Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P. 22727-22740 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Necessity of Machine Learning Algorithm in Health Issues: Prioritizing Lung Cancer Diagnosis

[1]Monalisa Hati , [2]Hindh K Nasar, [3]Abhinav V, [4]Ayeshaa.S , [5]Sourabh Kumar Singh, [6]Ayisha Raina, [7]Muhammed Bisher

[1]Assistant Professor, Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India
ssamit6@gmail.com
[2]Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India
[3]Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India
[4]Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India
[5]Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India
[6]Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India
[7]Department of Computer Science and Engineering, AMITY School of Engineering and Technology, AMITY University, Mumbai, Maharashtra, India

How to cite this article: Monalisa Hati, Hindh K Nasar, Abhinav V, Ayeshaa.S , Sourabh Kumar Singh, Ayisha Raina, Muhammed Bisher.(2024) Necessity of Machine Learning Algorithm in Health Issues: Prioritizing Lung Cancer Diagnosis. *Library Progress International*, 44(3), 22727-22740

Abstract:
Lung cancer is one of the leading causes of cancer-related deaths worldwide, underscoring the urgent need for early and accurate prediction to improve survival rates. This study presents a comparative analysis of the performance of various machine learning algorithms for lung cancer prediction. Several popular algorithms, including Support Vector Machines (SVM), Random Forest, k-Nearest Neighbors (k-NN), Logistic Regression, and Neural Networks, are evaluated using publicly available datasets. The models are assessed based on metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Key factors such as feature selection, data preprocessing, and hyperparameter tuning are also explored to optimize the performance of each algorithm. The results highlight the strengths and limitations of different techniques in handling complex lung cancer data, providing insights into the most suitable algorithms for clinical applications. This comparative approach aims to assist researchers and healthcare professionals in selecting robust models for early detection and personalized treatment strategies for lung cancer.

Keywords: Machine learning, lung cancer prediction, comparative analysis, model performance, early detection

INTRODUCTION:

Lung cancer is a major public health challenge, accounting for the highest number of cancer-related deaths worldwide. Despite advances in treatment, early detection remains critical in improving patient outcomes and survival rates. Traditional diagnostic methods, such as imaging techniques and biopsies, are often time-consuming, expensive, and prone to human error. In this context, machine learning (ML) algorithms have emerged as powerful tools for predictive analytics, offering the potential to enhance early detection and assist healthcare providers in clinical decision-making.

Machine learning techniques have shown significant promise in medical applications by analyzing large volumes of patient data to detect patterns that may not be apparent to human experts. For lung cancer prediction, a wide range of ML algorithms have been developed, including Support Vector Machines (SVM), Random Forests, Neural Networks, Logistic Regression, and ensemble methods. However, the performance of these algorithms can vary depending on factors such as data quality, feature selection, algorithm parameters, and the specific

characteristics of the dataset used. Therefore, it becomes essential to compare the performance of multiple algorithms to identify the most effective models for accurate lung cancer prediction.

This study aims to provide a comparative evaluation of several machine learning algorithms for lung cancer prediction, focusing on their predictive accuracy and reliability. By using publicly available datasets, we assess the strengths and limitations of each algorithm based on key performance metrics such as accuracy, sensitivity, specificity, precision, and the area under the ROC curve (AUC). Additionally, we explore the importance of feature engineering and hyperparameter tuning in optimizing model performance.

The primary objective of this research is to offer insights into the suitability of different algorithms for lung cancer prediction, helping researchers and practitioners select the most appropriate model for clinical applications. This comparative approach not only provides a benchmark for future studies but also supports the development of more effective and personalized diagnostic tools for early-stage lung cancer detection.

LITERATURE REVIEW

The application of machine learning (ML) in healthcare, particularly for cancer diagnosis, has grown rapidly over the last decade. In lung cancer prediction, ML algorithms have been leveraged to enhance early detection, risk assessment, and personalized treatment strategies. This section reviews recent research on various ML algorithms used in lung cancer prediction, focusing on their performance, advantages, and challenges, along with comparative studies that provide insights into selecting optimal models for clinical applications.

Lung Cancer Prediction Using ML Algorithms

Several studies have highlighted the potential of ML models to predict lung cancer more accurately than traditional diagnostic methods. Popular ML algorithms used in lung cancer research include Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (k-NN), Logistic Regression (LR), and Neural Networks (NN). These algorithms have been applied to datasets containing clinical, imaging, and genetic information to classify patients as high-risk or low-risk for lung cancer.

For example, **Kumar et al. (2021)** developed a predictive model using SVM, reporting high precision in identifying early-stage lung cancer. Similarly, **Ali et al. (2020)** demonstrated that Random Forest could achieve competitive accuracy, particularly in datasets with high-dimensional features such as CT scan data. However, these studies also identified challenges, such as sensitivity to data imbalance, where models tend to perform poorly with minority class samples.

Comparative Studies of Machine Learning Algorithms

Comparative studies are essential to identify the strengths and limitations of different ML algorithms. **Chen et al. (2019)** conducted a performance comparison between Logistic Regression, SVM, and k-NN on a lung cancer dataset. Their findings indicated that while SVM achieved higher accuracy, Logistic Regression was more interpretable, making it preferable for clinical use. Conversely, **Patil et al. (2022)** found that ensemble methods like XGBoost outperformed traditional models by reducing overfitting and achieving better generalization.

Another study by **Wang et al. (2020)** explored deep learning models, particularly Convolutional Neural Networks (CNNs), for image-based lung cancer detection. They demonstrated that CNNs could effectively extract features from CT scans and achieve high prediction accuracy, though the model required large datasets and significant computational power. Such findings highlight the trade-off between performance and resource requirements, which must be carefully considered in practical applications.

1. Machine learning algorithms in lung cancer prediction

Machine Learning Algorithms in Lung Cancer Prediction

Machine learning (ML) algorithms have demonstrated significant potential in the healthcare domain by uncovering patterns in complex datasets that facilitate accurate predictions and diagnostics. For lung cancer prediction, ML models aim to analyze clinical, imaging, and genomic data to distinguish malignant cases from benign ones. Various algorithms, ranging from traditional statistical models to more advanced neural networks, are employed to enhance early detection and improve patient outcomes. This section provides an overview of the key machine learning algorithms commonly used in lung cancer prediction, highlighting their working principles, strengths, and challenges.

Support Vector Machine (SVM):

SVM is a supervised learning algorithm that works well for binary classification tasks. It constructs hyperplanes to separate different classes with maximum margin. SVM has proven effective in lung cancer prediction, especially with smaller, structured datasets. However, it can struggle with large datasets and noisy data.

Random Forest (RF)

Random Forest is an ensemble learning method based on decision trees. It builds multiple trees during training and combines their outputs to improve prediction accuracy and reduce overfitting. RF is particularly useful for handling high-dimensional data and has been widely applied in lung cancer studies for feature selection and prediction.

Logistic Regression (LR)

Logistic Regression is a simple and interpretable algorithm often used for binary classification tasks, such as predicting the presence or absence of lung cancer. While it is easy to implement, it may struggle with non-linear patterns, making it less suitable for complex data compared to more advanced models.

k-Nearest Neighbors (k-NN)

k-NN is a non-parametric algorithm that assigns class labels based on the majority vote of the nearest neighbors in the feature space. It performs well with smaller datasets but is computationally expensive with large datasets. Moreover, its performance can be sensitive to feature scaling and the choice of hyperparameters, such as the number of neighbors.

Neural Networks (NN)

Neural networks, especially deep learning models, have gained prominence in medical diagnosis tasks. These models are capable of learning complex patterns from large, high-dimensional datasets such as radiological images. While neural networks offer high prediction accuracy, they require extensive computational resources and large amounts of labeled data for training.

XGBoost and Other Ensemble Methods

XGBoost, an optimized gradient boosting algorithm, is known for its speed and performance. It has become popular in recent years due to its ability to handle missing data, avoid overfitting, and provide high accuracy. Ensemble methods like XGBoost and stacking techniques combine the strengths of multiple algorithms to enhance predictive performance.

Naïve Bayes (NB)

Naïve Bayes is a probabilistic algorithm that assumes independence between features. While it is less accurate than other models for complex tasks, it can be useful for rapid predictions and works well with small datasets.

Deep Learning Models for Image-Based Prediction

Convolutional Neural Networks (CNNs) are widely used for image-based lung cancer detection. They excel at extracting features from radiological images such as CT scans and X-rays, providing high accuracy in detecting malignancies. However, these models require significant computational resources and large datasets to avoid overfitting.

2. DATASET PREPARATION AND ANALYSIS

A dataset for predicting lung cancer has been gathered from the original source. The dataset has 310 occurrences and 16 contributors in total. In the examples provided, the dataset characteristics are split between male and female. The 16 input feature qualities in the lung cancer research dataset that are utilized to predict lung cancer are all described in detail in Table 1. Habits and Symptoms are the two categories into which the qualities are separated. Habits and Symptoms can have positive or negative values, shown by the numbers 2 [yes] and 1 [no] correspondingly in Table 2. The dataset's provided instances had thirty-three duplicate items, which were eliminated prior to processing. After doing the instance frequency count, the positive case distribution was examined according to gender. Additionally, a gender-specific frequency analysis of the patient's behaviors and symptoms has been conducted. To determine how important one characteristic is in relation to the others, a Pearson's Correlation has also been shown as a heat map. The clinical dataset's characteristics were selected by specialists in this field in order to assess the efficacy of the cancer prediction system, which in turn assists patients in determining their risk for cancer at a reasonable cost and making decisions on the best course of treatment. Two sets of data—one for training (80%) and the other for testing (20%)—are separated. Ten-fold cross-validation was performed on each model during training. To do this, the training set was divided into two subsets: one for training and one for validation proportion of 10:1 to refine the qualities. The results of the 10 cross-validated models and the Area Under Curve (AUC) and Receiver Operating Curve (ROC) were used to determine the final accuracy measure.

Attribute	Description [values]	Values
Gender	Indicates gender of the patient	M [Male], F [Female]
Age	Age of patients	Numeric value
Smoking	Smoking habit of patient	2 [Yes], 1 [No]
Yellow_fingers	Patient has symptom of yellow finger	2 [Yes], 1 [No]
Anxiety	Patient having anxiety	2 [Yes], 1 [No]
Peer_pressure	Patient undergoes peer pressure	2 [Yes], 1 [No]
Chronic disease	Any chronic diseases identified	2 [Yes], 1 [No]
Fatigue	Patient having fatigue	2 [Yes], 1 [No]
Allergy	Patient facing any allergy	2 [Yes], 1 [No]
Wheezing	Breathing with a husky or whistling sound	2 [Yes], 1 [No]
Alcohol consuming	Patient is alcoholic	2 [Yes], 1 [No]
Coughing	Patient having cough problem	2 [Yes], 1 [No]
Shortness of breath	Patient facing shortness of breath	2 [Yes], 1 [No]
Swallowing difficulty	Patient having difficulty in swallowing	2 [Yes], 1 [No]
Chest pain	Patient having cough problem	2 [Yes], 1 [No]
Lung_cancer	Lung cancer detected in patient	Yes[Positive], No [Negative]

Table 1. Description of all 16 input attributes in lung cancer study dataset.

Types	Attributes	Values and description
Habits	Smoking	1 [No], 2 [Yes]
	Alcohol consuming	
Symptoms	Yellow fingers	
	Anxiety	
	Peer pressure	
	Fatigue	
	Chronic diseases	
	Allergy	
	Wheezing	
	Chest pain	
	Cough	
	Shortness of breath	
	Swallowing difficulty	

Table 2. List of patient's habits and symptoms in lung cancer study dataset.

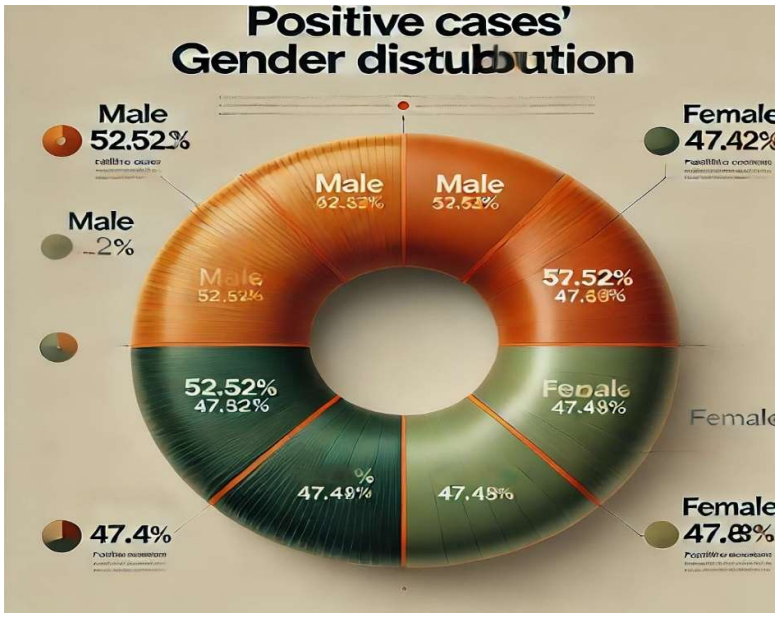


Figure 1. Positive case distribution gender-wise.

3. RESULTS AND DISCUSSION

Figure 1 illustrates that 52.52% of men and 47.48% of females are afflicted with the condition, while Figure 2 demonstrates that the majority of the distribution was found within the age range between 55 and 75 years. The data has mostly been studied based on positive and negative cases among males and females across the age distribution.

Next, the distribution of positive and negative samples of patient habits—namely, smoking and alcohol consumption—is examined. Of these, 54.2% of males and 45.80% of females have positive smoking cases, while 69.65% of males and 30.35 percent of females have positive alcohol consumption cases. The results of the gender-wise positive and negative case distribution over patient behaviors are shown in Figure 3.

The patient's symptoms serve as the basis for the third observation, which is yellow fingers, anxiety, long-term illness, exhaustion, chest discomfort, coughing, wheezing, shortness of breath, difficulty swallowing, and allergies. 42.5% of men and 57.5% of women have yellow fingers, 41.6% of men and 50.4% of women suffer from anxiety, 43.5% of men and 56.5% of women have a chronic illness, 66.2% of men and 33.8% of women have chest pain, 50.9% of men and 49.1% of women have fatigue, 57% of men and 43% of women have wheezing, 57% of men and 43 women have coughing, 52.3% of men and 47.7% of women have short breath, 46.6% of men and 52.4% of women have swal-lowing, and 58.2% of men and 41.8% of women have allergies, according to the comprehensive study. The gender-wise distribution of positive and negative instances over patient symptoms is shown in Figure 4.

This finding indicates that chronic coughing, yellow finger, and When examining data according to gender, disease, chest pain, and allergies are important symptoms to consider. But in order to understand the importance of each characteristic in connection to another, we have studied Pearson's correlation. Since alcohol intake was found to have a considerable impact (69.65%) on the detection of lung cancer, we conduct a correlation that takes alcohol consumption into account. The Pearson correlation coefficient (r) value has been explicitly analyzed using the thumb rule: if $r > 0.5$, it is Strong Positive; if $0.3 < r > 0.5$, it is Moderate Positive; and if $0 < r > 0.3$, it is Weak Positive. According to the association heat map in Figure 5, there is a moderate link between alcohol use and lung cancer as well as between it and allergies and chest discomfort.

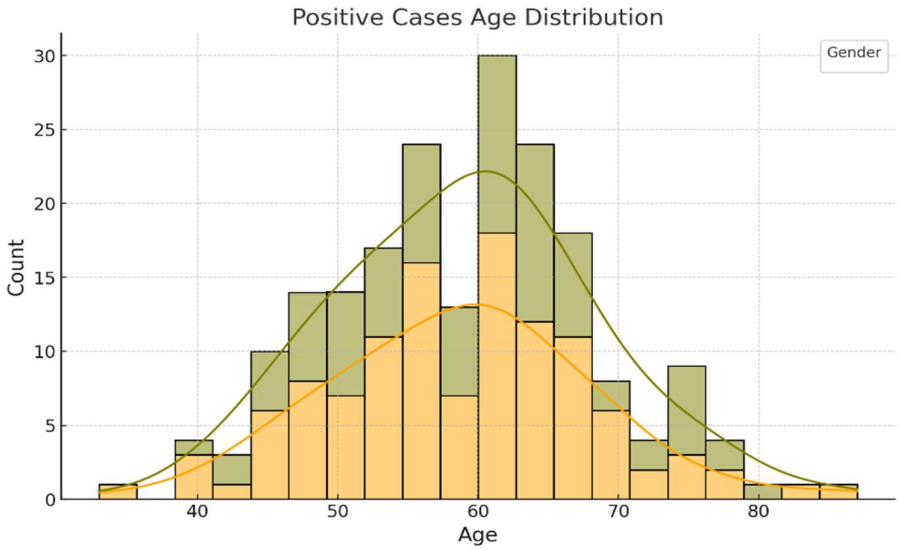
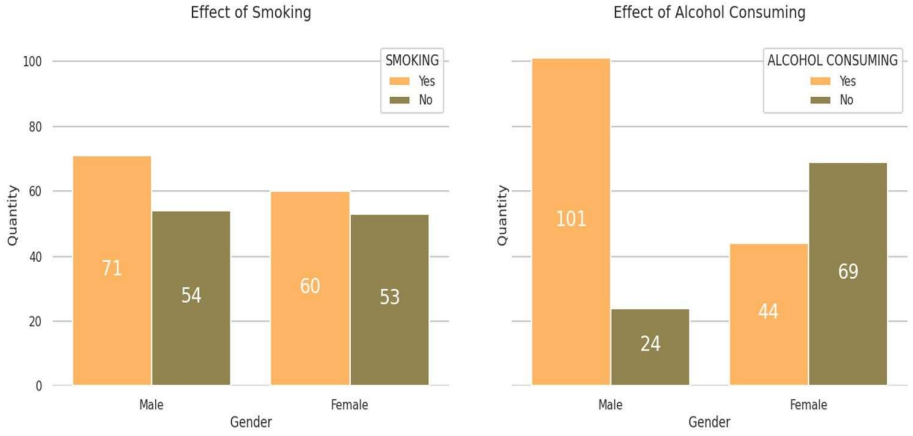


Figure 2. Positive case distribution age-wise over gender in the given dataset.

Figure 3. Positive and negative case distribution gender-wise over patient's habits.



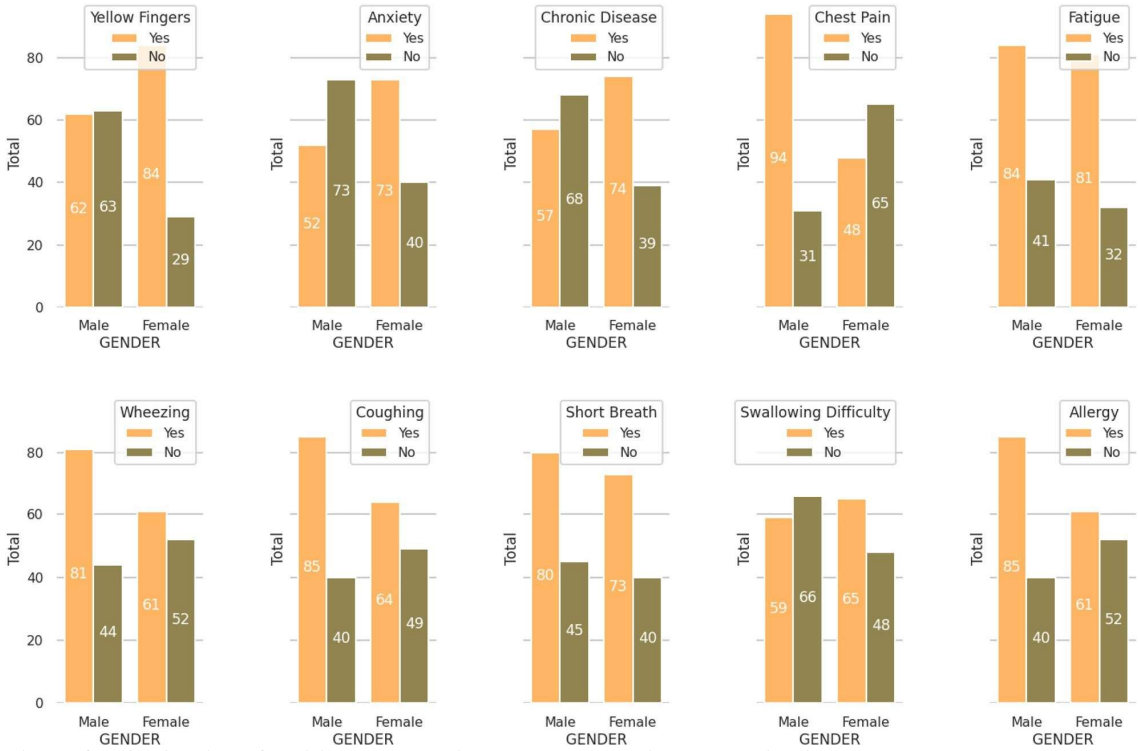


Figure 4. Distribution of positive and negative cases gender-wise over patient's symptoms.

Pearson Correlation of Features

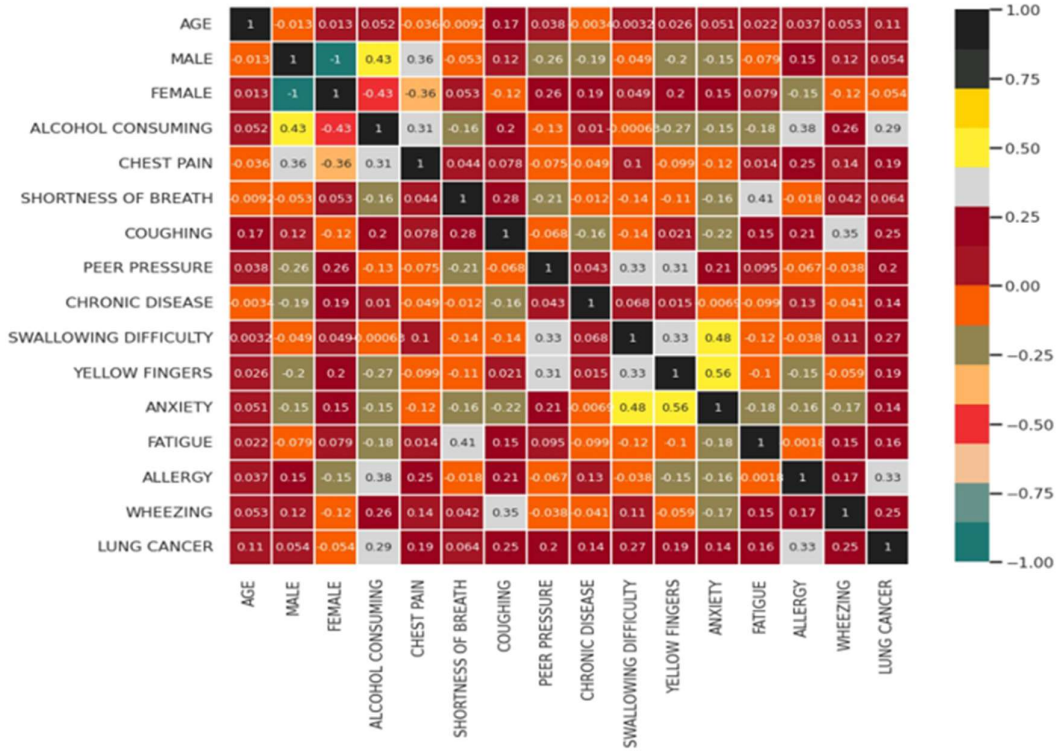


Figure 5. Correlation heat map for attributes considering alcohol consuming as habit of patient

We may now use several machine learning techniques to comprehend the algorithm's importance in this issue space.

We have determined a few learning algorithms for the prediction of lung cancer based on the literature review,

namely.
(1) Gaussian Naïve Bayes,
(2) Bernoulli Naïve Bayes,
(3) Logistic regression,
(4) Random forest,
(5) Support Vector Machine
(6), K-Nearest Neighbor
(7) High gradient enhancement
(8) Additional tree, Ensemble_1 with XGB and ADA, Ensemble_2 with Voting Classifier
(9) Ada boost
(10) Multilayer Perceptron (MLP).

3.1 **Comparison of performance of algorithms**

Following initial statistical analysis, we applied several machine learning algorithms to the lung cancer clinical dataset. The dataset has been condensed for the lung cancer prediction model based on the characteristics' correlation analysis. However, a thorough classification report has also been made available with each approach, and the ROC curve (receiver operating characteristic curve) and AUC (area under the ROC curve) have been taken into consideration in order to assess the effectiveness of the learning algorithms confusion matrix. The models were evaluated using the Jupyter v7.0.6 run environment with Python v3.11 support; the model's correctness is determined by the precision, recall, F1-score, and support in the confusion matrix, ROC/AUC, and classification report. Three sections make up the analysis. components: (1) Confusion Matrix, (2) ROC/AUC, and (3) Classification Report. The classification report provides the models' overall statistics, while the confusion matrix aids in calculating accuracy and recall, which in turn affects the F1-score and AUC. The confusion matrix for all of the machine learning models that have been described is shown in Figure 6, which also compares all of the machine learning techniques that have been addressed. The AUC graph, which compares all of the machine learning techniques presented on the confusion matrix, is shown in Figure 7 for each of the models that have been reviewed.

Also, the classification report for the accuracy of all the compared machine learning methods is shown in the set of Tables 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14. for the prediction of lung cancer. According to the comparison analysis, K-Nearest Neighbor has the best accuracy (92.86%), followed by Bernoulli Naïve Bayes and Gaussian Naïve Bayes (91.07%) in Table 15. Finally, we can say that the Bernoulli Naïve Bayes and K-Nearest Neighbor models perform better on the smaller dataset with binary features. They work better in datasets when features and characteristics are extremely independent. Other models could not do better for the dataset since they rely on correlation and the dataset's training/testing separation.

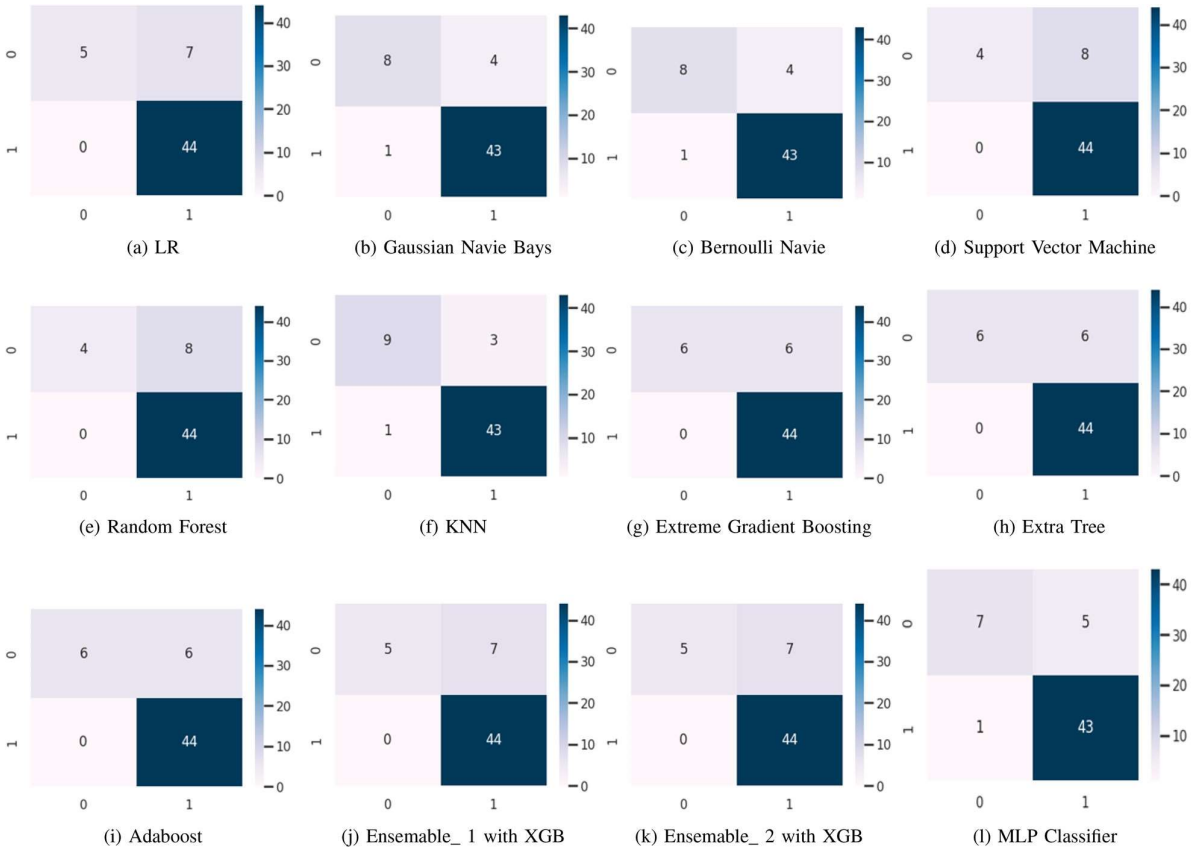
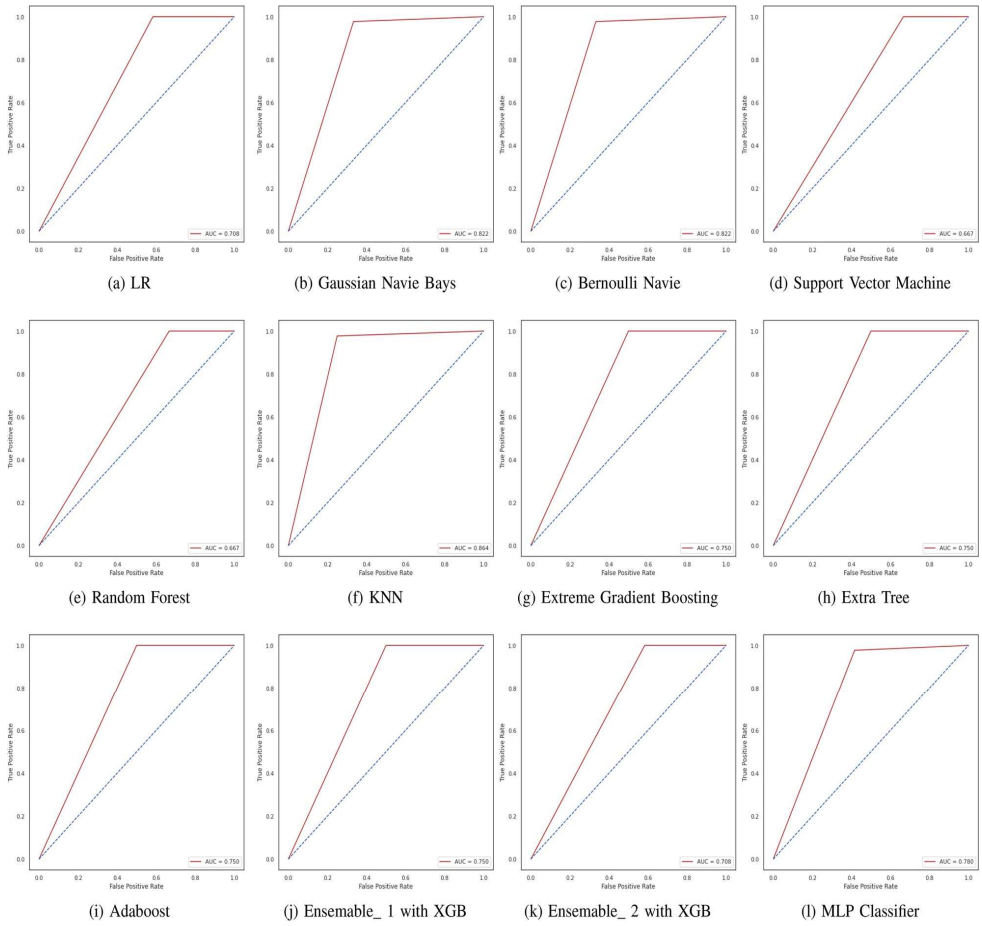


Figure 6. A comparative study of learning algorithm through confusion matrix over lung cancer dataset



(a) LR (b) Gaussian Navie Bays (c) Bernoulli Navie (d) Support Vector Machine

(e) Random Forest (f) KNN (g) Extreme Gradient Boosting (h) Extra Tree

(i) Adaboost (j) Ensemable_ 1 with XGB (k) Ensemable_ 2 with XGB (l) MLP Classifier

Figure 7. A comparative study of learning algorithm through ROC/AUC over lung cancer dataset

	Precision	Recall	f1-score	Support
0	1.00	0.42	0.59	12
1	0.86	1.00	0.93	44
Macro avg	0.93	0.71	0.76	56
Weighted avg	0.89	0.88	0.85	56
Accuracy			0.88	56

Table 3. Classification report for LR classifiers. The accuracy of logistic regression is 87.5%.

	Precision	Recall	f1-score	Support
0	0.89	0.67	0.76	12
1	0.91	0.98	0.95	44
Macro avg	0.90	0.82	0.85	56
Weighted avg	0.91	0.91	0.91	56
Accuracy			0.91	56

Table 4. Classification report for Gaussian Naive Bayes classifiers. The accuracy of Gaussian Naive Bayes is 91.07%.

	Precision	Recall	f1-score	Support
0	0.89	0.67	0.76	12
1	0.91	0.98	0.95	44
Macro avg	0.90	0.82	0.85	56
Weighted avg	0.91	0.91	0.91	56

Accuracy			0.91	56

Table 5. Classification report for Bernaulli Navie classifier. The accuracy of Bernoulli Naive Bayes is 91.07%.

	Precision	Recall	f1-score	Support
0	1.00	0.33	0.50	12
1	0.85	1.00	0.92	44
Macro avg	0.92	0.67	0.71	56
Weighted avg	0.88	0.86	0.83	56
Accuracy			0.86	56

Table 6. Classification report for SVM classifier. The accuracy of Support Vector Machine is 85.71%.

	Precision	Recall	f1-score	support
0	1.00	0.33	0.50	12
1	0.85	1.00	0.92	44
Macro avg	0.92	0.67	0.71	56
Weighted avg	0.88	0.86	0.83	56
Accuracy			0.86	56

Table 7. Classification report for Random Forest Classifiers. The accuracy of Random Forest Classifier is 85.71%.

	Precision	Recall	f1-score	Support
0	0.90	0.75	0.82	12
1	0.93	0.98	0.96	44
Macro avg	0.92	0.86	0.89	56
Weighted avg	0.93	0.93	0.93	56
Accuracy			0.93	56

Table 8. Classification report for K Nearest Neighbors Classifier. The accuracy of K Nearest Neighbors Classifier is 92.86%.

	Precision	Recall	f1-score	Support
0	1.00	0.50	0.67	12
1	0.88	1.00	0.94	44
Macro avg	0.94	0.75	0.80	56
Weighted avg	0.91	0.89	0.88	56
Accuracy			0.89	56

Table 9. Classification report for Extreme Gradient Boosting Classifier. The accuracy of extreme gradient boosting classifier is 89.29%.

	Precision	Recall	f1-score	Support
0	1.00	0.50	0.67	12
1	0.88	1.00	0.94	44
Macro avg	0.94	0.75	0.80	56
Weighted avg	0.91	0.89	0.88	56
Accuracy			0.89	56

Table 10. Classification report for Extra Tree Classifier. The accuracy of extra tree classifier is 89.29%.

	Precision	Recall	f1-score	Support
0	1.00	0.50	0.67	12
1	0.88	1.00	0.94	44
Macro avg	0.94	0.75	0.80	56
Weighted avg	0.91	0.89	0.88	56
Accuracy			0.89	56

Table 11. Classification report for Ada Boost Classifier. The accuracy of ada boost classifier is 89.29%.

	Precision	Recall	f1-score	Support

0	1.00	0.50	0.67	12
1	0.88	1.00	0.94	44
Macro avg	0.94	0.75	0.80	56
Weighted avg	0.91	0.89	0.88	56
Accuracy			0.89	56

Table 12. Classification report for Ensemble_1 with XGB and ADA Classifier. The accuracy of Ensemble_1 with XGB and ADA Classifier is 89.29%.

	Precision	Recall	f1-score	Support
0	1.00	0.42	0.59	12
1	0.86	1.00	0.93	44
Macro avg	0.93	0.71	0.76	56
Weighted avg	0.89	0.88	0.85	56
Accuracy			0.88	56

Table 13. Classification report for Ensemble_2 with Voting Classifier. The accuracy of Ensemble_2 with Voting Classifier is 87.5%.

	Precision	Recall	f1-score	Support
0	0.88	0.58	0.70	12
1	0.90	0.98	0.93	44
Macro avg	0.89	0.78	0.82	56
Weighted avg	0.89	0.89	0.88	56
Accuracy			0.89	56

S. No.	Model name	Accuracy (%)
1	Logistic Regression	87.5
2	Gaussian Naive Bayes	91.07
3	Bernoulli Naive Bayes	91.07
4	Support Vector Machine	85.71
5	Random Forest	85.71
6	K-Nearest Neighbors	92.86
7	Extreme Gradient Boosting	89.29
8	Extra Tree	89.29
9	ADA Boost	89.29
10	Ensemble_1 with XGB and ADA	89.29
11	Ensemble_2 with Voting Classifier	87.5
12	MLP	89.29

Table 15. A comparison of the accuracy of different learning algorithms applied over lung cancer

4. CONCLUSION

Lung cancer prediction can be helpful if the technique is effective once symptoms are identified and also corresponds with the patient's lifestyle and low-risk cancer status. Furthermore, depending on the patient's cancer risk level, the specialist may suggest the best course of action. However, when forecasting lung cancer in a patient, accuracy is crucial. After processing the 310 instances of raw data to identify positive cases by gender, each attribute's individual positive cases were compared by gender. According to a preliminary examination of the data,

yellow finger and allergies are the most common symptoms in a correlation research over alcohol drinking patterns. This study concentrated on thoroughly examining twelve possible machine types. The K-nearest neighbor and Bernoulli Naïve Bayes models, which perform as well as Gaussian Naïve Bayes, are determined to be appropriate learning algorithms with respective accuracy rates of 92.86% and 91.07%.

4.1 **LIMITATIONS AND FUTURE SCOPE**

The potential of several machine learning algorithms using textual clinical data for the early identification of lung cancer is demonstrated in this work. However, the limited dataset used in this study is dependent on the symptoms and behaviors of the patient. To examine the variation in the algorithm's performance, the study may be conducted on a bigger dataset. Additionally, a fine correlation might be created to increase the early-stage detection method's effectiveness. Furthermore, as the categorization was done based on symptoms and behaviors, this study may also be conducted on a bigger legitimate dataset that needs to have at least these 16 criteria. For bigger datasets, certain possible algorithms, including Ensemble 1 with XGB and ADA and Multilayer Perceptron (MLP), may be further examined. In addition to Based on statistical research and data observation, it is believed that men who regularly drink alcohol and experience symptoms like allergies and chest discomfort are more likely to get lung cancer. But for this understanding, which might result in the creation of a weighting system for the particular attribute in lung cancer diagnosis, we need an expert opinion. Data from Electronic Health Records (EHRs) might be essential for lung cancer early diagnosis.

The clinical data may thus be used to determine the similarities between the patient's parameters and the AUC that the applied model was able to obtain.

Further research is required to confirm this model acceptance process.

5. **Data availability**

The datasets generated and/or analysed during the current study are available in the Data source: https://www.kaggle.com .

6. **REFERENCES**

1. Organization, W. H. et al. A vision for primary health care in the 21st century: towards universal health coverage and the sustainable development goals (World Health Organization, Tech. Rep., 2018).
2. Yue, H., He, C., Huang, Q., Yin, D. & Bryan, B. A. Stronger policy required to substantially reduce deaths from pm2. 5 pollution in China. Nat. Commun. 11(1), 1462 (2020).
3. Organization, W.H. National cancer control programmes: Policies and managerial guidelines. World Health Organization, (2002).
4. Hamann, H. A., Ver Hoeve, E. S., Carter-Harris, L., Studts, J. L. & Ostroff, J. S. Multilevel opportunities to address lung cancer stigma across the cancer control continuum. J. Thoracic Oncol. 13(8), 1062–1075 (2018).
5. Valentine, T. R., Presley, C. J., Carbone, D. P., Shields, P. G. & Andersen, B. L. Illness perception profiles and psychological and physical symptoms in newly diagnosed advanced non-small cell lung cancer. Health Psychol. 41(6), 379 (2022).
6. Maurya, S.P., Ohri, A., & Gaur, S. Relevance of spatio-temporal data visualization techniques in healthcare system. in Geospatial Data Science in Healthcare for Society 5.0. Springer, 59–78 (2022).
7. Mithoowani, H. & Febbraro, M. Non-small-cell lung cancer in 2022: A review for general practitioners in oncology. Curr. Oncol.
8. 29(3), 1828–1839 (2022).
9. Miller, K. D. et al. Cancer treatment and survivorship statistics, 2022. CA Cancer J. Clin. 72(5), 409–436 (2022).
10. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. 13, 8–17 (2015).
11. Yang, Y., Xu, L., Sun, L., Zhang, P. & Farid, S. S. Machine learning application in personalised lung cancer recurrence and surviv- ability prediction. Comput. Struct. Biotechnol. J. 20, 1811–1820 (2022).
12. Pokkuluri, K.S., Usha Devi, N., & Mangalampalli, S. Dlcp: A robust deep learning with non-linear ca mechanism for lung cancer prediction. in Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, 2021. Springer, 299–305 (2022).
13. Alsinglawi, B. et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. Sci. Rep. 12(1), 607 (2022).
14. Venkatesh, S.P., & Raamesh, L. Predicting lung cancer survivability: A machine learning ensemble method on seer data, (2022).
15. Chauhan, A. et al. Detection of lung cancer using machine learning techniques based on routine blood

indices. in 2020 IEEE international conference for innovation in technology (INOCON). IEEE, 1–6. (2020)
16. Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. in 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST). IEEE 2018, 1–4 (2018).
17. R. Patra. Prediction of lung cancer using machine learning classifier. in Computing Science, Communication and Security: First International Conference, COMS2. Gujarat, India, March 26–27, 2020, Revised Selected Papers 1. Springer 2020, 132–142 (2020).
18. Earnest, A., Tesema, G. A. & Stirling, R. G. Machine learning techniques to predict timeliness of care among lung cancer patients.
19. Healthcare. 11(20), 2756 (2023).
20. Chandran, U. et al. Machine learning and real-world data to predict lung cancer risk in routine care. Cancer Epidemiol. Biomark. Prevent. 32(3), 337–343 (2023).
21. Qureshi, R. et al. Machine learning based personalized drug response prediction for lung cancer patients. Sci. Rep. 12(1), 18935 (2022).
22. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: Enhancing cancer research and clinical oncology. Nat. Cancer 3(9), 1026–1038 (2022).
23. Nahm, F. S. Receiver operating characteristic curve: Overview and practical use for clinicians. Korean J. Anesthesiol. 75(1), 25–36 (2022).
24. Muschelli, J. III. Roc and auc with a binary predictor: A potentially misleading metric. J. Classification 37(3), 696–708 (2020).
25. Dritsas, E. & Trigka, M. Lung cancer risk prediction with machine learning models. Big Data Cognit. Comput. 6(4), 139 (2022).

s