A Machine Learning Approach for Early Detection and Improved Decision-Making for Lung Cancer Diagnosis

BY

KM. ZAYEDUL HASAN
ID: 222-17-528

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Master of Science in Management Information Systems

Supervised By

Dr. Md Zahid Hasan

Associate Professor

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2024

APPROVAL

This Thesis titled "**A Machine Learning Approach for Early Detection and Improved Decision-Making for Lung Cancer Diagnosis**", submitted by **KM. ZAYEDUL HASAN** to the Department of Computing and Information System, Under CSE, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of MS in Management Information System and approved as to its style and contents. The presentation was held on 03 February, Saturday, 2024.

BOARD OF EXAMINERS

Professor Dr. Sheak Rashed Haider Noori — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md Zahid Hasan — **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

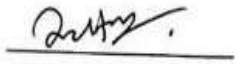
Sazzadur Ahamed — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Firoz Ahmed — **External Examiner**
Professor
Department of ICE
University of Rajshahi

DECLARATION

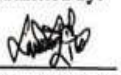
I hereby declare that, this thesis has been done by us under the supervision of **Dr. Md Zahid Hasan, Associate Professor, Department of CSE,** Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:



Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:



KM. ZAYEDUL HASAN
ID: 222-17-528
MS in MIS program
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

I really grateful and wish our profound our indebtedness to **Dr. Md Zahid Hasan** Associate Professor, Department of CSE, Daffodil International University. Deep Knowledge & keen interest of my supervisor in the field of **"Health Informatics"** to carry out this thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism , valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor and Head, Department of CSE**, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University.

I would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and passion of my parents.

ABSTRACT

Lung cancer is presently the leading cause of cancer-related mortalities worldwide. Environmental conditions, lifestyle habits, and genetics are the main causes of lung cancer. Early detection of lung cancer is pivotal in preventing its severe consequences. The integration of machine learning algorithms in the healthcare industry has led to significant advancements in disease diagnosis. These algorithms help medical professionals diagnose lung cancer accurately in the early stages. In this study, we propose using Quadratic Discriminant Analysis to improve the accuracy of lung cancer diagnosis by analyzing the symptoms of lung cancer patients. Our proposed technique is more suitable for diagnosing lung cancer with higher accuracy and precision compared to previous techniques. The methodology has demonstrated an impressive overall accuracy of 98% based on empirical results.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

Lung cancer is a major public health concern and one of the leading causes of death across the globe. According to the World Health Organization (WHO), lung cancer is responsible for 18% of global mortality, resulting in approximately 1.8 million deaths per year [1]. Consequently, early diagnosis strategies are necessary to detect and treat lung cancer globally.

The lung is a vital organ that collaborates with other organs and tissues to enable inhalation and exhalation within the respiratory system. The primary function of the lungs is to transfer oxygen from the air to the bloodstream while releasing carbon dioxide into the atmosphere through gas exchange. Lung cancer is a disease that occurs when cellular abnormalities in the lungs multiply uncontrollably, which can lead to severe and potentially life-threatening complications [2]. Early detection and prompt treatment of lung cancer markedly ameliorate the chances of a successful recovery.

Machine learning techniques have the potential to unveil hidden patterns from the clinical data that can significantly enhance decision-making in healthcare [3]. Early and accurate detection of illnesses is a crucial aspect of disease management, and machine learning can facilitate this process by enabling healthcare professionals to identify diseases during their early stages [4].

Initially, we utilized statistical analysis to determine the correlation between the risk factors of lung cancer, and then we examined the relationship between lung cancer and its features. Furthermore, we also explored the significance of the features of lung cancer using Chi-square test. Finally, we have evaluated four distinct machine learning classifiers, namely Linear Discriminant Analysis, Bootstrap Aggregating, Gradient Boosting, and Quadratic Discriminant Analysis. The Accuracy, Sensitivity, Precision, and F1-Score are employed to assess the performance of machine learning classifiers. These evaluation measures are crucial in determining the effectiveness and reliability of machine learning classifiers. Our research aims to assess the accuracy of various machine-learning approaches in correctly predicting lung cancer patients from clinical datasets.

1.2 Motivation

The primary aim of utilizing machine learning in lung cancer diagnosis is to revolutionize healthcare by saving lives. There are several key motivations:

- Early detection
- Personalized medicine
- Cost-effective treatment plans
- Data-driven improved decision-making.

1.3 Research Objectives

The research objective of machine learning in lung cancer is to implement algorithms and models that can effectively analyze medical data to anticipate the risk of lung cancer. The primary objectives are enlisted following:

- To compare the diagnostic performance of several machine learning algorithms for lung cancer.
- To compute the correlation among the risk factors of lung cancer.
- To calculate and evaluate the importance of the factors contributing to the onset of lung cancer.

1.4 Report Structure

There are five chapters in the report, each of which will be presented with guidelines. We covered the introduction of this research in Chapter 1. The importance of early detection of lung cancer using machine learning is emphasized in Chapter 1. Additionally, the motivation and objectives of this research are also discussed. In Chapter 2, we reviewed the existing literature on the diagnosis of lung cancer. Furthermore, the scope of the problem and challenges of the research are also covered. The research methodology is highlighted in Chapter 3. This chapter covers data collection procedures, data statistics, and algorithm of machine learning classifiers. The results and performance of all the machine learning classifiers of our research are explored in Chapter 4. The results of the research, the conclusion, and future work are presented in Chapter 5.

CHAPTER 2

BACKGROUND

2.1 Introduction

The analysis of previous research presents findings on diverse clinical datasets of lung cancer. Various methods and techniques such as data mining, machine learning, and deep learning were used to analyze and predict these datasets.

2.2 Related Works

Firstly, R. Patra [5] implemented different machine learning algorithms to predict lung cancer, including Neural Network, Radial Basis Function Network, Support Vector machine, Logistic Regression, Random Forest, J48, Naive Bayes, and K-nearest neighbors. The Radial Basis Function Network was found to have the highest accuracy of 81.25% on lung cancer data.

K. Danjuma [6] evaluated machine learning classification schemes for post-operative life expectancy in lung cancer patients. He applied three machine learning algorithms namely J48, Naive Bayes, and Multilayer Perceptron. The Multilayer Perceptron achieved 82.3% accuracy for clinical prognosis as the best classifier.

ACC Omar et al. [7] applied various machine learning techniques to predict lung cancer such as K-Nearest Neighbors, Support Vector Machine, Multilayer Perceptron, and RBF Classifier. The best classifier was the Multilayer Perceptron, which achieved an accuracy of 90%.

MI Faisal et al. [8] employed eight machine learning algorithms with Support Vector Machine, C4.5 Decision tree, Multi-Layer Perceptron, Neural Network, Naïve Bayes, Random Forest, Majority Voting, and Gradient-boosted Tree for the prediction of lung cancer and the suggested Gradient-boosted Tree outclassed the other ML algorithms with the accuracy of 90%.

M. Mamun et al. [9] executed four ensemble techniques namely XGBoost, LightGBM, Bagging, and AdaBoost for detecting lung cancer and the XGBoost attained the highest accuracy of 94.42%.

Finally, E. Dritsas [10] implemented 14 machine learning algorithms including Naïve Bayes, Bayesian Network, Logistic Regression, Logistic Model Tree, Support Vector Machine, Stochastic Gradient Descent , J48, Random Tree, Rotation Forest , Reduced Error Pruning Tree, Random Forest , AdaBoostM1, Artificial Neural Network, and Multi-layer Perceptron to predict lung cancer. The recommended Rotation Forest reached the most accuracy of 97.1%.

2.1 Scope of The Problem

The diagnosis of lung cancer using a machine learning approach is a complex and challenging task. We required a systematic technique to succeed, but there are flaws in that strategy. In this study, we utilized a public dataset of lung cancer that had limited data. Machine learning models may face difficulties when dealing with small sample sizes in datasets [11]. Insufficient sample size can lead to overfitting and poor model generalization in machine learning. Our training model, although effective, has been developed with limited data and underlying patterns. Therefore, there is a possibility of confusion with new sample data for diagnosing lung cancer.

2.2 Challenges

Our main goal is to increase the accuracy of our developed model to the highest possible level. However, to achieve this, we require a larger clinical dataset of lung cancer patients from local hospitals in our country. The primary dataset will replace the current secondary dataset that we have, which has a limited number of samples.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

An overview of the research and the methodologies used for study is discussed in this section. It includes the dataset, research strategy, and the structure of the study. It facilitates a reliable framework to diagnose lung cancer patients in the earlier stage. To efficaciously diagnose lung cancer, several steps must be taken. The flowchart of the proposed framework is shown in Fig. 3.1.1.

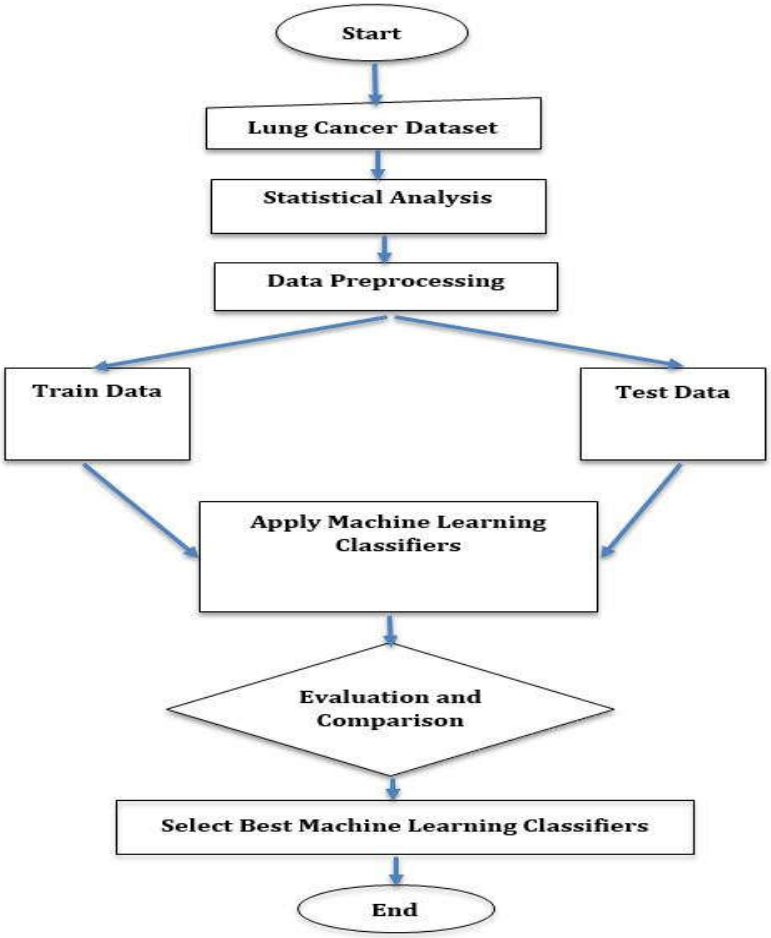


Fig. 3.1.1. The flowchart of the proposed framework.

3.2 Dataset

We utilized the overt lung cancer dataset for our research [12]. The dataset has 309 samples with 15 independent variables and one dependent variable. The feature information of the dataset is exhibited in Table 1.

Table 1: The feature information of the lung cancer dataset.

Feature	Data Type
Gender	Qualitative
Age	Quantitative
Smoking	Qualitative
Yellow Finger	Qualitative
Anxiety	Qualitative
Peer Pressure	Qualitative
Chronic Disease	Qualitative
Fatigue	Qualitative
Allergy	Qualitative
Wheezing	Qualitative
Alcohol Consuming	Qualitative
Coughing	Qualitative
Shortness of Breath	Qualitative
Swallowing Difficulty	Qualitative
Chest Pain	Qualitative
Lung Cancer	Qualitative

3.3 Statistical Analysis

The dataset contains 270 lung cancer patients and 39 healthy controls of the samples. Figure 3.3.1 depicts the information of the class variable Lung Cancer. According to the dataset, 87% of the samples have lung cancer, whereas 13% do not.

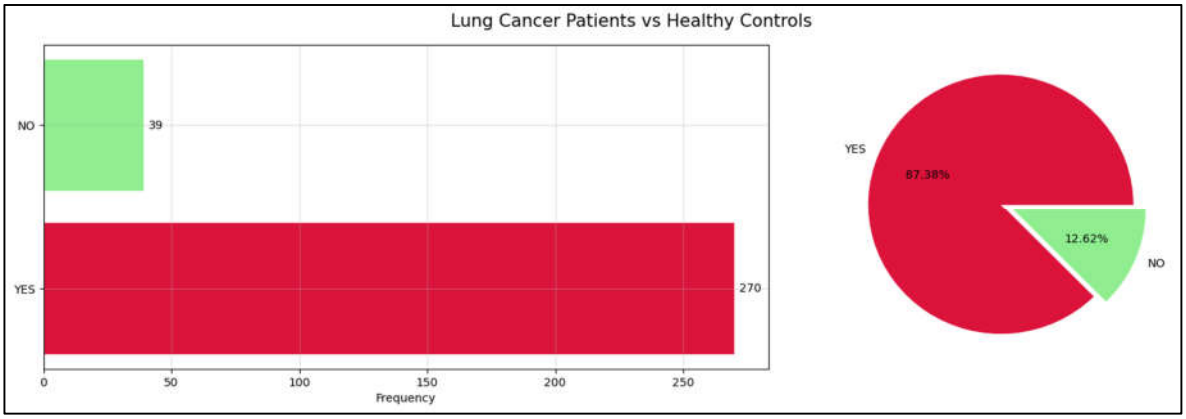


Fig. 3.3.1. The comparison between lung cancer patients and healthy controls.

The Pearson correlation analysis is utilized to identify significant attributes in this dataset that impact the classification of a lung cancer patient [13]. The Pearson's correlation among the independent variables is delineated in a pyramidal heat-map in the Fig. 3.3.2. The independent features of pairs (Yellow Finger, Anxiety), (Swallowing Difficulty, Anxiety), (Shortness of Breathiness, Fatigue) have a stronger positive correlation and their values are .60, .48, and .42 correspondingly. The independent variables of pairs (Gender, Alcohol Consuming) and (Yellow Finger, Alcohol Consuming) have a stronger negative correlation and their values are -.45, and .40 respectively.

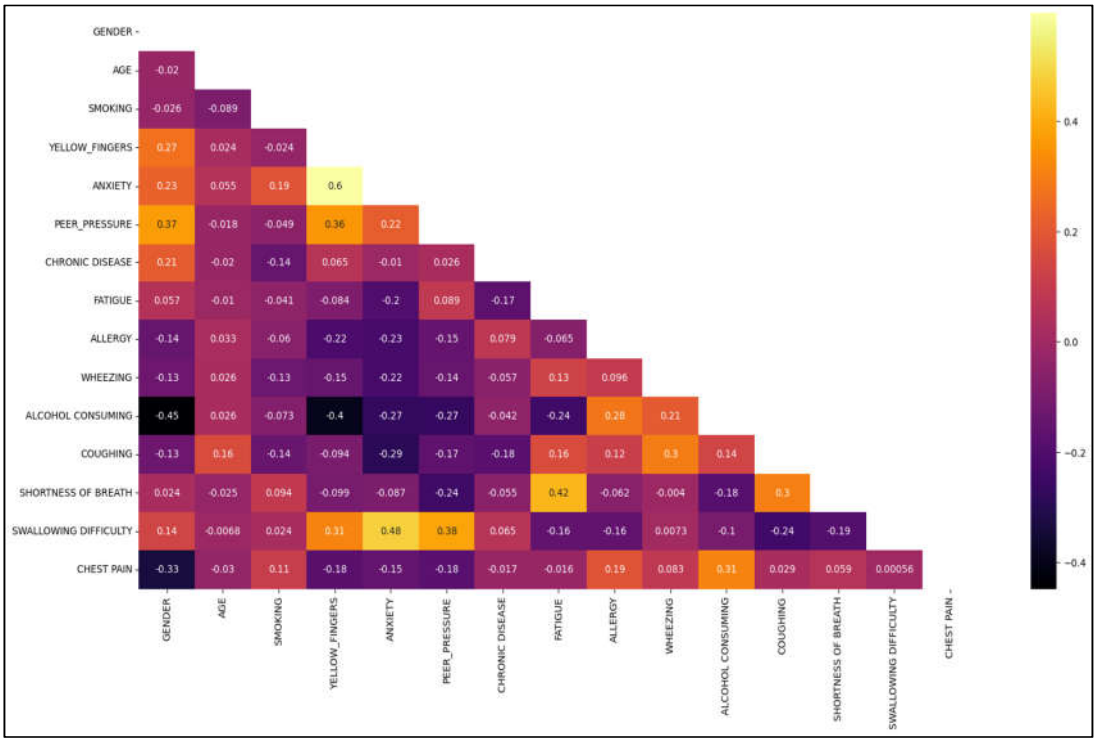


Fig. 3.3.2. The Pearson's correlation among the independent features.

The correlation of independent features to dependent feature is showed in Table 2. Independent features Allergy, Alcohol Consuming, Swallowing Difficulty, Coughing, and Wheezing are positively correlated with the dependent feature Lung Cancer and values are .33, 0.29, 0.26, .25 and .25 respectively.

Table 2: The correlation of independent features to dependent feature.

Feature	**Correlation with Lung Cancer**
Gender	-.067
Age	.089
Smoking	.058
Yellow Finger	.18
Anxiety	.14
Peer Pressure	.19
Chronic Disease	.11
Fatigue	.15
Allergy	.33
Wheezing	.25
Alcohol Consuming	.29
Coughing	.25
Shortness of Breath	.061
Swallowing Difficulty	.26
Chest Pain	.19

We utilized the chi-square test to determine the feature score according to the importance based on the relationship between a categorical variable and a class variable [14]. The feature score of association with lung cancer using the Chi-square test according to their Feature Score is presented in Table 3. The most significant features namely Allergy, Alcohol Consuming, Swallowing Difficulty, Wheezing, Coughing, Peer Pressure and Chest Pain, and their feature score 14.71, 11.40, 11.06, 8.51, 8.03, 5.35, and 4.96 respectively.

Table 3: The feature score of association with lung cancer using Chi-square test.

Feature	Feature Score of Association with Lung Cancer
Allergy	14.71
Alcohol Consuming	11.40
Swallowing Difficulty	11.06
Wheezing	8.51
Coughing	8.03
Peer Pressure	5.35
Chest Pain	4.96
Yellow Fingers	4.37
Anxiety	3.25
Fatigue	2.29
Chronic Disease	1.88
Gender	.73
Smoking	.45
Shortness of Breath	.40

3.4 Data Preprocessing

As part of our data preprocessing, we removed 33 duplicate samples from the lung cancer dataset to enhance its practicality [15]. After deleting 33 samples out of 303 samples, we scaled the remaining 276 samples for the improvement of the machine learning performance [16].

3.5 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is used for classification and recognition in machine learning, and it was introduced by Fisher [17] in 1936. To classify lung cancer patients, the primary goal of LDA is to augment the distance between the means of separated classes while reducing the variance within each class [18].

Algorithm: Linear Discriminant Analysis

Step 1: Input:

X: Lung cancer dataset, which contains of n samples and m classes

y: class label for each sample

Step 2: Calculate the mean of each class μ_k

Step 3: Calculate the scatter matrices:

Intra-class scatter matrix: S_{wc}

Inter-class scatter matrix: S_{bc}

Step 4: Calculate Eigenvalues and Eigenvector:

$S_{bc}^{-1} * S_{wc} * v = \lambda * v$; v and λ are the eigenvector and the eigenvalue

Step 5: Select Discriminative Feature :

$X_{new} = X * T$; T is the matrix with selected top eigenvector

Step 6: Classification:

After transforming the features, LDA utilizes them for classification employing other classifiers.

3.6 Bootstrap Aggregating

An ensemble-based Bootstrap Aggregating merges predictions from several models and learns from unlike subsets of data using bootstrapping to improve accuracy by reducing variance and over-fitting [19].

Algorithm: Bootstrap Aggregating
Step 1: Input: X: Lung cancer dataset
Step 2: Set initial weight to lung cancer dataset
Step 3: Normalize the weights
Step 4: To train the classification model
Step 5: To choose the based on lowest error
Step 6: To update the weight of training data of classification model
Step 7: To do classification

3.7 Gradient Boosting

Another ensemble-based Gradient Boosting was used in our study. It uses sequential training to combine weak learners with strong learners [20]. It takes input as a pair of data points (x_a, y_b) for the training and gives an output of classification tree T_G for the diagnosis of lung cancer.

3.8 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is a more generalized version of Linear Discriminant Analysis (LDA). In contrast to LDA, QDA computes separate convergence matrices for each class. [21]. The Quadratic Discriminant Function is represented in the following Equation 1.

$$\delta_i = \frac{1}{2} \log|\sum_i| - \frac{1}{2} (x - \sum_i)^T \sum_i^{-1} (x - \sum_i) + log\, \pi_i \qquad (1)$$

Algorithm: Quadratic Discriminant Analysis

Step 1: Input:

X: Data Matrix of lung cancer, which contains of N samples$[x_i]_{i=1}^{N}$

z: class label for each sample

Step 2: Compute the mean of each class μ_i

Step 3: Compute the scatter matrices

Step 4: Compute the covariance matrix for each class $\sum_i$

Step 5: Computer the inverse covariance for each class $\sum_i^{-1}$

Step 6: For all (class w_i, i = 1, 2, …, c) do

Calculate the quadratic discriminant function (δ_i) as in equation (1)

End for

Step 7: Do the classification

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Classification Performance Assessment

The evaluation measures are used to assess the classification model. The utilization of the evaluation measures is performed to determine the efficacy of the classification model [22]. The instances of Lung Cancer patients are labeled as the detected class as well as the healthy controls are designated as the non-detected class. Here,

- Positive (P): Lung Cancer patient
- Negative (N): Healthy Control
- True Positive (TP): Lung Cancer patient is classified as a detected class.
- False Positive (FP): Healthy control is categorized as a detected class.
- True Negative (TN): Healthy control is recognized as a non-detected class.
- False Negative (FN): Lung Cancer patient is classified as a non-detected class.

The evaluation measures are defined as below:

- Accuracy = $\frac{TP + TN}{TP + FN + TN + FP}$
- Sensitivity = $\frac{TP}{TP + FN}$
- Precision = $\frac{TP}{TP + FP}$
- F1-score = $\frac{2TP}{2TP + FP + FN}$

4.2 Experimental Results

The four machine learning classifiers, Linear Discriminant Analysis (LDA), Bootstrap Aggregating (Bagging), Gradient Boosting (GBoost), and Quadratic Discriminant Analysis (QDA), are trained using all features for distinguishing the Lung Cancer patients and the healthy controls. The Python library known as scikit-learn is used to perform the machine learning classification tasks [23]. These machine learning classifiers were trained and tested on 180 and 56 instances of the Lung Cancer dataset. Confusion matrices [24] are used to evaluate four classification models. The results of the classification models are shown in Table 4 - Table 7.

Table 4 shows that Linear Discriminant Analysis identifies 42 of 47 Lung Cancer patients are diagnosed as Detected, while 5 of 9 healthy controls are Non-Detected, hence the TP, FN, TN, and FP of LDA are 42, 5, 5, and 4.

Table 4: The Confusion Matrix of Linear Discriminant Analysis.

Linear Discriminant Analysis			
Lung Cancer		Class of Prediction	
		Detected	*Non-Detected*
Actual Class	*Detected*	**42**	5
	Non-Detected	4	**5**

The TP, FN, TN, and FP of Bootstrap Aggregating are 43, 4, 6, and 3, as Table 5 specifies that Bagging distinguishes 44 out of 47 Lung Cancer patients as Detected and 6 out of 9 healthy controls as Non-Detected.

Table 5: The Confusion Matrix of Bootstrap Aggregating.

Bootstrap Aggregating			
Lung Cancer		Class of Prediction	
		Detected	*Non-Detected*
Actual Class	*Detected*	**43**	4
	Non-Detected	3	**6**

Table 6 exhibits that Gradient Boosting recognizes that 7 of 9 healthy controls are Non-Detected and 44 of 47 Lung Cancer patients are Diagnosed as Detected; for these reasons, the TP, FN, TN, and FP of Random Forest are 44, 3, 7, and 2.

Table 6: The Confusion Matrix of Gradient Boosting.

Gradient Boosting			
Lung Cancer		Class of Prediction	
		Detected	*Non-Detected*
Actual Class	*Detected*	**44**	3
	Non-Detected	2	**7**

Table 7 shows that Quadratic Discriminant Analysis classifies 47 of 47 Lung Cancer patients are diagnosed as Detected, while 8 of 9 healthy controls are Non-Detected, as a result the TP, FN, TN, and FP of Quadratic Discriminant Analysis are 47, 0, 8, and 1.

Table 7: The Confusion Matrix of Quadratic Discriminant Analysis.

Quadratic Discriminant Analysis			
Lung Cancer		Class of Prediction	
		Detected	*Non-Detected*
Actual Class	*Detected*	**47**	0
	Non-Detected	1	**8**

Figure 4.2.1 shows the performance of all classification models, with Quadratic Discriminant Analysis outperforming the others. Quadratic Discriminant Analysis predicts the maximum number of True Positives and True Negatives as well as the minimum number of False Positives and False Negatives.

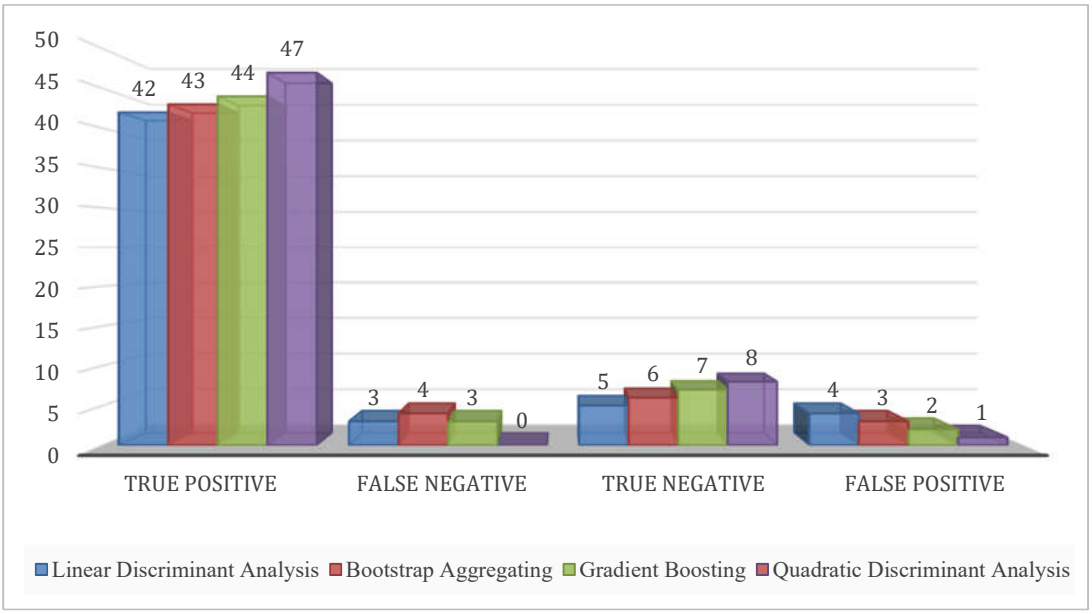


Fig. 4.2.1. The classification results using machine learning classifiers.

4.3 Discussion

The performance of the machine learning classifiers using the evaluation measures is itemized in Table 8. Quadratic Discriminant Analysis achieved the maximum accuracy of 98% whereas Linear Discriminant Analysis, Bootstrap Aggregating, Gradient Boosting achieved 84%, 88%, and 91% respectively.

While the sensitivity of LDA, Bagging, and GBoost was 89%, 91%, and 94% correspondingly, QDA accomplished the highest sensitivity of 100%.

QDA achieved the highest precision at 98%, while LDA, Bagging, and GBoost achieved 91%, 93%, and 95% respectively.

Finally, QDA reached the maximum F1-Score of 99%, whereas Linear Discriminant Analysis, Bootstrap Aggregating, and Gradient Boosting achieved 90%, 92%, and 95%, respectively.

Table 8: The classification performances using machine learning classifiers.

	Accuracy	Sensitivity	Precision	F1-Score
Linear Discriminant Analysis	.84	.89	.91	.90
Bootstrap Aggregating	.88	.91	.93	.92
Gradient Boosting	.91	.94	.95	.95
Quadratic Discriminant Analysis	.98	1	.98	.99

The comparisons of related works are presented in Table 9. These comparisons clearly show that the best classifier for predicting lung cancer is Quadratic Discriminant Analysis (QDA).

Table 9: The comparisons of related works for the diagnosis of lung cancer.

Study	Best Classifier Based on Machine Learning Technique	Accuracy
R. Patra [4]	Radial Basis Function Network	81.25%
K. Danjuma [5]	Multilayer Perceptron	82.3%
ACC Omar [6]	Multilayer Perceptron	90%
MI Faisal et al. [7]	Gradient-boosted Tree	90%
M. Mamun et al. [8]	XGBoost	94.42%
E. Dritsas [9]	Rotation Forest	97.1%
Our Study	Quadratic Discriminant Analysis	98%

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The early and precise detection of lung cancer is crucial for effective treatment. In this regard, machine learning algorithms based on a patient's symptoms have emerged as a promising tool for early diagnosis. In this study, various machine learning algorithms were analyzed to determine the most effective model for detection. The Quadratic Discriminant Analysis algorithm was found to have the highest accuracy rate of 98% and a precision rate of 100%. Early detection of lung cancer can significantly reduce associated costs and mitigate risks. Thus, the proposed system holds promise in providing a viable solution that can help improve the health outcomes of patients with lung cancer.

5.2 Future Work

We will collect a comprehensive clinical dataset of lung cancer patients from regional hospitals in our country. After this procedure is finished, we aim to implement a cloud-based approach to support clinicians in diagnosing lung cancer with higher accuracy.

REFERENCE

[1] "Lung Cancer," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer [Accessed: 24-Jan-2024].

[2] da Silva, A. L. G., Bresciani, M. J., Karnopp, T. E., Weber, A. F., Ellwanger, J. H., Henriques, J. A. P., ... & Possuelo, L. G. (2015). DNA damage and cellular abnormalities in tuberculosis, lung cancer and chronic obstructive pulmonary disease. *Multidisciplinary Respiratory Medicine*, *10*, 1-5.

[3] Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A machine learning approach to improving dynamic decision making. *Information Systems Research*, *25*(2), 239-263.

[4] Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020). Early detection of Parkinson's disease using deep learning and machine learning. *IEEE Access*, *8*, 147635-147646.

[5] Patra, R. (2020). Prediction of lung cancer using machine learning classifier. In *Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1* (pp. 132-142). Springer Singapore.

[6] Danjuma, K. J. (2015). Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. *arXiv preprint arXiv:1504.04646*.

[7] Omar, A. A. C., & Nassif, A. B. (2023, February). Lung Cancer Prediction using Machine Learning based Feature Selection: A comparative Study. In *2023 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 1-6). IEEE.

[8] Faisal, M. I., Bashir, S., Khan, Z. S., & Khan, F. H. (2018, December). An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer. In *2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST)* (pp. 1-4). IEEE.

[9] Mamun, M., Farjana, A., Al Mamun, M., & Ahammed, M. S. (2022, June). Lung cancer prediction model using ensemble learning techniques and a systematic review analysis. In *2022 IEEE World AI IoT Congress (AIIoT)* (pp. 187-193). IEEE.

[10] Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), 139.

[11] Ying, X. (2019, February). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168, p. 022022). IOP Publishing.

[12] "Lung Cancer Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer [Accessed: 24-Jan-2024].

[13] Kumar, S., & Chong, I. (2018). Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. *International journal of environmental research and public health*, *15*(12), 2907.

[14] Thaseen, I. S., Kumar, C. A., & Ahmad, A. (2019). Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers. *Arabian Journal for Science and Engineering*, *44*, 3357-3368.

[15] Huang, J., Li, Y. F., & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, *67*, 108-127.

[16] Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

[17] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179-188.

[18] Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., Green, R., Fiorillo, A., ... & Cesarelli, M. (2020). Linear discriminant analysis and principal component analysis to predict coronary artery disease. *Health informatics journal*, *26*(3), 2181-2192.

[19] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123-140.

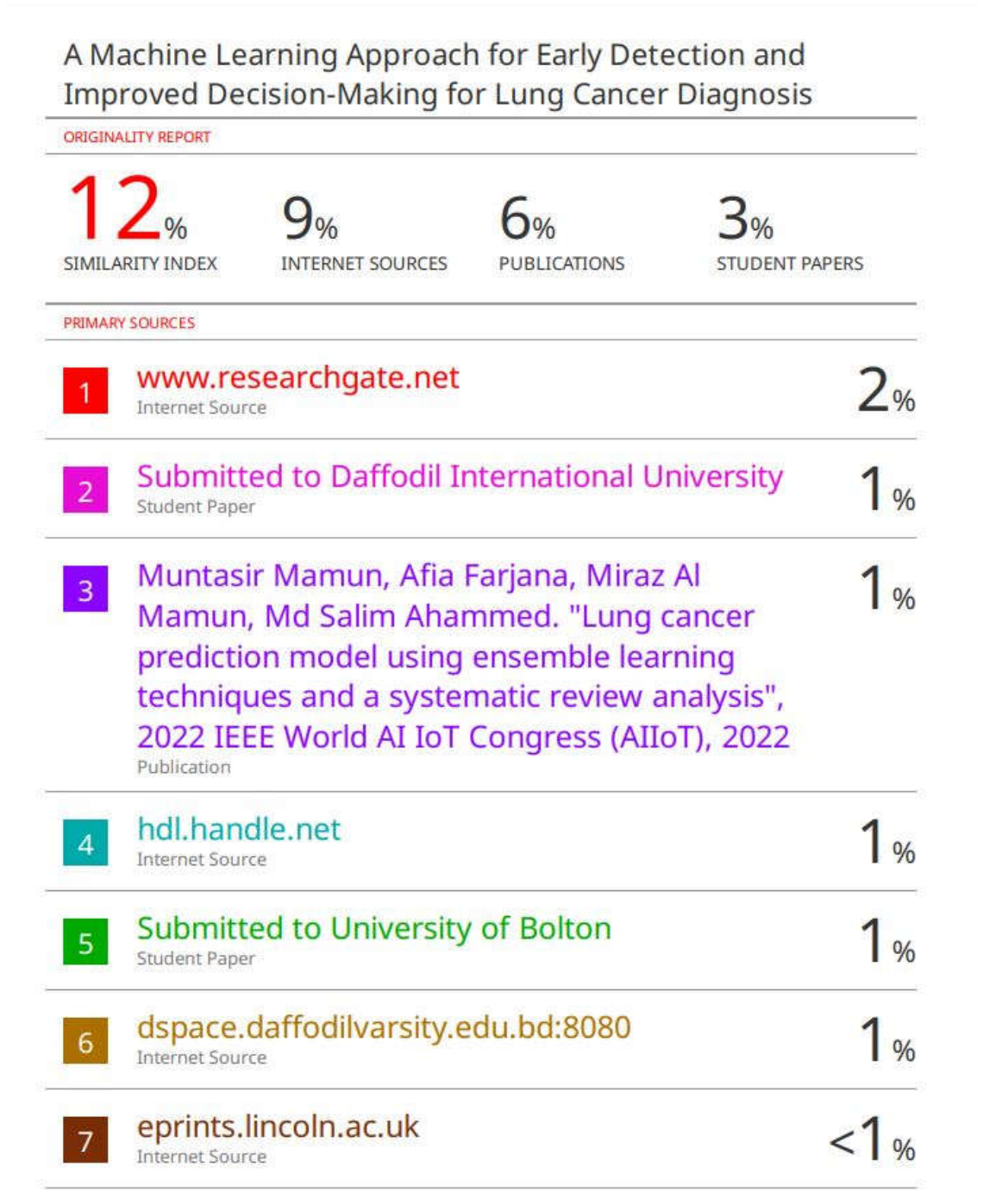
[20] Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2017, June). A statistical approach to predict flight delay using gradient boosted decision tree. In *2017 International conference on computational intelligence in data science (ICCIDS)* (pp. 1-5). IEEE.

[21] Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145-180.

[22] Zubair Hasan, K. M., & Zahid Hasan, M. (2019). Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease. In *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2018, Volume 1* (pp. 415-426). Springer Singapore.

[23] Hasan, K. Z., Datta, S., Hasan, M. Z., & Zahan, N. (2019, February). Automated prediction of heart disease patients using sparse discriminant analysis. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-6). IEEE.

[24] Japkowicz, N. (2006, July). Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning* (pp. 6-11). University of Ottawa.



A Machine Learning Approach for Early Detection and Improved Decision-Making for Lung Cancer Diagnosis

ORIGINALITY REPORT

12%	9%	6%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	Submitted to Daffodil International University Student Paper	1%
3	Muntasir Mamun, Afia Farjana, Miraz Al Mamun, Md Salim Ahammed. "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis", 2022 IEEE World AI IoT Congress (AIIoT), 2022 Publication	1%
4	hdl.handle.net Internet Source	1%
5	Submitted to University of Bolton Student Paper	1%
6	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
7	eprints.lincoln.ac.uk Internet Source	<1%