Lung Cancer Prediction Using Machine Learning Techniques

Saif Al Rumhi1, Raza Hasan[1], Saqib Hussain[1#] and Jitendra Pandey[1#]

[1]Middle East College, Muscat, Oman
[#]Advisor

ABSTRACT

A thorough analysis and assessment of several research projects using machine learning algorithms to create prediction models for the diagnosis of lung cancer. The main aim of this research is to identify innovative insights into artificial intelligence techniques that can enhance cancer prognosis, early detection, and overall health outcomes. The study utilized a dataset comprising 309 individuals with and without lung cancer, selected based on family history, and incorporated multiple attributes such as gender, age, smoking, allergies, among others. The most important features of the lung cancer dataset were selected using logistic regression, and the findings were validated using k-fold 10 cross-validation. The proposed model demonstrated superior performance compared to other machine learning techniques, achieving an accuracy of 92.2% and an AUC of 93.6%. The development of a reliable method for lung cancer early detection is the main goal of this study.

Introduction

Lung cancer is a prominent cause of death for people of both genders and is now one of the most lethal illnesses [1]. According to research and statistics, cancer is the second most prevalent cause of death, yet doctors still have difficulty detecting it. In addition to family history, smoking, bad lifestyle choices, and genetic factors are the main causes of lung cancer [2][3].

The American Cancer Society predicts that diagnoses of cancer reached over 1.8 million in 2018, affecting about 40% Americans in their lives at some point [1] [4].

Unfortunately, lung cancer cannot be avoided, and only 15.5% of patients survive five years following their diagnosis. Despite the fact that lung cancer normally starts in the lungs, it can sometimes show early signs before spreading. Early detection of tumors can facilitate the selection of an appropriate course of treatment [2] [4].

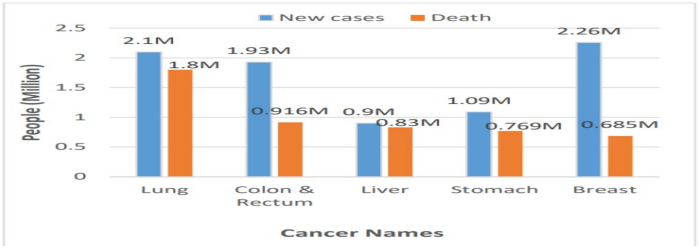

Figure 1. Cancer in 2020 (New cases against death)[5]

Table.1. Lung cancer cases and deaths over the globe[6]

Region	Population	Cancer Cases	Deaths

Asia	60%	66%	57.3%
Europe	9.0%	23.4%	20.3%
America	13.3%	21.0%	14.4%
Africa	17%	9-10%	7.3%

The utilization of information mining practices has demonstrated potential in the prediction of lung cancer mortality risk [7], through the analysis of symptoms and treatment properties. This study employs seven machine learning (ML) techniques to provide a comprehensive analysis of lung cancer prediction, evaluating outcomes across multiple categories, including accuracy, precision, AUC, and a confusion matrix that summarizes the prediction results in matrix form. These robust evaluation metrics are effective for many classification problems and are crucial in selecting the optimal model for lung cancer prediction. Different techniques, including Decision Trees (DT), K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM), are used to categorize the algorithms for this disease.

Related Work

Considerable research has been undertaken to forecast the likelihood of lung cancer. The investigation into determining the probability of developing lung cancer has garnered noteworthy attention and exertion. ML, an AI software innovation, enables self-directed learning to occur within a component based on its own previous work or data extrapolations. As it advances, the software engages in intricate decision-making processes and assimilates knowledge from past actions. Presented below is a brief overview of studies that detail the utilization of various ML techniques for the detection of lung cancer. Majority of the studies used the dataset obtained from Data World website having 1000 dataset with 25 attributes.

The authors of a study published in [8] conducted a comparison of the predictive capabilities of two ML algorithms, specifically SVM and RF, for the purpose of predicting lung cancer. The dataset utilized in the study was obtained from Data World. The findings of the study revealed that the SVM algorithm demonstrated superior performance in comparison to the RF algorithm, achieving an accuracy rate of 98%, precision rate, recall rate, and a F1-score of 100%, all while executing in a mere 0.010 seconds.

In a study conducted by the authors in [9], a hybrid approach of machine and ensemble learning techniques was employed to identify and forecast lung cancer. The evaluation of multiple classifiers was carried out using K-fold 10 cross-validation, and the gradient-boosted tree algorithm was identified as the most effective classifier, with an accuracy rate of approximately 90% with precision, recall and F1-score achieved above 85% on the UCI repository, comprising of 32 instances and 57 attributes.

In a study conducted by the authors in reference [10], four distinct ML algorithms, namely NB, SVM, DT, and LR, were utilized to develop predictive models for lung cancer. The dataset employed in the research was sourced from the Data World website. The findings revealed that the SVM algorithm demonstrated the most superior accuracy rate of 99.2%.

In the study conducted by the authors in [11], the RF algorithm was suggested as a viable method for detecting lung cancer through the utilization of conventional cancer indicators. The efficacy of the model was assessed through the examination of data from 277 patients at Lanzhou University, resulting in an AUC of 99.01%, recall of 96.3%, and accuracy of 95.70%. However, it is important to note that certain assumptions may have been made or potential errors could have arisen during the analysis those are given below:

1. Lung Cancer Research may use biased or unrepresentative data. Patients with lung cancer from one hospital may not represent the total population [6] [9].
2. Small sample size: Lung cancer studies may have a small sample size, especially if they use rare data or are expensive or difficult to execute [7].

3. Incomplete or inconsistent data: Lung cancer studies with various sources or missing values may have incomplete or inconsistent data. These results may be erroneous or incomplete [8].
4. Retrospective and non-long-term lung cancer studies may have limited potential for predicting outcomes. This can hinder therapeutic efficacy assessment [11].

Methodology

In order to achieve optimal outcomes in the process of data mining, it is imperative to effectively manage the attributes of symptoms used for diagnosing a disease. For the purposes of this study, the "Lung Cancer Data set" was utilized, consisting of 309 cases and 16 variables, obtained from Kaggle. The objective of this study was to predict lung cancer where the diagnosis is described by the attributes [12]. The methodology employed in this study incorporated various datasets for lung cancer detection, including Gender, Age, etc. The selected dataset was thoroughly pre-processed to remove duplicate columns and handle missing values, utilizing various ML techniques such as SVM, DT, LR, KNN, etc. These ML methods were used and evaluated using 10-fold cross validation on the chosen dataset. For each model that was chosen, a confusion matrix was used to determine the optimal values based on accuracy, recall, and F1-score. Additionally, a Venn diagram was used to illustrate similarities and contrasts between defined concepts. Correlation analysis was also employed to determine the degree of association between two or more variables. However, limitations were encountered during data collection, including incomplete data and potential inaccuracies, which may have impacted the accuracy of predictions.

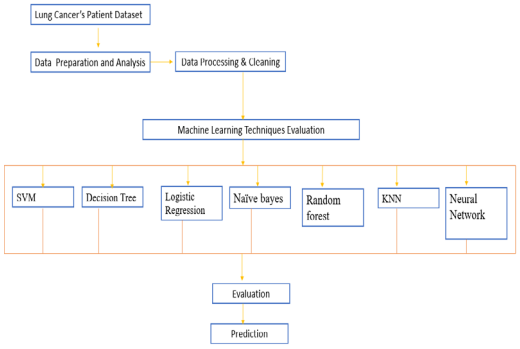

Figure 2. Model analysis

Table 2. Dataset Attribute

No.	Column	Non-Null Count	Dtype
1	Gender	309 non-null	object
2	Age	309 non-null	int64
3	Smoking	309 non-null	int64
4	Yellow Fingers	309 non-null	int64

5	Anxiety	309 non-null	int64
6	Peer Pressure	309 non-null	int64
7	Chronic Disease	309 non-null	int64
8	Fatigue	309 non-null	int64
9	Allergy	309 non-null	int64
10	Wheezing	309 non-null	int64
11	Alcohol Consuming	309 non-null	int64
12	Coughing	309 non-null	int64
13	Shortness of Breath	309 non-null	int64
14	Swallowing Difficulty	309 non-null	int64
15	Chest pain	309 non-null	object
16	Patient ID	309 non-null	int64

Machine Learning Algorithms

Support Vector Machine (SVM)

SVM is a prominent classification technique that is widely used in supervised ML models. SVM operates by sorting data points in the input set into predetermined buckets, and it is capable of classifying transactions into one or more categories to enhance efficiency and accuracy [13]. In the medical field, SVM has been utilized to accurately classify individuals with cardiac illness [7]. The performance of the SVM algorithm is presented in Table 3.

Table 3. Performance evaluation of SVM algorithm

Class	Precision	Recall	F1-Score
No	0.65	0.51	0.57
Yes	0.93	0.96	0.95

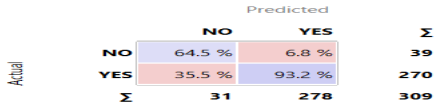

Figure 3. SVM Confusion Matrix

Decision Tree

The DT method differs from other algorithms in that it may be applied to both classification and regression problems. Its primary goal is to build a training model using straightforward decision rules deduced from historical data, which can then be used to predict the value or class of a target variable. DTs are frequently utilized in decision analysis, particularly in operation research, to direct a system towards a goal while also serving as a contingency plan. There are two types of decision trees: those that employ categorical variables and those that employ continuous variables [6][7]. The performance of the Decision Tree algorithm is presented in Table 4.

Table 4. Performance evaluation of Decision Tree algorithm

Class	Precision	Recall	F1-Score
No	0.45	0.54	0.49
Yes	0.93	0.90	0.92

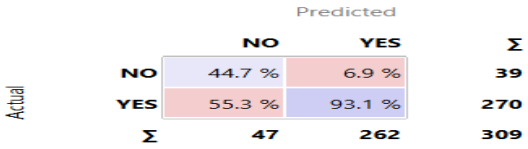

Figure 4. Decision Tree Confusion Matrix

Logistic Regression

The methodology for ascertaining the likelihood of a particular result necessitates the utilization of an input variable. Logistic regression models are frequently employed for this purpose, typically involving a binary outcome that can be articulated as either affirmative or negative, true or false, or comparable dichotomous alternatives [6]. Despite its nomenclature, logistic regression is predominantly utilized as a classification model rather than a regression model. It is a straightforward and efficacious approach for situations where binary and linear classifications are required [14], and it exhibits satisfactory performance with linearly separable categories [15].

Table 5. Performance evaluation of Logistic Regression

Class	Precision	Recall	F1-Score
No	0.73	0.62	0.67
Yes	0.95	0.97	0.96

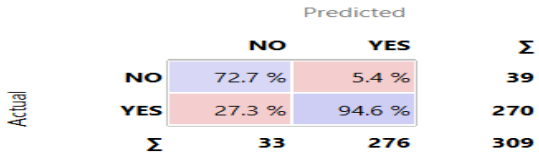

Figure 5. Logistic Regression Confusion Matrix

Naïve Bayes Algorithm

The NB algorithm is a classification technique that utilizes Bayes' theorem and is frequently employed for text categorization when trained on a substantial dataset. It is renowned for its exceptional efficiency and simplicity, enabling it to rapidly generate ML models for prediction purposes. This algorithm is classified as a probabilistic classifier since it makes predictions based on the likelihood of an object. [16].

Table 6. Performance evaluation of Naïve Bayes algorithm

Class	Precision	Recall	F1-Score
No	0.60	0.64	0.62
Yes	0.95	0.94	0.94

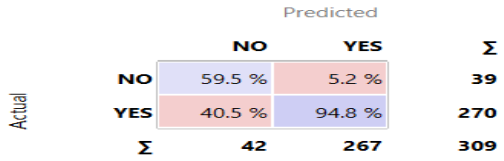

Figure 6. Naïve Bayes Confusion Matrix

Random Forest Algorithm

The RF technique, a form of supervised ML, is commonly employed for classification and regression tasks. It utilizes ensemble learning, which amalgamates multiple classifiers to improve model performance in intricate scenarios [6]. In contrast to alternative algorithms, RF necessitates less training time and can effectively manage large datasets, while still achieving high precision in outcome prediction. Furthermore, it can maintain accuracy even when confronted with incomplete data [14].

Table7. Performance evaluation of Random forest algorithm

Class	Precision	Recall	F1-Score
No	0.68	0.49	0.57
Yes	0.93	0.97	0.95

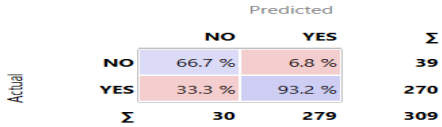

Figure 7. Random forest Confusion Matrix

K Nearest Neighbour (KNN)

The KNN algorithm is a supervised ML approach that relies on the concept of similarity. It is a straightforward and uncomplicated algorithm that presupposes that new data points are comparable to existing ones and allocates them to the category that is most alike. The algorithm retains all accessible data and utilizes similarity to classify new data points into the appropriate category [5]. Although it can be employed for both regression and classification tasks, the

KNN algorithm is primarily utilized for classification [17]. Due to its capacity to rapidly classify newly generated data, the KNN algorithm is a potent tool in the realm of ML [18].

Table 8. Performance evaluation of KNN algorithm

Class	Precision	Recall	F1-Score
No	0.89	0.97	0.93
Yes	0.50	0.21	0.29

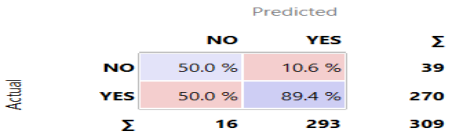

Figure 8. KNN Confusion Matrix

Neural Network

NN enhanced deep learning methodologies known as artificial neural networks or simulated neural networks [2]. ANNs are fashioned to emulate the configuration and communication patterns of the human brain, particularly the interconnections among biological neurons [19]. They comprise of node layers having an input layer along with one or more hidden layers and finally an output layer. The precision of ANNs is enhanced with the utilization of training data over a period of time [5].

Table 9. Performance evaluation of Neural Network algorithm

Class	Precision	Recall	F1-Score
No	0.64	0.59	0.61
Yes	0.94	0.95	0.95

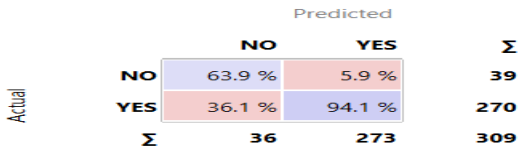

Figure 9. Neural Network Confusion Matrix

Results and Discussion

The "Lung Cancer" dataset used in this study has 309 occurrences, 16 attributes including 15 class attribute and 1 prediction attributes which was downloaded from Kaggle. To assess the effectiveness of the proposed approach, a 10-fold cross-validation was executed, where "K" represented the number of folds. The model's efficacy was evaluated using performance measures. The results of ML techniques applied to this dataset are presented in Table 10.

Table 10. ML different models performance for lung cancer prediction.

Model	Accuracy (%)	Precision (%)	Re-call (%)	Score (%)	AUC (%)
KNN	87.4	84.4	87.4	85.0	79.4
Decision Tree	85.8	87.0	85.8	86.3	73.5
SVM	90.3	89.5	90.3	89.8	91.0
Random Forest	90.6	89.8	90.6	90.1	92.4
Neural Network	90.6	90.3	90.6	90.5	92.9
Naïve Bayes	90.0	90.3	90.0	90.1	91.5
Logistic Regression	92.2	91.8	92.2	92.0	93.6

The researchers employed a correlation analysis to explore the connection between the feature variables and target variables. To accomplish this, they utilized a correlation matrix to gauge the extent of association between two or more variables. Correlation matrices are valuable tools for summarizing data and identifying more intricate investigations. Based on the findings of the analysis, the dataset on "Lung Cancer" displayed a robust negative correlation with "peer pressure" and "yellow fingers," along with a feeble negative correlation with "smoking" and "age." Conversely, a strong positive correlation was observed between lung cancer and "chest pain" and "alcohol consumption."

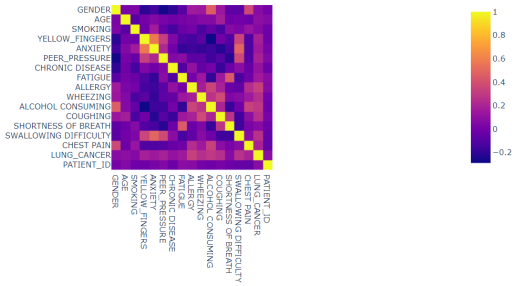

Figure 10. Correlation Matrix

To assess the internal validity of the individual early lung cancer risk predictors, a Matplotlib Venn diagram was utilized. This shows the following:

- It is significantly more likely to get lung cancer if one BOTH smokes and consumes alcohol (3 vs 90).
- It is significantly more likely to get lung cancer if one EITHER smokes OR consumes alcohol (16 vs 65 and 4 vs 75).
- There is a substantial possibility that one may still have lung cancer without smoking or alcohol consumption (40 cases).

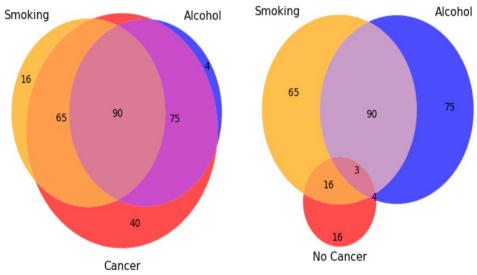

Figure 11. Matplotlib-venn diagram

Venn diagrams are a valuable tool for elucidating well-defined concepts by visually representing their similarities and differences. This is particularly advantageous for coordinate concepts, which share some or many attributes, as a Venn diagram can effectively highlight the distinguishing characteristics between them. Additionally, recall, precision, accuracy, and the AUC-ROC curve were evaluated using a confusion matrix as a machine learning (ML) technique. The confusion matrix presents a summary of predictions in matrix form, indicating the number of correct and incorrect predictions for each class. This facilitates the identification of classes that are being misclassified as another class, and the relevant formulae were utilized accordingly [20]. TP represents a person diagnosed with lung cancer. FP represents positive report but no Lung Cancer. TN represents a negative report stating that there is no evidence of lung cancer. FN represents a negative report that results in lung cancer diagnosis [8].

Table 11. A model-derived confusion matrix for different ML.

Algorithm	TP	TN	FP	FN
KNN	8	262	8	31
Neural Network	23	257	13	16
Random Forest	19	261	9	20
SVM	20	259	11	19
Tree	21	244	26	18
Naïve Bayes	25	253	17	14
Logistic Regression	24	261	9	15

Fig.11 shows the accuracy scores of different ML techniques used to predict cancer. Logistic regression achieved the best accuracy of 92.2%, whereas the random forest and Neural Network both achieved 90.6% accuracy. SVM and Naïve Bayes with accuracies of 90.3% and 90.0%, respectively. Finally, both the KNN and Decision tree had the lowest scores.

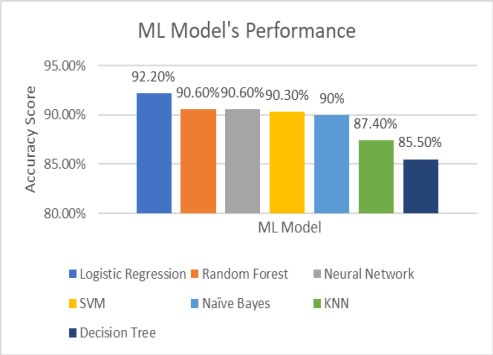

Figure 11. Accuracy score for each ML Technique.

The inclusion of Accuracy transforms the AUC into a second statistic that assesses how effectively a model can differentiate between various types of data and aids in identifying which model performs best [17]. Based on several simulations, this probability curve explores the TP and FP Rate. An indicator used to gauge a model's capacity to distinguish between positive and negative categories is the AUC. With a range of values from 0 to 1, a higher AUC value is preferred [8]. A test with a value of 0 is fully incorrect, whereas a test with a value of 1 is entirely accurate. An AUC of 0.5 indicates no discrimination in classifying individuals with cancer or illness based on the test [21]. The AUC range illustrates the accuracy of the test. AUCs between 0.7 and 0.8 are regarded as excellent, while those between 0.8 and 0.9 are regarded as outstanding. AUC values over 0.9 are regarded as having very high performance [5]. Figures 12, 13, 14, and 15 show the AUC curves and mean results from the K-fold 10 cross-validation for the group models. This work has a sizable potential influence on the early detection and treatment of lung cancer. By identifying lung cancer sooner providing individualized treatment plans, this study can benefit patients and society.CT, X-ray, and MRI images may be used to detect lung cancer trends using ML algorithms. individualized treatment programs depending on the features of the patient. Genetic information, medical history, and treatment results may all be examined by ML algorithms to find patterns and correlations that can guide clinicians in choosing the best course of therapy. Medical professionals and machine learning models must work together to offer safe and efficient patient care. The limitations that might affect this study are the small sample size, limited ability to predict outcomes, and bias in the data.

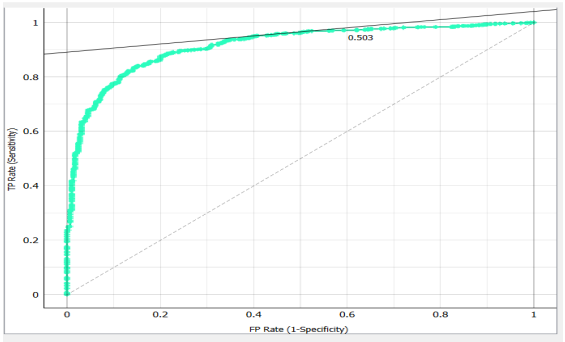

Figure 12. AUC of Logistic Regression

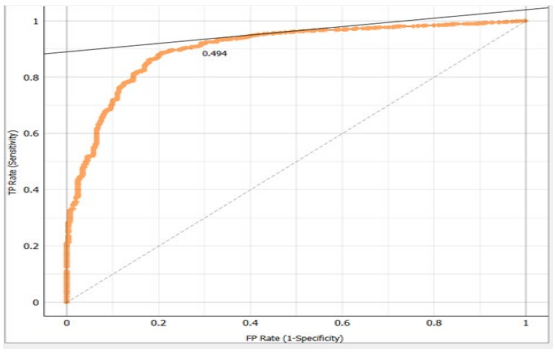

Figure 13. AUC of Naïve Bayes

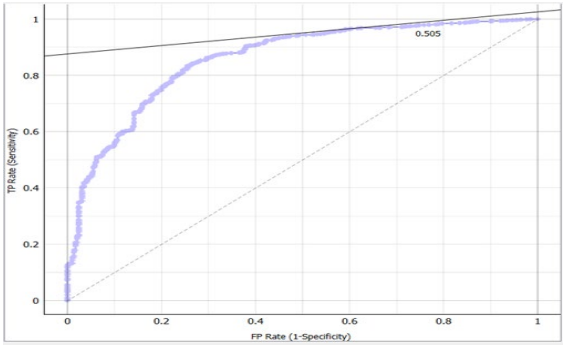

Figure 14. AUC of SVM

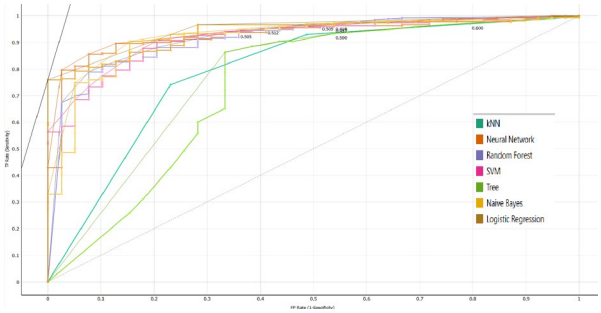

Figure 15. ROC curve for Seven different models.

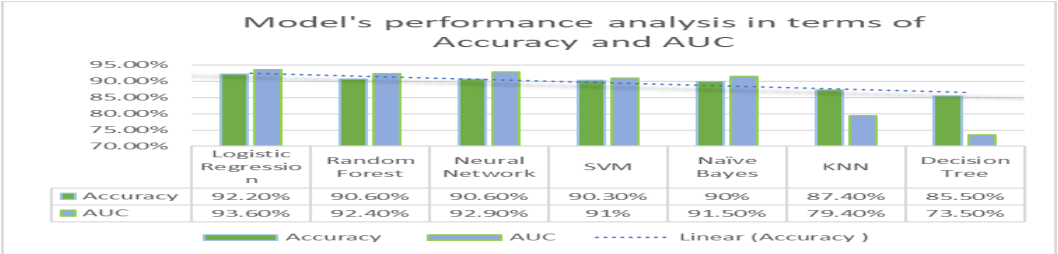

Figure 15. Analysis of ML model's performance in terms of Accuracy and AUC

Conclusion

In this study, we examined the effectiveness of our ML models and reviewed them along with several cautionary papers on lung cancer prediction. To choose the best preventative or suppressive procedure, it is essential to identify the lung cancer. This study was conducted on 309 patients with different cancer symptoms such as chest pain, age, smoking, and chronic disease. Seven different ML techniques were used: KNN, SVM, NB, DT, RF, NN, and LR, to obtain the best model that can be used for lung cancer prediction. Ten folds were used for cross validation is 10 for the model validation evaluation process. According to our model's examination, "logistic regression" performed better than all the models, and it is considered that our model system has the highest accuracy of 92.2%, precision of 91.8%, AUC of 93.6%, score of 92.0%, and recall of 92.2%. Our findings are promising for early lung cancer prediction and might help people with similar symptoms to undergo a lung cancer medical check to avoid destroying their lungs. In addition, doctors could benefit from our study by applying the same algorithms to a group of people who are expected to have lung cancer. Although our findings are encouraging, we could still predict even better outcomes. ML techniques are engaging and powerful, and ML can accurately and sensitively recognize cancer types. Data size and the possibility of missing data were discovered as limitations throughout the data collection process. (Missing values, even the absence of a piece or a sizable portion of the data, might restrict its applicability.) and data accuracy, which could lead to erroneous predictions.

Our results are promising; however, a larger and more balanced dataset may improve them. In the future, our approach may aid in early lung cancer diagnosis, treatment, and biomedical research. We may also use a deep neural network lung sound detection collaborative learning algorithm, cancer diagnosis, heart failure prediction, and other diseases using ML algorithms.

References

[1] A. S. Sakr, "Automatic Detection of Various Types of Lung Cancer Based on Histopathological Images Using a Lightweight End-to-End CNN Approach," Institute of Electrical and Electronics Engineers (IEEE), Jan. 2023, pp. 141–146. doi: 10.1109/esolec54569.2022.10009108.

[2] "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques." [Online]. Available: www.ijcsit.com

[3] G. Paliwal and U. Kurmi, "A Comprehensive Analysis of Identifying Lung Cancer via Different Machine Learning Approach," in *Proceedings of the 2021 10th International Conference on System Modeling and Advancement in Research Trends, SMART 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 691–696. doi: 10.1109/SMART52563.2021.9675304.

[4] T. Christopherp and J. Jamera Banup, "Study of Classification Algorithm for Lung Cancer Prediction," 2016. [Online]. Available: www.ijiset.com

[5] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *2022 IEEE World AI IoT Congress, AIIoT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.

[6] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, "Lung Cancer prediction using machine learning: A comprehensive approach," in *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)*, IEEE, 2020, pp. 108–115.

[7] T. Kadir and F. Gleeson, "Lung cancer prediction using machine learning and advanced imaging techniques," *Transl Lung Cancer Res*, vol. 7, no. 3, p. 304, 2018.

[8] "Lung Cancer Detection Using Chi-Square Feature Selection and Support Vector Machine Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 3, pp. 2050–2060, Jun. 2021, doi: 10.30534/ijatcse/2021/801032021.

[9] *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*. IEEE, 2018.

[10] N. Krishnan, M. Karthikeyan, Thiagarajar College of Engineering, Institute of Electrical and Electronics Engineers. Madras Section. Podhigai Subsection, Institute of Electrical and Electronics Engineers. Madras Section. Signal Processing/Computational Intelligence/Computer Joint Societies Chapter., and Institute of Electrical and Electronics Engineers, *2018 IEEE International Conference on Computational Intelligence and Computing Research: 2018 December 13-15: venue: Thiagarajar College of Engineering, Madurai, Tamilnadu, India.*

[11] J. Wu *et al.*, "A machine learning method for identifying lung cancer based on routine blood indices: Qualitative feasibility study," *JMIR Med Inform*, vol. 7, no. 3, Jul. 2019, doi: 10.2196/13476.

[12] Institute of Electrical and Electronics Engineers, Denki Gakkai (1888), IEEE Engineering in Medicine and Biology Society. Thailand Chapter., and T. International Computer Science and Engineering Conference (22nd: 2018: Chiang Mai, *BMEiCON - 2018: the 11th Biomedical Engineering International Conference: November 21-24, 2018, Chaing Mai, Thailand.*

[13] "Lung Cancer Prediction using Data Mining Techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 12301–12305, Nov. 2019, doi: 10.35940/ijrte.d9914.118419.

[14] D. Chauhan and V. Jaiswal, "An efficient data mining classification approach for detecting lung cancer disease," in *Proceedings of the International Conference on Communication and Electronics Systems, ICCES 2016*, Institute of Electrical and Electronics Engineers Inc., 2016. doi: 10.1109/CESYS.2016.7889872.

[15] A. Singh, R. Kumar, and R. Rastogi, "Study of Machine Learning Models for the Prediction and Detection of Lungs Cancer," in *Proceedings of the 2022 11th International Conference on System Modeling and Advancement in Research Trends, SMART 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1243–1248. doi: 10.1109/SMART55829.2022.10047610.

[16] D. M. Abdullah, A. M. Abdulazeez, and A. B. Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 141–149, 2021.

[17] M. Phillips *et al.*, "Prediction of lung cancer using volatile biomarkersinbreath 1," IOS Press, 2007.

[18] H. T. Sadeeq, S. Y. Ameen, and A. M. Abdulazeez, "Cancer Diagnosis based on Artificial Intelligence, Machine Learning, and Deep Learning," in *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 656–661. doi: 10.1109/3ICT56508.2022.9990784.

[19] A. Priyanga, M. Phil, and S. Prakasam, "Effectiveness of Data Mining-based Cancer Prediction System (DMBCPS)," 2013.

[20] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2018.