Smart Lung Cancer Prediction Using Machine Learning Algorithms

Fahad Ahmed[1*] , Hamza Muneer[2]

[1] Department of Computer Science, National College of Business Administration economics, Lahore, 54000, Pakistan

[2] Department of Computer Science, NFC-Institute of Engineering & Technology Multan, Multan, 61000, Pakistan

[*]Corresponding Author: Email: fahad.ahmed@ncbae.edu.pk

Abstract: Lung cancer is a commonly diagnosed type of cancer. It is a highly life-threatening disease. An accurate prediction of lung cancer can reduce the death rates as accurate prediction can help doctors early in their decision-making to start the treatment of the patients. In this article, three machine learning (ML) algorithms, random forest (RF), support vector machine (SVM), and XGBoost, are utilized. The proposed model's performance was evaluated using a confusion matrix. The proposed model achieved a high testing accuracy of 96.77% with the XGBoost algorithm. This study highlights the potential of using ML algorithms to enhance the accuracy of lung cancer prediction.

Keywords: Machine learning (ML); Random forest (RF); Support vector machine (SVM); XGBosst; Lung cancer

1 Introduction

One of the most common factors of death in humans worldwide is lung cancer, a disease that affects the respiratory system. The symptoms caused by lung cancer include expectoration, shortness of breath, chest pain, anorexia, fever, hemoptysis, and weight loss [1], [2].

The most crucial factor in a clinical decision-making procedure for cancer patients is a precise prognosis and survival time estimate. Predicting a patient's survival from a single moment would be highly beneficial since it allows medical experts to propose treatments based on their expected longevity.

These days, machine learning (ML) is crucial for early medical disease diagnosis and prediction, assuring human safety. The diagnosis strategy is made more deterministic and more accessible by ML. Recently, ML has already taken over the medical industry. ML approaches are currently being adopted by all countries in the healthcare industry. It is possible to investigate disease detection with the use of ML. Feature extraction is a critical application of ML that includes some of the following: The actual information container of every disease is its attributes. ML facilitates the processing of genuine features or information, makes data analysis more accessible, and identifies the actual cause of medical problems. It aids in the diagnosis of diseases by medical professionals. Three ML algorithms, i.e., random forest (RF), support vector machine (SVM), and XGBoost, are utilized in this study to predict lung cancer.

Section 2 briefly reviews previous work. Section 3 presents the proposed methodology. Section 4 presents the experiment and results, while section 5 offers a conclusion and future work.

2 Previous Work

Artificial intelligence (AI) [15-19] has been playing a vital role in the healthcare sector in predicting diseases for the past few years [2-6]. ML algorithms, a sub-branch of AI, have been increasingly employed to classify, diagnose, and predict lung cancer. Numerous investigations have revealed the effectiveness of ML in this field.

Faisal et al. [7] assessed the ML algorithms and ensembles to predict cancer at an early stage. Gradient-boosted Tree attained 90% accuracy and surpassed all other individual and ensemble classifiers based on performance evaluations.

The author [8] analyzes different ML classifiers to categorize lung cancer data into benign and malignant categories. The suggested RBF technique achieved a high accuracy of 81.25% and is regarded as a

practical approach for lung cancer prediction.

The authors [9] utilized ensemble approaches, like XGBoost, LightGBM, Bagging, and AdaBoost, to improve the prediction accuracy of lung cancer. According to the research, the XGBoost appraoch accomplished the best among the ensemble techniques, achieving an accuracy of 94.42%.

3 Proposed Methodology

The architecture of the proposed lung cancer prediction model is shown in Figure 1. After the data acquisition of the lung cancer dataset, it goes through data preprocessing. The dataset is checked for duplicate and missing values in data preprocessing. There are no missing values in the dataset, while there are some duplicate values. The duplicate values are dropped to make predictions fair. After successful data preprocessing, the dataset is split into training and testing. 80% of the dataset is used for training, while 20% is used for testing. Three ML algorithms are utilized to train the dataset.

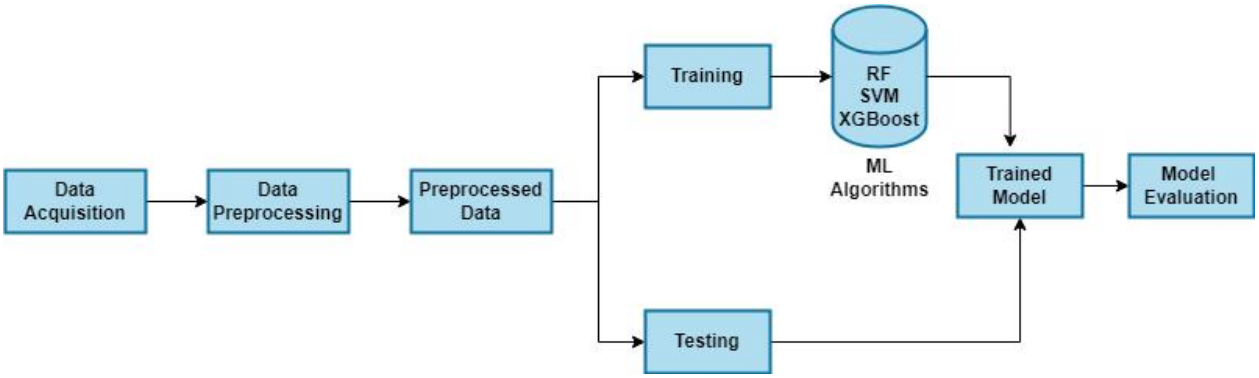

Figure 1. The architecture of the proposed lung cancer prediction model

After successful training, the testing dataset set is applied to the trained model to evaluate the model. The model evaluation exhibits ML algorithm prediction accuracies.

3.1 Dataset of the Lung Cancer

The dataset of lung cancer was acquired from Kaggle Respiratory [10]. Figure 2 displays the attributes of the dataset with their information.

Sr. no.	Attributes	Information
1	Gender	M(male), F(female)
2	Age	Age of the patient
3	Smoking	YES=2, NO=1
4	Yellow fingers	YES=2 , NO=1
5	Anxiety	YES=2, NO=1
6	Peer pressure	YES=2, NO=1
7	Chronic Disease	YES=2, NO=1
8	Fatigue	YES=2 , NO=1
9	Allergy	YES=2, NO=1
10	Wheezing	YES=2, NO=1
11	Alcohol	YES=2, NO=1
12	Coughing	YES=2, NO=1
13	Shortness of Breath	YES=2, NO=1
14	Swallowing Difficulty	YES=2, NO=1
15	Chest pain	YES=2, NO=1
16	Lung Cancer	YES, NO

Figure 2. Attributes of the dataset

To keep things simple, from the attribute of smoking to the attribute of chest pain, the digit of 2 is converted into 1 while the digits of 1 are converted into 0. Similarly, gender attributes and lung cancer are converted into integer form from object form. In the case of attribute gender, digit 1 is utilized for males, while digit 0 is utilized for females. In the case of attribute lung cancer, digit 1 is utilized for yes, while digit 0 is utilized for no.

3.2 RF

RF is an ML algorithm utilized for classification and regression tasks. During the training phase, RF constructs multiple decision trees (DTs). RFs are widely utilized in ML for their performance, simplicity, and scalability [11].

3.3 SVM

SVM is also an ML algorithm for classification and regression tasks [12], [13]. It classifies the dataset into two different groups. SVMs are utilized in text classification, spam detection, anomaly detection, and image classification [14].

3.4 XGBoost

XGBoost stands for extreme gradient boosting. It is an open-source ML library designed for gradient boosting. It is an ensemble learning technique that combines multiple DTs to construct a robust predictive model.

4 Experiment and results

The experiments and subsequent analysis of the study results are presented in this section. Two performance metrics, accuracy, and misclassification rate, are used to evaluate the performance of the proposed model (equations 1-2).

$$\text{Accuracy} = \frac{TP+TN}{+\ +\ +} * 100 \quad (1)$$

$$\text{Misclassification rate} = \frac{FP+FN}{+\ +\ +} * 100 \quad (2)$$

TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative. Figure 3 displays the testing confusion matrices of the ML algorithms.

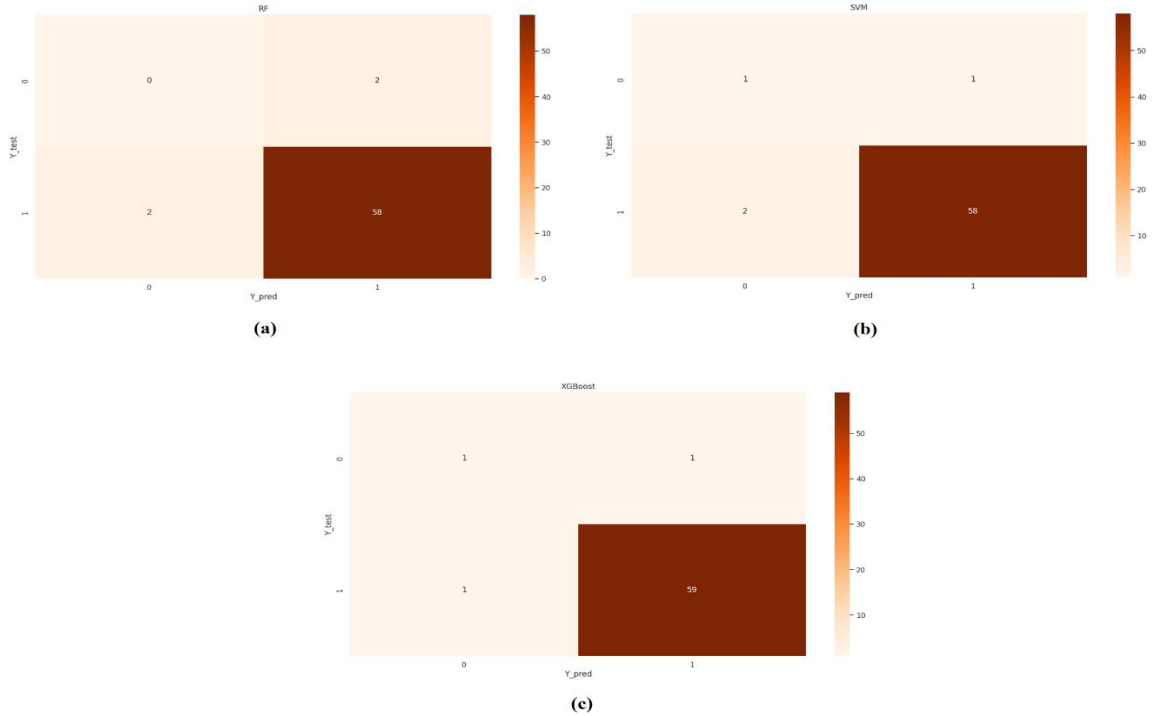

Figure 3. Testing confusion matrices of the ML algorithms, (a) RF, (b) SVM, (c) XGBoost

Figure 4 shows the comparison of the ML algorithm's prediction accuracies. It can be seen from the bar chart that XGBoost has the highest prediction accuracy among the ML algorithms that were applied.

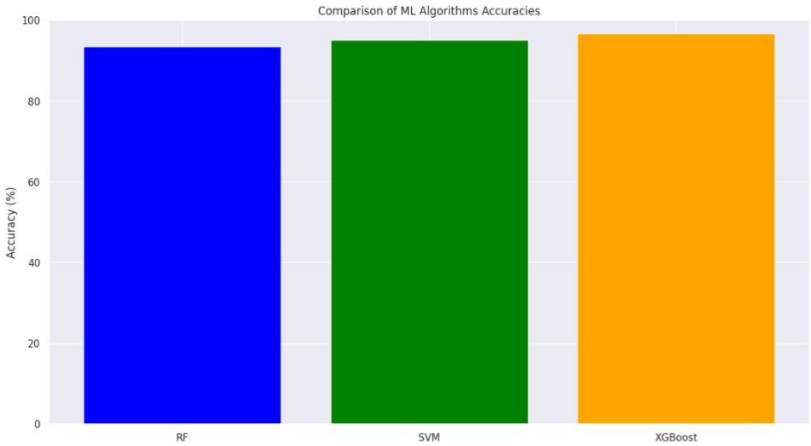

Figure 4. Comparison of ML algorithms prediction accuracies

Table 1 depicts a comparison of the proposed model with previous methodologies. The proposed model has better accuracy than the previous methodologies.

Table 1: Comparison with previous methodologies

Author	Year	Methodologies	Accuracy (%)	Misclassification rate (%)
Faisal et al. [7]	2018	Gradient-boosted Tree	90	10
Patra [8]	2020	RBF	81.25	18.75
Mamun et al. [9]	2022	XGBoost	94.42	5.58
Proposed model	2023	RF	93.55	6.45
		SVM	95.16	4.84
		XGBoost	96.77	3.23

5 Conclusion and Future Work

In conclusion, leveraging three ML algorithms for lung cancer prediction yielded promising results with notable testing accuracies. XGBoost achieved the highest accuracy of 96.77% among the ML algorithms. This proposed model will enhance the rate of lung cancer timely to start treatment earlier.

For future work, explainable artificial intelligence (XAI) techniques can be integrated with ML algorithms further to explain the decision-making of lung cancer prediction.

References

[1] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global epidemiology of lung cancer," *Annals of Global Health*, vol. 85, no. 1, pp. 1–16, 2019, doi: 10.5334/aogh.2419.

[2] S. H. Bradley, M. P. T. Kennedy, and R. D. Neal, "Recognising Lung Cancer in Primary Care," *Advances in Therapy*, vol. 36, pp. 19–30, 2019, doi: 10.1007/s12325-018-0843-5.

[3] Asghar, M. I., Ahmed, F., & Khan, S., "Harnessing Machine Learning Techniques for Intelligent Disease Prediction," International Journal of Computational and Innovative Sciences, vol. 2, no.3, pp. 1-6, 2023.

[4] Ahmed, F., Asif, M. and Saleem, M., "Identification and Prediction of Brain Tumor Using VGG-16 Empowered with Explainable Artificial Intelligence," International Journal of Computational and Innovative Sciences, vol. 2, no. 2, pp. 24-33, 2023.

[5] Athar, A., Asif, R.N., Saleem, M., Munir, S., Al Nasar, M.R. and Momani, A.M., "Improving Pneumonia Detection in chest X-rays using Transfer Learning Approach (AlexNet) and Adversarial Training," In 2023 International Conference on Business Analytics for Technology and Security (ICBATS), pp. 1-7, 2023.

[6] Sajjad, G., Khan, M.B.S., Ghazal, T.M., Saleem, M., Khan, M.F. and Wannous, M., "An Early Diagnosis of Brain Tumor Using Fused Transfer Learning," In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) , pp. 1-5, 2023.

[7] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "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)*, Karachi, Pakistan, 2018, pp. 1–4. doi: 10.1109/ICEEST.2018.8643311.

[8] R. Patra, "Prediction of Lung Cancer Using Machine Learning Classifier," in *Computing Science, Communication and Security: First International Conference, COMS2 2020*, Gujarat, India: Springer Singapore, 2020, pp. 26–27. doi: https://doi.org/10.1007/978-981-15-6648-6_11.

[9] 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)*, Seattle, WA, USA, 2022, pp. 187–193. doi: 10.1109/AIIoT54504.2022.9817326.

[10] "Lung Cancer Detection." https://www.kaggle.com/datasets/jillanisofttech/lung-cancer-detection.

[11] "Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science." https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[12] "What is a support vector machine? | Definition from WhatIs." https://www.techtarget.com/whatis/definition/support-vector-machine-SVM.

[13] "Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science." https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[14] "Support Vector Machine (SVM) Algorithm - GeeksforGeeks." https://www.geeksforgeeks.org/support-vector-machine-algorithm/.

[15] Ahmed, F., Asif, M. and Saleem, M., 2023. Identification and Prediction of Brain Tumor Using VGG-16 Empowered with Explainable Artificial Intelligence. International Journal of Computational and Innovative Sciences, 2(2), pp.24-33.

[16] Saleem, M., Khan, M.S., Issa, G.F., Khadim, A., Asif, M., Akram, A.S. and Nair, H.K., 2023, March. Smart Spaces: Occupancy Detection using Adaptive Back-Propagation Neural Network. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-6). IEEE.

[17] Athar, A., Asif, R.N., Saleem, M., Munir, S., Al Nasar, M.R. and Momani, A.M., 2023, March. Improving Pneumonia Detection in chest X-rays using Transfer Learning Approach (AlexNet) and Adversarial Training. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-7). IEEE.

[18] Abualkishik, A., Saleem, M., Farooq, U., Asif, M., Hassan, M. and Malik, J.A., 2023, March. Genetic Algorithm Based Adaptive FSO Communication Link. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-4). IEEE.

[19] Sajjad, G., Khan, M.B.S., Ghazal, T.M., Saleem, M., Khan, M.F. and Wannous, M., 2023, March. An Early Diagnosis of Brain Tumor Using Fused Transfer Learning. In 2023 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-5). IEEE.