, one must note the limitations of our dataset's size. Exploring larger datasets in the future might provide more comprehensive and generalizable insights. Our next steps involve either delving into more extensive datasets or redirecting our focus toward harnessing ML for detecting other diseases, all in a bid to make significant contributions to global health.

Acknowledgements: None.

Contributorship: Syed Muhammad Nabeel contributed to conceptualization, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, and project administration. Sibghat Ullah Bazai contributed to conceptualization, investigation, data curation, writing—original draft preparation, writing—review and editing, supervision, project administration. Nada Alasbali contributed to methodology, software, resources, writing—original draft preparation, and funding acquisition. Yifan Liu contributed to data curation and formal analysis. Muhammad Imran Ghafoor contributed to software and formal analysis. Rozi Khan contributed to software, formal analysis, and funding acquisition. Chin Soon Ku contributed to validation, writing—review and editing and funding acquisition. Jing Yang contributed to data curation, writing—review and editing, project administration. Sana Shahab contributed to formal analysis, resources, writing—review and editing and funding acquisition. Lip Yee Por contributed to writing—review and editing and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: Ethical approval was not required for this study, as we analyzed publicly available data that was not collected from human subjects but from publicly available symptom checker apps.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through the large group research project under grant number RGP.2/550/44. This research was also supported by the KW IPPP (Research Maintenance Fee) Individual/Centre/Group (RMF1506-2021), Universiti Malaya, Malaysia, and the UTAR Financial Support for Journal Paper Publication Scheme through Universiti Tunku Abdul Rahman, Malaysia.

Guarantor: None.

Informed consent: A publicly available dataset from Kaggle was utilized. This dataset is detailed and has been recently used, making patient consent unnecessary

ORCID iDs: Chin Soon Ku https://orcid.org/0000-0003-0793-3308
Jing Yang https://orcid.org/0000-0002-8132-8395
Lip Yee Por https://orcid.org/0000-0001-5865-1533

References

1. Berk Ş, Kaya S, Akkol EK, et al. A comprehensive and current review on the role of flavonoids in lung cancer–experimental and theoretical approaches. *Phytomedicine* 2022; 98: 153938.
2. What is Metastasis?: Image Details - NCI Visuals Online n.d. https://visualsonline.cancer.gov/details.cfm?imageid=12501 (accessed December 11, 2022).
3. Benusiglio PR, Fallet V, Sanchis-Borja M, et al. Lung cancer is also a hereditary disease. *Eur Respir Rev* 2021; 30: 210045.
4. Schabath MB and Cote ML. Cancer progress and priorities: lung cancer. *Cancer Epidemiol Biomarkers Prev* 2019; 28: 1563–1579.
5. Barnes C, Bray F, Drope J, et al. *Global Cancer Facts & Figures*. 4th ed. Atlanta, GA: American Cancer Society, 2018, pp. 1–73. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-fact-sand- figures/global-cancer-facts-and-figures-4th-edition.pdf (accessed December 11, 2022).
6. Lung Cancer Awareness Month 2022 – IARC n.d. https://www.iarc.who.int/news-events/lung-cancer-awareness-month-2022 (accessed December 11, 2022).
7. Lam S and Tammemagi M. Contemporary issues in the implementation of lung cancer screening. *Eur Respir Rev* 2021; 30: 200288.
8. Nooreldeen R and Bach H. Current and future development in lung cancer diagnosis. *Int J Mol Sci* 2021; 22: 8661.
9. Raoof SS, Jabbar MA and Fathima SA. Lung cancer prediction using machine learning: a comprehensive approach. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), IEEE, 2020, pp. 108–115. https://doi.org/10.1109/ICIMIA48430.2020.9074947.
10. Mamun M, Farjana A, Al Mamun M, et al. Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. 2022 IEEE World AI IoT Congress (AIIoT), IEEE, 2022, pp. 187–193. https://doi.org/10.1109/AIIoT54504.2022.9817326.
11. Abdu-Aljabar RD and Awad OA. A comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier. *IOP Conf Ser Mater Sci Eng* 2021; 1076: 012048.
12. Anil Kumar C, Harish S, Ravi P, et al. Lung cancer prediction from text datasets using machine learning. *Biomed Res Int* 2022; 2022: 1–10.
13. Aamir M, Bazai SU, Bhatti UA, et al. Applications of Machine Learning in Medicine: Current Trends and Prospects. In 2023 Global Conference on Wireless and Optical Technologies (GCWOT), 2023, pp. 1–4. IEEE.
14. Bankar A, Padamwar K and Jahagirdar A. Symptom analysis using a machine learning approach for early stage lung cancer. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), IEEE, 2020, pp. 246–250. https://doi.org/10.1109/ICISS49785.2020.9315904.
15. Feng S, Liu Q, Patel A, et al. Automated pneumothorax triaging in chest X-rays in the New Zealand population using deep-learning algorithms. *J Med Imaging Radiat Oncol* 2022; 66: 1035–1043.
16. Binson VA, Subramoniam M, Sunny Y, et al. Prediction of pulmonary diseases with electronic nose using SVM and XGBoost. *IEEE Sens J* 2021; 21: 20886–20895.
17. Ausawalaithong W, Thirach A, Marukatat S, et al. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. 2018 11th Biomedical Engineering International Conference (BMEiCON), IEEE, 2018, pp. 1–5. https://doi.org/10.1109/BMEiCON.2018.8609997.
18. Lynch CM, Abdollahi B, Fuqua JD, et al. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *Int J Med Inform* 2017; 108: 1–8.
19. Faisal MI, Bashir S, Khan ZS, et al. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. 2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST), IEEE, 2018, pp. 1–4. https://doi.org/10.1109/ICEEST.2018.8643311.
20. El Guabassi I, Bousalem Z, Marah R, et al. Towards an artificial intelligence framework for early diagnosis and prediction of lung cancer. 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE, 2022, pp. 1–6. https://doi.org/10.1109/IRASET52964.2022.9738317.
21. Mustafa Abdullah D, Mohsin Abdulazeez A and Bibo Sallow A. Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Qubahan Academic J* 2021; 1: 141–149.
22. Pradhan KS, Chawla P and Tiwari R. HRDEL: high ranking deep ensemble learning-based lung cancer diagnosis model. *Expert Syst Appl* 2023; 213: 118956.
23. Yamini B, Sudha K, Nalini M, et al. Predictive modelling for lung cancer detection using machine learning techniques. 2023 8th International Conference on Communication and Electronics Systems (ICCES), IEEE, 2023, pp. 1220–1226. https://doi.org/10.1109/ICCES57224.2023.10192648.
24. Singh D, Khandelwal A, Bhandari P, et al. Predicting lung cancer using XGBoost and other ensemble learning models. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023, pp. 1–6. https://doi.org/10.1109/ICCCNT56998.2023.10308301.
25. Chaturvedi P, Jhamb A, Vanani M, et al. Prediction and classification of lung cancer using machine learning techniques. *IOP Conf Ser Mater Sci Eng* 2021; 1099: 012059.
26. Nishio M, Nishizawa M, Sugiyama O, et al. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS One* 2018; 13: e0195875.
27. Makaju S, Prasad PWC, Alsadoon A, et al. Lung cancer detection using CT scan images. *Procedia Comput Sci* 2018; 125: 107–114.
28. Bharathy S, Pavithra R and Akshaya B. Lung cancer detection using machine learning. 2022 International Conference

on Applied Artificial Intelligence and Computing (ICAAIC), IEEE; 2022, pp. 539–543. https://doi.org/10.1109/ICAAIC53929.2022.9793061.

29. Thirunavukkarasu MK and Karuppasamy R. Forecasting determinants of recurrence in lung cancer patients exploiting various machine learning models. *J Biopharm Stat* 2023; 33: 257–271.
30. Puneet CA. Detection of lung cancer using machine learning techniques based on routine blood indices. 2020 IEEE International Conference for Innovation in Technology (INOCON), IEEE, 2020, pp. 1–6. https://doi.org/10.1109/INOCON50539.2020.9298407.
31. Patra R. Prediction of lung cancer using machine learning classifier. *Commun Comput Inf Sci* 2020; 1235 CCIS: 132–142.
32. Dritsas E and Trigka M. Lung cancer risk prediction with machine learning models. *Big Data Cogn Comput* 2022; 6: 39.
33. Vieira E, Ferreira D, Neto C, et al. Data mining approach to classify cases of lung cancer. *Adv Intell Syst Comput* 2021; 1365 AIST: 511–521.
34. Alsinglawi B, Alshari O, Alorjani M, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci Rep* 2022; 12: 07.
35. Venkatesh C and Bojja P. A dynamic optimization and deep learning technique for detection of lung cancer in CT images and data access through internet of things. *Wirel Pers Commun* 2022; 125: 2621–2646.
36. Prabhpreet Kaur V. Lung cancer detection using chi-square feature selection and support vector machine algorithm. *Int J Adv Trends Comput Sci Eng* 2021; 10: 2050–2060.
37. Sim J, Kim YA, Kim JH, et al. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. *Sci Rep* 2020; 10: 10693.
38. Wu J, Zan X, Gao L, et al. A machine learning method for identifying lung cancer based on routine blood indices: qualitative feasibility study. *JMIR Med Inform* 2019; 7: e13476.
39. Lung Cancer Detection/Kaggle n.d. https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection (accessed December 11, 2022).
40. Stapelfeld C, Dammann C and Maser E. Sex-specificity in lung cancer risk. *Int J Cancer* 2020; 146: 2376–2382.
41. Brady GC, Roe JWG, O'Brien M, et al. An investigation of the prevalence of swallowing difficulties and impact on quality of life in patients with advanced lung cancer. *Support Care Cancer* 2018; 26: 515–519.
42. Al-Bander B, Fadil YA and Mahdi H. Multi-Criteria decision support system for lung cancer prediction. *IOP Conf Ser Mater Sci Eng* 2021; 1076: 012036.
43. Malinowska K. The relationship between chest pain and level of perioperative anxiety in patients with lung cancer. *Pol J Surg* 2018; 90: 23–27.
44. Detterbeck FC, Lewis SZ, Diekemper R, et al. Executive summary. *Chest* 2013; 143: 7S–37S.
45. Kantor ED, Hsu M, Du M, et al. Allergies and asthma in relation to cancer risk. *Cancer Epidemiol Biomarkers Prev* 2019; 28: 1395–1403.
46. O'Keeffe LM, Taylor G, Huxley RR, et al. Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis. *BMJ Open* 2018; 8: e021611.
47. Leshargie CT, Alebel A, Kibret GD, et al. The impact of peer pressure on cigarette smoking among high school and university students in Ethiopia: a systemic review and meta-analysis. *PLoS One* 2019; 14: e0222572.
48. Phillips M, Bauer TL and Pass HI. A volatile biomarker in breath predicts lung cancer and pulmonary nodules. *J Breath Res* 2019; 13: 036013.
49. Avancini A, Sartori G, Gkountakos A, et al. Physical activity and exercise in lung cancer care: will promises be fulfilled? *Oncologist* 2020; 25: e555–e569.
50. Bagnall A and Cawley GC. On the use of default parameter settings in the empirical evaluation of classification algorithms 2017; arxiv.1703.06777. https://doi.org/10.48550/arxiv.1703.06777.