Turkish Journal of Engineering
https://dergipark.org.tr/en/pub/tuje
e-ISSN 2587-1366

Machine learning-based lung cancer diagnosis

Mahmut Dirik *[1]

[1]Sirnak University, Computer Engineering Department, Türkiye

Keywords
Lung cancer
Early diagnosis
Machine learning
Classification

Research Article
DOI: 10.31127/tuje.1180931

Received: 27.09.2022
Revised: 22.12.2022
Accepted: 27.12.2022
Published:27.02.2023

Abstract

Cancer is one of the leading health problems, occurring in various organs and tissues of the body, and its incidence is increasing worldwide. Lung cancer is one of the deadliest types of cancer. Due to its worldwide prevalence, increasing number of cases, and deadly consequences, early detection of lung cancer, as with all other cancers, greatly increases the chances of survival. As with all other diseases, the diagnosis of cancer is only possible after the appearance of various symptoms and an examination by specialists. Known symptoms of lung cancer are shortness of breath, coughing, wheezing, jaundice in the fingers, chest pain, and difficulty swallowing. The diagnosis is made by an expert on site based on these symptoms and additional tests. The aim of this study is to detect the disease at an earlier stage based on the symptoms present, to assess more cases with less time and cost, and to achieve results in new situations that are as successful or even faster than those of human experts by deriving them from existing data using different algorithms. The aim is to develop an automated model that can detect early-stage lung cancer based on machine learning methods. The developed model includes nine different machine learning algorithms (NB, LR, DT, RF, GB, and SVM). The success of the classification algorithms used was evaluated using the metrics of accuracy, sensitivity, and precision calculated using the parameters of the confusion matrix. The results obtained show that the proposed model can detect cancer with a maximum accuracy of 91%.

1. Introduction

Lung cancer is the formation of a mass (tumor) in the lung by cells that are structurally normal lung tissue but that multiply uncontrollably. The mass formed here first grows around itself and, in later stages, spreads via the blood to surrounding tissue or distant sites (liver, bones, brain, etc.) and causes damage. This spread is called metastasis. Lung cancer is responsible for the largest proportion of deaths from malignant diseases worldwide [1-6]. The International Agency for Research on Cancer (IARC) provides estimates of incidence and mortality rates for 36 specific cancers in 185 countries and for all cancers combined for the year 2020, according to its latest estimates of the global cancer burden as of December 15, 2020 [6-7]. Worldwide, the total number of cancer patients still alive within 5 years of cancer diagnosis (5-year prevalence) is estimated at 50.6 million. The most common cancer worldwide is breast cancer in women (11.7%), followed by lung cancer (11.4%), colorectal cancer (10.0%), prostate cancer (7.3%), and stomach cancer (5.6%) [6]. The causes of cancer are diverse and range from behavioral characteristics such as high body mass index, tobacco, and alcohol use to physical carcinogens such as exposure to ultraviolet rays and radiation, including certain biological and genetic carcinogens [8]. Malaise, fatigue, nausea, a persistent cough, difficulty breathing, weight loss, muscle pain, and bleeding and bruising are among the most common cancer symptoms [3, 9, 10]. Again, none of these symptoms are cancer-specific, and not every patient has them all. Without a comprehensive diagnostic examination such as a computed tomography scan (CT) [11], magnetic resonance imaging (MRI) [12], 13], positron emission tomography (PET) [14], ultrasound, or a biopsy, it is impossible to detect the presence of cancer. In the early stages, those affected often show few or no symptoms. As with all other malignancies, timely and early detection of lung cancer is critical due to its prevalence, high mortality rate, and increasing incidence. Clinicians want to know the actual relationship between observations, interventions, and

* **Corresponding Author**
*(scai.journal@gmail.com) ORCID ID 0000-0003-1718-5075

Cite this article
Dirik, M. (2023). Machine learning-based lung cancer diagnosis. Turkish Journal of Engineering, 7(4), 322-330

outcomes (outputs). In other words, they need a model to detect, classify, or predict disease. Currently, this information is based on clinical trials and clinicians' experience. Reviewing medical records to determine the best treatment options for patients is also very time-consuming. A good estimation and classification model simplifies the whole process.

The diagnosis of cancer is made by doctors who are experts in the field by interpreting the observable symptoms and the results of examinations. Intensive studies in the field of artificial intelligence [15–17], together with rapidly developing technology, have paved the way for machines to make the right decisions in many areas, just like human experts. The branch of artificial intelligence that enables predictions about new situations by using correlations obtained from data by interpreting the data of the current situation is called machine learning (ML) [18–20]. It is a set of techniques that enable rapid and accurate decisions to be made for new and different situations by learning from data that has been studied in a particular subject area. These techniques have enabled successful applications in healthcare [18, 19] and many other fields. ML has a wide range of applications, from disease detection in pathology to intelligent systems that can prescribe conventional drugs when evaluated based on the patient's symptoms [21]. When provided with high quality and sufficient data, it can deliver results that are as accurate as those of human experts, even faster and more powerful.

When it comes to methods of diagnosing disease, computer technology and artificial intelligence have shown incredible potential in the diagnostic industry, offering a powerful alternative to traditional diagnostic methods. Diagnosing a particular disease requires taking a sample from a patient, running a series of tests on those samples, converting the results into an interpretable form, and finally having a trained person make a decision based on those results. If the samples taken from a patient are digital or have been digitized in some way, they can be analyzed by machines. They can then be provided with a dataset containing decisions about similar situations in the past. In machine learning, making decisions based on information obtained from past scenarios is called "supervised learning" [22]. Over the last three decades, many supervised learning algorithms have been developed that are ideal for working with biomedical data. By using biomedical data [23], artificial intelligence can offer a new dimension in the field of medical diagnosis [24] and is increasingly becoming a viable alternative to traditional diagnostic methods. Although AI models appear promising on paper and in controlled experiments, they are not yet reliable enough to be trusted with life-changing decisions. Of course, some simple diagnostic procedures are only performed by machines with little or no human intervention. Yet AI methods often still struggle in practice. These challenges are being overcome by collecting more practical data, developing new and improved learning algorithms, and rigorously testing new models.

In this study, we developed a ML-based classification model based on a performance comparison of algorithms that can diagnose lung cancer. The results showed that the model is able to classify associated lung cancer with high reliability. In the following sections, we describe this approach to cancer diagnosis and provide the necessary graphs, tables, charts, and other drawings to facilitate interpretation.

The rest of the article is structured as follows: Chapter 2 explains the principles of the methods and techniques required to build this model. This section explains the dataset, the machine learning algorithm, and the performance evaluation metrics used. Finally, chapter 3 presents the results and discussion.

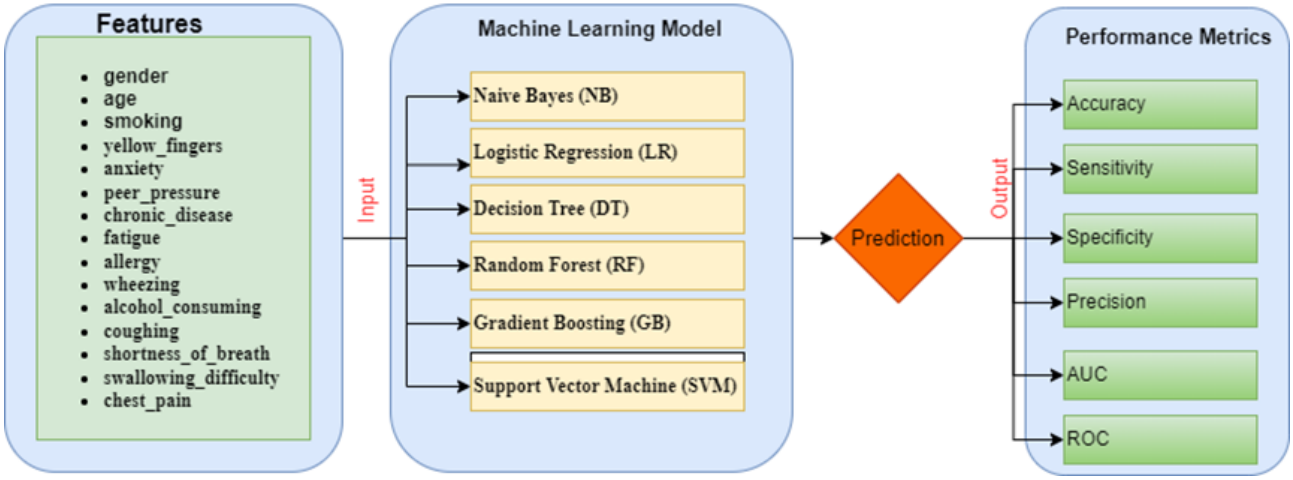

Figure 1. Flowchart of the process for machine learning based lung cancer diagnosis

2. Method

Early detection of lung cancer is thought to increase survival rates and reduce cancer-related deaths worldwide. People at high risk often undergo annual radiological screening by computed tomography, the current method of clinical lung cancer diagnosis. The performance of CT screening is not satisfactory due to its high cost and high prevalence of false-positive results [25]. In this study, the predictive biomarkers for the diagnosis of lung cancer were a persistent or worsening cough, blood or bloody sputum when coughing, chest pain that worsens when coughing or laughing while breathing deeply, loss of appetite, weakness, fatigue,

weight loss, hoarseness, shortness of breath, and recurrent or persistent lung infections such as bronchitis or pneumonia. These data are combined using machine learning techniques to find diagnostic indicators of early-stage lung cancer. Figure 1 shows the overall framework of the proposed architecture.

2.1. Dataset

The effectiveness of the cancer prediction system helps people know their cancer risk at a low cost and helps them make the right decision based on their cancer risk status. The data used in this study [25] consists of responses from 309 different people from a lung cancer survey conducted in 2013. In addition to the basic information about the individuals in the dataset, there are several observable anomalies about their harmful habits and health. The dataset contains information on 15 characteristics. Depending on these features, a label is generated indicating whether lung cancer was diagnosed in the returns of the same individuals. The features in the dataset and the correlation between them are listed below. The 16th feature in the dataset was obtained as feedback from participants. A study was conducted to estimate the 16th feature using the first 15 features related to lung cancer in the dataset. Figure 2 shows a diagram of the general distribution of the data.

The statistical method used to determine whether a linear relationship exists between numerical measurements of data in a data set and, if so, the direction and strength of that relationship, is called correlation. If the correlation coefficient is negative, there is an inverse relationship between the two variables. If the correlation coefficient is positive, there is a correct relationship between the two variables, i.e., "if one variable increases, so does the other." Table 1 shows the correlation of the data used.

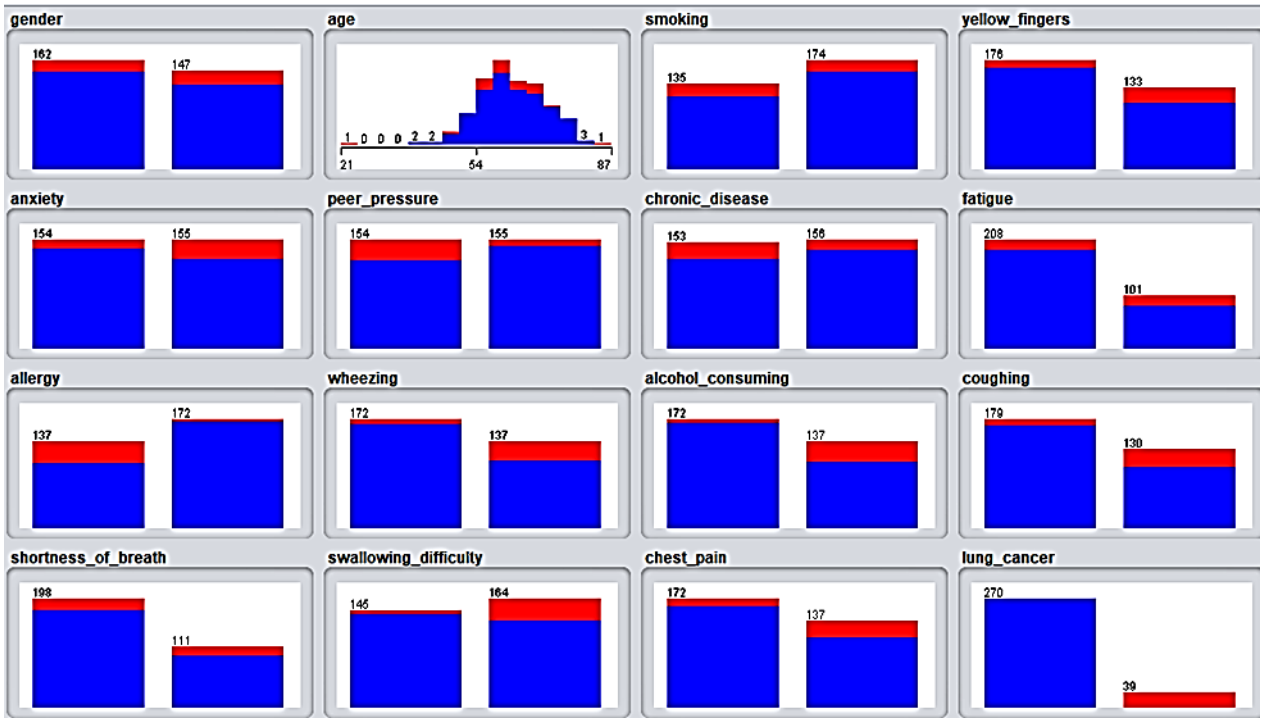

Figure 2. Visualize All attributes

Table 1. Lung cancer data correlation table

Attributes	age	alcohol...	allergy ...	anxiety ...	chest_p...	chronic...	coughin...	fatigue ...	gender ...	lung_ca...	peer_pr...	shortne...	smokin...	swallo...	wheezi...	yellow_...
age	1	0.059	0.028	-0.053	-0.018	-0.013	0.170	0.013	0.021	0.089	0.019	-0.018	-0.084	0.001	0.055	0.005
alcohol_co...	0.059	1	0.344	0.166	0.331	0.002	0.203	-0.191	0.454	0.289	-0.160	-0.179	-0.051	0.009	0.266	-0.289
allergy = yes	0.028	0.344	1	0.166	0.239	0.106	0.190	0.003	0.154	0.328	-0.082	-0.030	0.002	0.062	0.174	-0.144
anxiety = no	-0.053	0.166	0.166	1	0.114	0.010	0.226	0.189	0.152	-0.145	-0.217	0.144	-0.160	0.489	0.192	-0.566
chest_pain ...	-0.018	0.331	0.239	0.114	1	-0.037	0.084	-0.011	0.363	0.190	-0.095	0.024	0.120	-0.069	0.148	-0.105
chronic_dis...	-0.013	0.002	0.106	0.010	-0.037	1	-0.175	-0.111	-0.205	0.111	0.049	-0.026	-0.142	-0.075	-0.050	0.041
coughing = ...	0.170	0.203	0.190	0.226	0.084	-0.175	1	0.147	0.133	0.249	-0.089	0.277	-0.129	0.158	0.374	-0.013
fatigue = yes	0.013	-0.191	0.003	0.189	-0.011	-0.111	0.147	1	-0.084	0.151	0.078	0.442	-0.030	0.133	0.142	-0.118
gender = m...	0.021	0.454	0.154	0.152	0.363	-0.205	0.133	-0.084	1	0.067	-0.276	-0.065	0.036	0.078	0.141	-0.213
lung_cance...	0.089	0.289	0.328	-0.145	0.190	0.111	0.249	0.151	0.067	1	0.186	0.061	0.058	-0.260	0.249	0.181
peer_press...	0.019	-0.160	-0.082	-0.217	-0.095	0.049	-0.089	0.078	-0.276	0.186	1	-0.220	-0.043	-0.367	-0.069	0.323
shortness_...	-0.018	-0.179	-0.030	0.144	0.024	-0.026	0.277	0.442	-0.065	0.061	-0.220	1	0.061	0.161	0.038	-0.106
smoking = ...	-0.084	-0.051	0.002	-0.160	0.120	-0.142	-0.129	-0.030	0.036	0.058	-0.043	0.061	1	-0.031	-0.129	-0.015
swallowing...	0.001	0.009	0.062	0.489	-0.069	-0.075	0.158	0.133	0.078	-0.260	-0.367	0.161	-0.031	1	-0.069	-0.346
wheezing = ...	0.055	0.266	0.174	0.192	0.148	-0.050	0.374	0.142	0.141	0.249	-0.069	0.038	-0.129	-0.069	1	-0.079
yellow_fing...	0.005	-0.289	-0.144	-0.566	-0.105	0.041	-0.013	-0.118	-0.213	0.181	0.323	-0.106	-0.015	-0.346	-0.079	1

2.2. Machine Learning Classifiers

Machine learning (ML) [26–28] is the process of instructing computers to use data more efficiently and effectively through reinforcement learning. It refers to the supervised learning process used in classification, whereby the software learns from incoming data and then applies this knowledge to categorize future observations. Classification techniques are used to determine the classification of the data. Similar to the regression model, the categorization model predicts future outcomes. In this study on ML-based lung cancer detection, we used the following algorithms: Naive Bayes, Logistic Regression, Fast Large Margin, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machin.

2.2.1. Naive Bayes (NB)

Naive Bayes is a simple learning procedure developed by Thomas Bayes [29] that applies Bayes' rule and makes a strong assumption about the conditional independence of features with respect to class. Naive Bayes is widely used in practice because of its computational efficiency and several other advantageous properties. The quantitative component of the Bayesian network consists of three basic components: probability theory, Bayes' theorem, and conditional probability functions. Bayes' theorem starts from the premise that the conditional probability is proportional to the probability of events occurring. This makes it easy to represent the probability distribution in graphical models as conditional dependence or independence [30, 31].

2.2.2. Logistic Regression (LR)

Logistic regression (LR) is a simpler and more accurate method for dealing with binary and linear classification problems, i.e., modelling the probability of a discrete outcome as a function of an input variable. It is a basic classification model that works effectively for classes that can be linearly separated [32].

2.2.3. Decision Tree (DT)

The decision tree method is a supervised learning technique that can handle categorical and numerical data under supervision and corresponds to the ideal node and edge tree for classification problems. Each node in the tree indicates the class of the problem, while each edge reflects the result of the analysis. This classifier is a predictive machine learning model that illustrates the relationship between the values of the dataset and the features. The goal of decision-making is to determine the best option considering the entire probability distribution. Each branch of the decision tree represents the possible value of a different category. The nodes are determined based on entropy measurements of the features in the dataset. The root node is the feature with the highest entropy [33–35].

2.2.4. Random Forest (RF)

Random Forest was developed by Leo Breiman [36] to create a community of estimators by generating a set of decision trees in randomly selected data subdomains, where each tree has the same value and depends on the values of an independently generated random vector. Random forests, also known as random choice forests, are a type of ensemble learning technique used to solve classification, regression, and other problems that require training a large number of decision trees. The output of the random forest is the class of the classification problem that is chosen by the majority of the trees [37, 38].

2.2.5. Gradient Boosting (GB)

The method was developed by combining the concepts of gradient descent and boosting. improves the results of decision trees using the gradient descent algorithm. Splitting the dataset into multiple sub-datasets as in a Random Forest is not done in this approach. A decision tree is created from the existing dataset and a new decision tree is created based on its errors [39–41].

2.2.6. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression problems. However, it is mainly used for classification problems. Draws a line to separate points that lie on a plane [42–44]. The objective is that this line has the maximum distance between the points of the two classes. SVM was developed by Vapnik et al [45].

2.3. Performance Metrics

The purpose of the performance evaluation is to analyze the effectiveness of the algorithms used and show the usability of the system. The output values are compared with the actual values to validate a classification approach. In this study, the confusion matrix [46] was used to measure the classification success of the proposed approach, and six different performance measures were used, namely accuracy, area under the curve (AUC), precision, F-measure, sensitivity, and specificity. The corresponding calculation formulas can be found in Table 2. In the table, TP (True Positive) represents the number of correctly classified positive data, FP (False Positive) represents the number of misclassified positive data, TN (True Negative) represents the number of correctly classified negative data, and FN (False Negative) represents the number of misclassified negative data [46].

The Receiver Operating Characteristic Curve (ROC) is characterized by the AUC. AUC is a graphical representation of the false-positive rate (FPR) and the true-positive rate (TPR) at different confidence levels. Since AUC is not based on a discontinuity number, it is a more reliable measure of overall performance than accuracy [51].

Table 2. Performance metrics

Description	*Formula*	*References*
Accuracy	$ACC = \frac{TP + TN}{TP + FP + TN + FN}$	[47] [48], [49]
***Sensitivity* (*Recall*)**	$RCL = \frac{TP}{TP + FN}$	[48], [49]
Specificity	$SPC = \frac{TN}{FP + TN}$	[48], [49]
Precision	$PRE = \frac{TP}{TP + FP}$	[48], [49]
F-1 Score	$\mathrm{FSC} = 2 * \frac{PRE.RCL}{PRE + RCL}$	[48], [49]
Area Under the Curve	$\mathrm{AUC} = \frac{1}{2}.(RCL + SPC)$	[50]

3. Results and Discussion

Lung cancer is one of the cancers that causes the most deaths among cancers because it is not diagnosed until the late stages. For this reason, it is very important to detect lung cancer at an early stage. As in many other fields, machine learning is now delivering successful results in health research. The use of computer-aided diagnosis systems in the early detection of lung cancer can lead to more accurate and faster results. The classification system implemented in this study is able to provide a practical assessment of the risk status of lung cancer based on the information provided by the individuals themselves. In this study, Naive NB, LR, DT, RF, GB, and SVM were classified by six different machine learning methods to accurately diagnose critical disease-related features of lung cancer.

In the study on evaluating the performance of machine learning classifiers, we present the results of the metrics obtained based on the complexity matrix of the case with the highest average classification success. The values ACC, RCL, SPC, PRE, FSC, AUC, and ROC of the classifiers used are shown in Figures 3–10.

4. Conclusion

A summary graph of the accuracy rates achieved as a result of the classification process can be found in Figure 3. Although all six methods showed high success in classifying the data, it was found that the most successful methods were NB and SVM, with a rate of 91%. The ROC curves of each classifier are shown in Figure 10. The ROC curve is one of the most commonly used metrics to evaluate the performance of machine learning algorithms. It explains how good the model is at prediction. One of the most commonly used metrics is the AUC curve. AUC stands for "Area Under the ROC Curve." The area of this field is AUC. The larger the area covered; the better machine learning models can distinguish between certain classes. The ideal value for AUC is 1. The classification system implemented in this study is able to provide a practical assessment of the risk status of lung cancer based on the information provided by the people themselves. In this way, it guides people to take the necessary precautions in time. Due to the methods used, high diagnostic accuracy is aimed for. In addition, this study will help doctors who will use the system to provide initial information for the diagnosis of lung cancer. Methods with high accuracy are necessary to achieve developments in such matters. This study focuses on effective decision-support systems that can help them diagnose more easily and accurately. In this context, the effectiveness and accuracy of expert systems and various artificial intelligence techniques were evaluated. As a result of this evaluation, NB and SVM showed high predictive performance in the problem of diagnosing lung diseases. In conclusion, NB and SVM methods can be successfully used in the diagnosis of lung diseases and are preferred by physicians to make decisions about the disease.

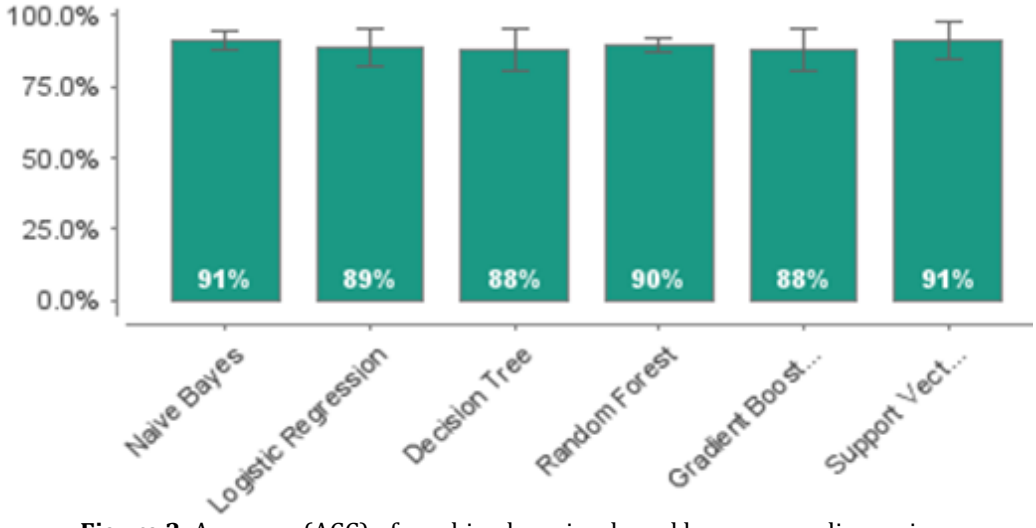

Figure 3. Accuracy (ACC) of machine learning-based lung cancer diagnosis

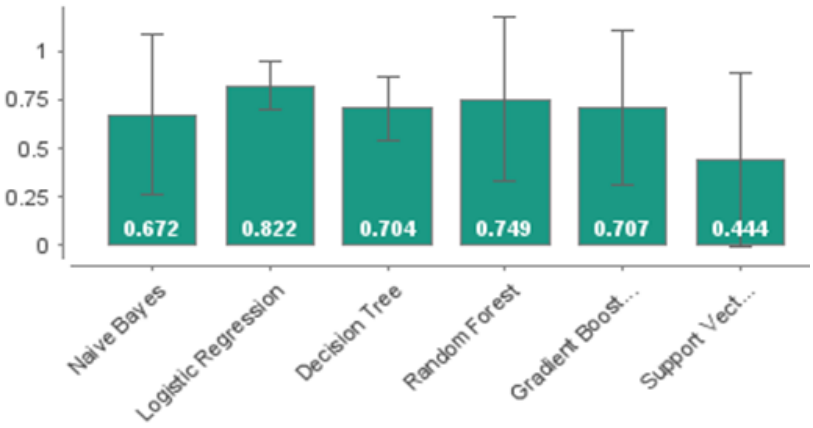

Figure 4. Area Under the Curve (AUC) of machine learning-based lung cancer diagnosis

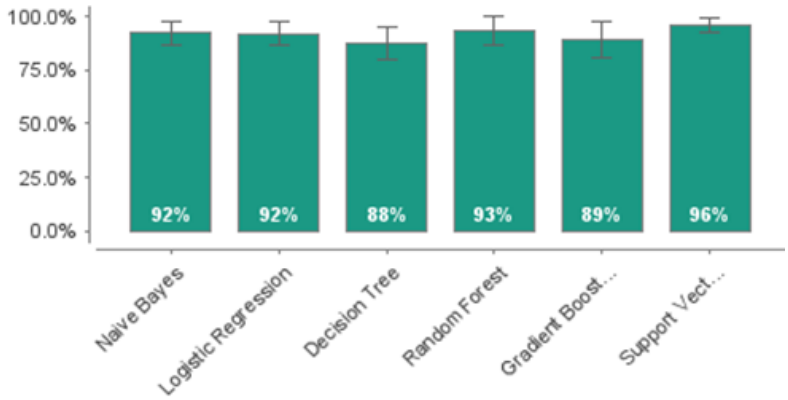

Figure 5. Precision of machine learning-based lung cancer diagnosis

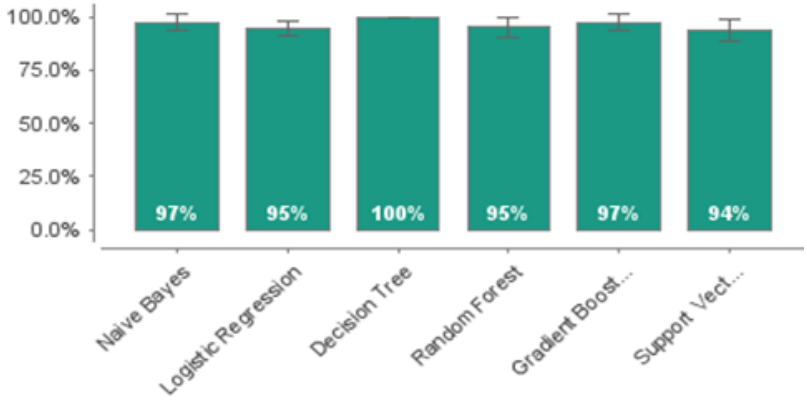

Figure 6. Sensitivity/Recall of machine learning-based lung cancer diagnosis

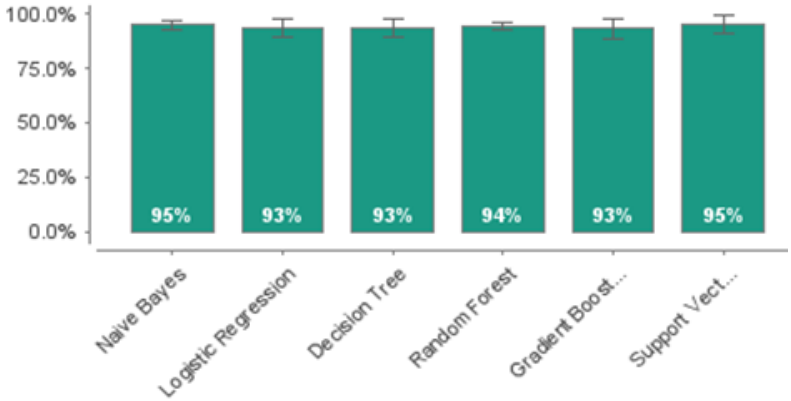

Figure 7. F Measure of machine learning-based lung cancer diagnosis

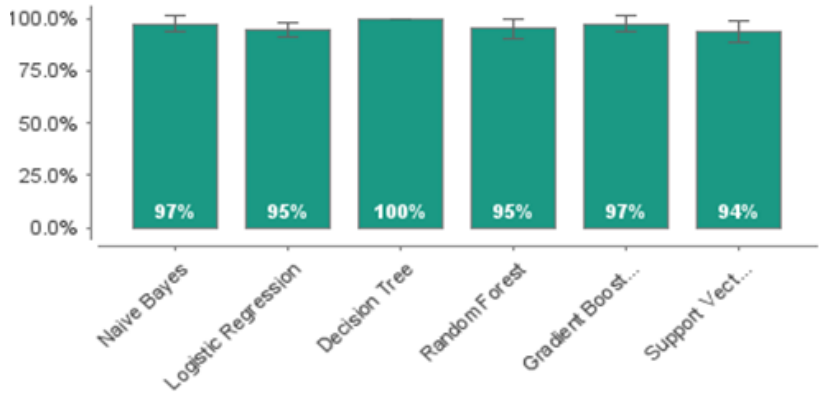

Figure 8. Sensitivity of machine learning-based lung cancer diagnosis

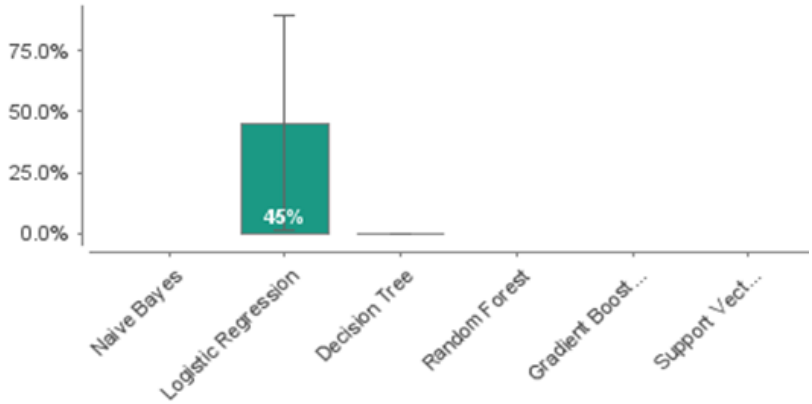

Figure 9. Specificity of machine learning-based lung cancer diagnosis

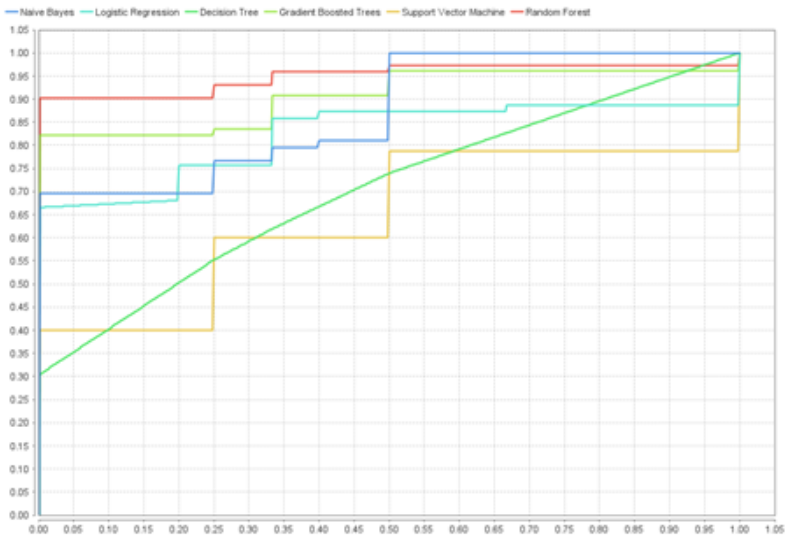

Figure 10. ROC Comparison of machine learning-based lung cancer diagnosis

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., ... & Leung, E. L. H. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, *14*(1), 100907. https://doi.org/10.1016/j.tranon.2020.100907
2. Chiu, H. Y., Chao, H. S., & Chen, Y. M. (2022). Application of artificial intelligence in lung cancer. *Cancers*, *14*(6), 1370. https://doi.org/10.3390/cancers14061370
3. Masud, M., Sikder, N., Nahid, A. A., Bairagi, A. K., & AlZain, M. A. (2021). A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors*, *21*(3), 748. https://doi.org/10.3390/s21030748
4. https://www.mohw.gov.tw/cp-4650-50697-2.html
5. https://www.who.int/news-room/fact-sheets/detail/cancer
6. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global

cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249. https://doi.org/10.3322/CAAC.21660
7. https://gco.iarc.fr/
8. https://www.who.int/news-room/fact-sheets/detail/cancer
9. Rock, C. L., Thomson, C., Gansler, T., Gapstur, S. M., McCullough, M. L., Patel, A. V., ... & Doyle, C. (2020). American Cancer Society guideline for diet and physical activity for cancer prevention. *CA: a cancer journal for clinicians*, *70*(4), 245-271. https://doi.org/10.3322/CAAC.21591
10. Shakeel, P. M., Tolba, A., Al-Makhadmeh, Z., & Jaber, M. M. (2020). Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks. *Neural Computing and Applications*, *32*, 777-790. https://doi.org/10.1007/S00521-018-03972-2/FIGURES/8
11. Bruno, F., Granata, V., Cobianchi Bellisari, F., Sgalambro, F., Tommasino, E., Palumbo, P., ... & Barile, A. (2022). Advanced Magnetic Resonance Imaging (MRI) Techniques: Technical Principles and Applications in Nanomedicine. *Cancers*, *14*(7), 1626. https://doi.org/10.3390/CANCERS14071626
12. Zhang, Y., Wang, R., Hu, J., Qin, X., Chen, A., & Li, X. (2022). Magnetic resonance imaging (MRI) and computed topography (CT) analysis of Schatzker type IV tibial plateau fracture revealed possible mechanisms of injury beyond varus deforming force. *Injury*, *53*(2), 683-690. https://doi.org/10.1016/J.INJURY.2021.09.041
13. Grootjans, W., Rietbergen, D. D., & van Velden, F. H. (2022, May). Added value of respiratory gating in positron emission tomography for the clinical management of lung cancer patients. In *Seminars in Nuclear Medicine*. WB Saunders. https://doi.org/10.1053/J.SEMNUCLMED.2022.04.006
14. Kooli, C., & Al Muftah, H. (2022). Artificial intelligence in healthcare: a comprehensive review of its ethical concerns. *Technological Sustainability*, 1(2), 121-131. https://doi.org/10.1108/TECHS-12-2021-0029
15. Sun, L., Gupta, R. K., & Sharma, A. (2022). Review and potential for artificial intelligence in healthcare. *International Journal of System Assurance Engineering and Management*, *13*(Suppl 1), 54-62. https://doi.org/10.1007/S13198-021-01221-9/FIGURES/6
16. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., & Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, *9*(8), 220638. https://doi.org/10.1098/RSOS.220638
17. Rastogi, M., Vijarania, D., & Goel, D. (2022). Role of Machine Learning in Healthcare Sector. *Neha, Role of Machine Learning in Healthcare Sector (August 20, 2022)*. https://doi.org/10.2139/SSRN.4195384
18. Lawson, C. E., Martí, J. M., Radivojevic, T., Jonnalagadda, S. V. R., Gentz, R., Hillson, N. J., ... & Martin, H. G. (2021). Machine learning for metabolic engineering: A review. *Metabolic Engineering*, *63*, 34-60. https://doi.org/10.1016/J.YMBEN.2020.10.005
19. Das, S., Biswas, S., Paul, A., & Dey, A. (2018). AI Doctor: An intelligent approach for medical diagnosis. In *Industry Interactive Innovations in Science, Engineering and Technology: Proceedings of the International Conference, I3SET 2016* (pp. 173-183). Springer Singapore. https://doi.org/10.1007/978-981-10-3953-9_17/COVER
20. Bukhari, S. U. K., Syed, A., Bokhari, S. K. A., Hussain, S. S., Armaghan, S. U., & Shah, S. S. H. (2020). The histological diagnosis of colonic adenocarcinoma by applying partial self supervised learning. *MedRxiv*, 2020-08. https://doi.org/10.1101/2020.08.15.20175760
21. Shakeel, P. M., Tolba, A., Al-Makhadmeh, Z., & Jaber, M. M. (2020). Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks. *Neural Computing and Applications*, *32*, 777-790. https://doi.org/10.1007/S00521-018-03972-2/FIGURES/8
22. Das, S., Biswas, S., Paul, A., & Dey, A. (2018). AI Doctor: An intelligent approach for medical diagnosis. In *Industry Interactive Innovations in Science, Engineering and Technology: Proceedings of the International Conference, I3SET 2016* (pp. 173-183). Springer Singapore. https://doi.org/10.1007/978-981-10-3953-9_17/COVER
23. Zhao, W., Yang, J., Sun, Y., Li, C., Wu, W., Jin, L., ... & Li, M. (2018). 3D deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. *Cancer research*, *78*(24), 6881-6889. https://doi.org/10.1158/0008-5472.CAN-18-0696
24. https://data.world/josh-nbu/lung-cancer/workspace/file?filename=survey+lung+cancer+%281%29.csv
25. Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. *Informatics in Medicine Unlocked*, 100924. https://doi.org/10.1016/J.IMU.2022.100924
26. Mohammadi, F. G., Shenavarmasouleh, F., & Arabnia, H. R. (2022). Applications of machine learning in healthcare and internet of things (IOT): a comprehensive review. *arXiv preprint arXiv:2202.02868*. https://doi.org/10.48550/arxiv.2202.02868
27. Subasi, A. (2020). *Practical machine learning for data analysis using python*. Academic Press. https://doi.org/10.1016/B978-0-12-821379-7.00003-5
28. Bellhouse, D. R. (2004). The Reverend Thomas Bayes, FRS: a biography to celebrate the tercentenary of his birth. https://doi.org/10.1214/088342304000000189
29. Itoo, F., & Singh, S. (2021). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*, 1503-1511. https://doi.org/10.1007/s41870-020-00430-y

30. Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Naive Bayes for regression. *Machine Learning*, *41*, 5-25.
31. LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658
32. Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H., Alqarni, A. A., Alsharif, N., ... & Alzahrani, M. Y. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *Journal of Healthcare Engineering*, *2021*. https://doi.org/10.1155/2021/1004767
33. Aggrawal, R., & Pal, S. (2020). Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease. *SN Computer Science*, *1*(6), 344. https://doi.org/10.1007/S42979-020-00370-1/TABLES/5
34. Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, *68*(4), 2488-2507. https://doi.org/10.1080/03772063.2020.1713916
35. Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792. https://doi.org/10.1890/07-0539.1
36. Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, *13*(1), 1063-1095.
37. Lingwal, S., Bhatia, K. K., & Tomer, M. S. (2021). Image-based wheat grain classification using convolutional neural network. *Multimedia Tools and Applications*, 80,35441–35465. https://doi.org/10.1007/s11042-020-10174-3
38. Biau, G., Cadre, B., & Rouvière, L. (2019). Accelerated gradient boosting. *Machine learning*, *108*, 971-992. https://doi.org/10.1007/S10994-019-05787-1/TABLES/5
39. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2
40. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21. https://doi.org/10.3389/FNBOT.2013.00021/XML/NLM
41. Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press. https://doi.org/10.1017/CBO9780511801389
42. Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis. *SSRN Electronic Journal. https://doi.org/10.2139/SSRN.1424949*
43. Rivas-Perea, P., Cota-Ruiz, J., Chaparro, D. G., Venzor, J. A. P., Carreón, A. Q., & Rosiles, J. G. (2012). Support vector machines for regression: a succinct review of large-scale and linear programming formulations. *International Journal of Intelligence Science*, 03(01), 5–14. https://doi.org/10.4236/ijis.2013.31002
44. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.
45. https://devopedia.org/confusion-matrix
46. Isabella, S. J., Srinivasan, S., & Suseendran, G. (2020). An efficient study of fraud detection system using Ml techniques. *Intelligent Computing and Innovation on Data Science*, 59-67. https://doi.org/10.1007/978-981-15-3284-9_8
47. Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, *8*, 25579-25587. https://doi.org/10.1109/ACCESS.2020.2971354
48. Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-9). IEEE. https://doi.org/10.1109/ICCNI.2017.8123782
49. Dirik, M., & Gül, M. (2021). Dynamic optimal ANFIS parameters tuning with particle swarm optimization. *Avrupa Bilim ve Teknoloji Dergisi*, (28), 1083-1092. https://doi.org/10.31590/ejosat.1012888
50. Lin, T. H., & Jiang, J. R. (2021). Credit card fraud detection with autoencoder and probabilistic random forest. *Mathematics*, *9*(21), 2683. https://doi.org/10.3390/math9212683
51. Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering*, *160*, 182-193. https://doi.org/10.1016/j.petrol.2017.10.028