International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Exploring the Efficacy of Machine Learning Classifiers in Lung Cancer Prognosis and Risk Assessment

G.Sireesha

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: Lung cancer remains a significant public health challenge worldwide, with various risk factors contributing to its development. This study aims to investigate the association between age, smoking status, alcohol consumption, coughing, allergies, and lung cancer using a comprehensive dataset. **Objectives:** Our analysis utilized a dataset containing information on individuals diagnosed with lung cancer, including demographic factors and lifestyle habits. Machine learning now days has a great influence to health care sector because of its high computational capability for early prediction of the diseases with accurate data analysis. **Method:** The lung cancer prediction was analysed using different machine learning classification algorithms such as Naive Bayes, Random Forest (RF) and K- Nearest Neighbour (KNN) to boost the performance. **Findings:** Among various metrics, we used accuracy and confusion matrix to analyze different machine learning classifiers to explore the relationships between these parameters and lung cancer incidence. **Novelty:** These methodologies have been used to determine lung cancer patient survival rates and to assist clinicians in providing accurate prognosis. The key objective of this paper is the early diagnosis of lung cancer by examining the performance of classification algorithms. Finally this paper contributes understanding of the disease, improving diagnostic methods, developing more effective treatments, or offering new approaches for patient care.

Keywords: *Smoking history, Cancer diagnosis, Risk factors, early detection and occupational Hazards.*

1. Introduction

In recent years, lung cancer is a common cancer across the globe. For the early prediction of lung cancer, medical practitioners and researchers require an efficient predictive model, which will reduce the number of deaths. Machine learning makes the system to find the solution of problem with own learning strategies [1]. ML classifies into three categories such as unsupervised learning, supervised learning, Reinforcement learning. Supervised learning identifies two processes under its umbrella, one is classification and another is regression. Classification is process in which input data is processed and categorized in to certain group. Machine learning now days plays a crucial role for detection and prediction of medical diseases at early stages of safe human life. Machine Learning makes diagnosis process easier and deterministic. Every county is now adopting machine learning techniques in their health care sector. With the application of machine learning the actual detection of diseases can be explored. Preliminary findings suggest a significant association between smoking status and lung cancer, with smokers exhibiting a higher risk compared to non-smokers. Key parameters of interest included age at diagnosis, smoking history, alcohol consumption patterns, presence of chronic coughing, and history of allergies [2]. Lung cancer cannot be prevented but its risk can be reduced. So detection of lung cancer at the earliest is crucial for the survival rate of patients. The number of chain- smokers is directly proportional to the number of people affected with lung cancer additionally advanced age was identified as a significant risk factor, particularly among older individuals. Interestingly, while alcohol consumption showed some correlation with lung cancer incidence, the relationship was less pronounced than that observed with smoking. As a result of the fact that the majority of patients are diagnosed at a more advanced stage, lung cancer is the primary cause of death resulting from cancer [3]. There is currently no chance of a successful treatment being developed. Lung cancer is consistently ranked as one of the most lethal forms of the disease, regardless of whether a country is industrialized or developing [4]. The incidence of lung cancer in developing countries is on the rise as a result of a longer life expectancy, more urbanization, and the adoption of Western lifestyles. The early detection of cancer and the survival of people with the disease are both essential to the control of lung cancer. Machine learning techniques have been successfully implemented to analyse large electronic health records, to identify risk factors for complications using decision trees and cross-validation techniques [5]. Numerous techniques have been developed and work on the detection of lung cancer is still under way. If lung cancer is detected at an early stage, the American Cancer Society estimates that a patient has a 47 percent chance of surviving the disease. Lung cancer originates from lung and spreads up to brain and spreads Lung cancer is categorized in to two major group. One is non-small cell lung cancer and another is

VNR Vignana Jyothi Institute of Engineering and Technology Hyderabad, India.
** Corresponding Author Email: sireesha.ganga@gmail.com*

small cell lung cancer. Some of the symptoms which are associated with the patients like severe chest pain, dry cough, breathlessness, weight loss etc. Looking into the cultivation of cancer and its causes doctors give stress more on smoking and second-hand smoking as if the primary causes of lung cancer. Treatment of lung cancer involves surgery, chemotherapy, radiation therapy, Immune therapy etc The early detection of cancer and the survival of people with the disease are both essential to the control of lung disease To achieve the highest level of detection accuracy, which is approaching 100% other technologies must yet be because the diagnostic process calls for using human intelligence to make crucial decisions, it costs time and money. It is widely accepted that machine learning is capable of significantly contributing to this essential task [6]. Exploring different methods of machine learning methods to diagnose lung cancer will be a prime aim in this paper.

2. Proposed Model:

Lung cancer is consistently ranked as one of the most lethal forms of the disease, regardless of whether a country is industrialized or developing. For the early prediction of lung cancer, medical practitioners and researchers require an efficient predictive model, which will reduce the number of deaths. This section shows an accurate classification and prediction of lung cancer using technology that is enabled by machine learning [7]. This paper proposes a lung cancer prediction model by using three classifiers namely KNN, Random forest and Naïve Bayes were employed to categorise a dataset after the creation of a machine learning pipeline. The accuracy and F1 ratings of these classifiers are then used to assess their performance and the hyper-parameters for each machine learning techniques were optimized.

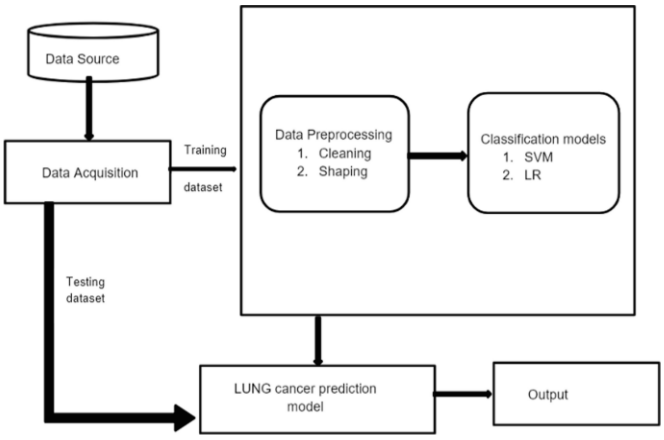

Fig: 1 Flow diagram for Lung cancer prediction

3. Classification Techniques:

3.1 Decision Trees:

Decision trees are supervised machine learning tools that split data based on parameters into nodes and leaves to construct a tree-like model. Information gain is used for node division. Key parameters include the following: Maximum depth: controls tree depth for capturing complex relationships while preventing overfitting. Minimum samples to split: requires a certain sample count before node splitting, preventing premature splits [8]. Minimum samples in a leaf: sets the minimum samples needed in a leaf node, curbing over fitting. Maximum leaf nodes: sets the upper limit for leaf nodes, aiding in avoiding over fitting. Splitting criterion: chooses metrics like “GINI” or “entropy” to assess split quality based on impurity or uncertainty. Class weight: adjusts class influence to manage imbalanced data by assigning weights, with “balanced” being an automated option.

3.2 Random Forest:

A random forest employs multiple decision trees and averages their predictions. Cross validation selects variables and cut-offs. Each tree considers a random subset of predictors for decorrelation[10]. Tuneable parameters for model optimization include the following: number of trees, maximum, minimum samples to split, and splitting criterion.

3.3 Naive Bayes:

Naive Bayes models can be optimized by adjusting key parameters: Smoothing parameter: Manages probability smoothing. Lower alpha means less smoothing, risking over fitting. Higher alpha increases smoothing, aiding prevention of overfitting and rare feature impact [11]. Prior probabilities: Sets initial class probabilities. Defaults based on training data frequencies. Adjust for prior knowledge of class distribution.

$P(c|x) = (P(x|c) P(c) / P(x)$ where

- $P(c|x)$ is posterior probability

of class(c) given predictor (x).

- $P(c)$ is the prior probability of a class.
- $P(x|c)$ is the likelihood which is the probability of a predictor given class.
- $P(x)$ is prior probability of a predictor.

3.4 K-Nearest Neighbour:

The main parameter for the k-nearest neighbour (k-NN) algorithm: k: The count of nearest neighbours considered for prediction. Its choice impacts the bias–variance trade-off. Smaller k leads to flexible, low-bias but high-variance models. Larger k yields rigid, high-bias but low-variance models. k’s value is tuned through cross validation or other methods.

4. Results and Discussion:

The Input data consists of missing values. So it is required to prepossess the data such that the missing values have been replaced with the most occurrence value of the corresponding column. The prepossessed data is converted into suitable form for classification using different classifier approach [12]. Generally in confusion matrix Accuracy, Recall, Precision and F-Measure are the key process parameter for classification. Machine learning (ML) helps for easy of data analysis and process the real attributes or information and finds the actual problem creator of diseases. It helps medical expert to find the root cause of diseases. Machine learning now a day's have already dominated medical field. Every county is now adopting machine learning techniques in their health care sector. With the application of machine learning the actual detection of diseases can be explored. Classification accuracy is the measure of number of correct prediction made out from total number of prediction. These parameters depend on some specific outcome. Those are TP (True Positive) which is the correctly predicted event values and TN (True Negative) is correctly predicted no event values. Similarly FP (False Positive) is incorrectly predicted event values and FN (False Negative) for incorrectly predicted no event values [13]. The goal of this research is to evaluate the capacity of Random forest, Naïve Bayes and KNN algorithms to predict the survival rates of lung cancer patients. Additionally, it compares the accuracy, precision, recall, and F1 score of the two algorithms. These methodologies have been used to determine lung cancer patient survival rates and to assist clinicians in providing accurate prognosis. The user interface enables the user to enter the necessary parameters and decide if the patient has cancer.

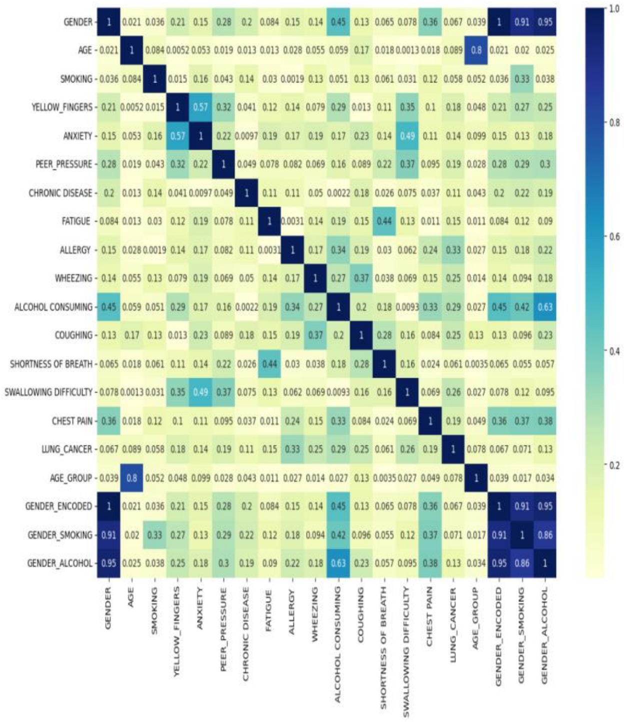

Fig: 2 Correlation Matrix

5. EVALUATION METRICS:

A confusion matrix is utilized in this project to evaluate the precision, recall, and F1 score. Confusion matrices are helpful because they allow you to compare values such as True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) in a direct manner. For performance comparison three parameters, accuracy, sensitivity, and specificity are used.

5.1 Precision

The number of TP divided by the number of TP plus the number of FP equals precision.

$$Precision = TP/ (TP + FP)$$

5.2 Recall

The number of TP divided by the number of TP plus the number of FN equals recall.

$$Recall = TP/ (TP + FN)$$

5.3 F1-score

The F1-score is the harmonic mean of the precision and recall.

$$F1=2*((Precision*Recall) / (Precision+Recall))$$

5.4 Accuracy

$$Accuracy = TP + TN / (TP +TN+ FP+FN)$$

5.5 Specificity

The number of TN divided by the number of TN plus the number of FP equals specificity

$$Specificity = TN/ (TN + FP)$$

Results of different machine learning predictors are shown in Figure: 3 the accuracy of random forest and KNN is nearer.

Table: 1 Classifiers output

S. No	Classifiers	Lung Cancer Yes(1) or No(0)	Precision	Recall	F1-score	Accuracy
1	RF	0	0.5	0.5	0.5	0.98
		1	0.98	0.98	0.98	
2	KNN	0	0.5	0.5	0.5	0.98
		1	0.98	0.98	0.98	
3	NB	0	0.5	0.33	0.4	0.97
		1	0.97	0.98	0.97	

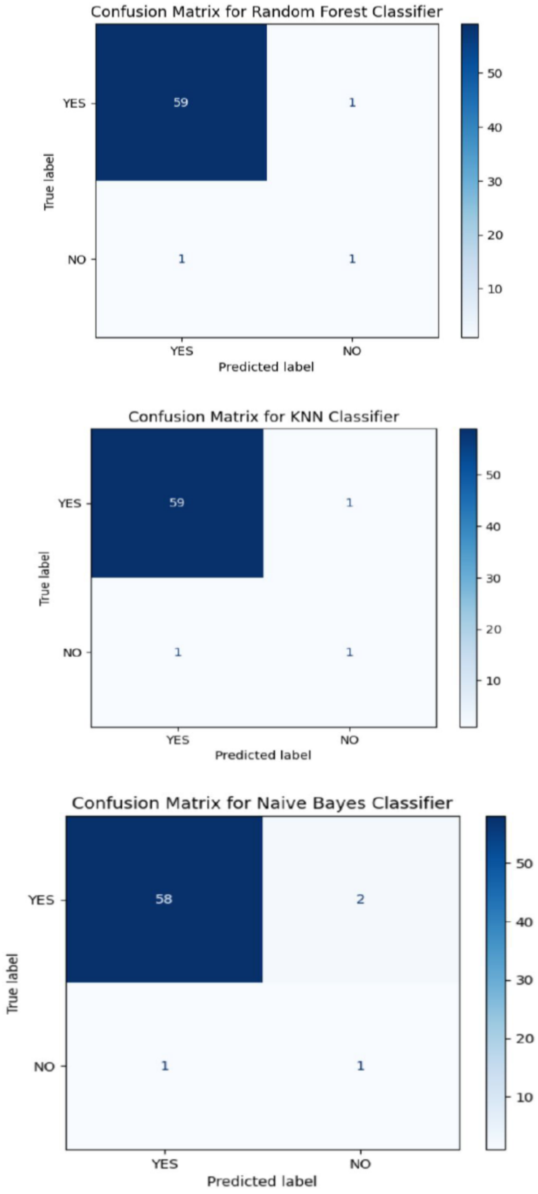

Fig: 3 Confusion matrices for Machine learning

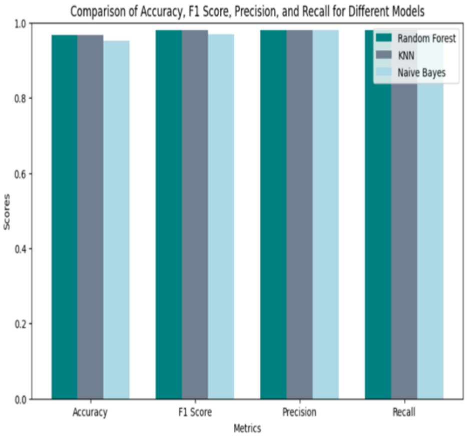

Fig: 4 Accuracy graphs of Machine Learning Techniques for Lung cancer

6. Conclusion:

By identifying key determinants of lung cancer incidence, our findings contribute to the growing body of knowledge aimed at improving early detection, prevention, and personalized treatment strategies for this devastating disease. Furthermore advancements in genetic profiling, biomarkers, and advanced imaging techniques are enhancing diagnostic precision to individual patient needs. As continued interdisciplinary research and collaboration are essential to developing innovative treatment approaches and improving survival rates. Finally this extensive procedure focused to offer more productive and personalized care, minimizing the burden of lung cancer on individuals and health care systems alike.

Conflicts of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization, Methodology, Software, Field Data curation, Writing-Original draft preparation, Software, Validation., Field study

References

[1] Patel, D., Shah, Y., Thakkar, N., Shah, K., Shah, M.: Implementation of artificial intelligence techniques for cancer detection. Augmented Human Res. 5(1), 6 (2020)

[2] Bhatia S., Sinha Y., Goel L. *Soft Computing for Problem Solving.* Singapore: Springer; 2019. Lung cancer detection: a deep learning approach; pp. 699–705.

[3] Nikita Banerjee, Subhalaxmi Das “Machine Learning Techniques for Prediction of Lung Cancer” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020.

[4] Liu C., Hu S. C., Wang C., Lafata K., Yin F. F. Automatic detection of pulmonary nodules on CT images with YOLOv3: development and evaluation using simulated and patient data. *Quantitative Imaging in Medicine and Surgery*. 2020;10(10):1917–1929.

[5] Chaudhury S., Krishna A. N., Gupta S., et al. Effective image processing and segmentation-based machine learning techniques for diagnosis of breast cancer.

[6] *Computational and Mathematical Methods in Medicine*. 2022; 2022:6.

[7] Siegel R. L., Miller K. D., and Jemal A., Cancer statistics, 2019, *Cancer Journal for Clinicians*. (2019) **69**, no. 1, 7–34

[8] Halder A., Kumar A. Active learning using Fuzzy-Rough Nearest Neighbor classifier for cancer prediction from microarray gene expression data. *Journal of Biomedical Informatics*. 2020;34(1):p. 2057001.

[9] Pradeep, K., Nuveen, and N.: Lung cancer survivability prediction based on performance using classification techniques of support vector machines, c4. 5 and naive bayes algorithms for healthcare analytics. Procedia computer science 132, 412–420 (2018)

[10] Kadir,T., Gleeson, F.Lung “Lung cancer prediction using machine learning and advanced imaging techniques” Res.2018 Jun;7(3):304-312.

[11] B. Harangi, ``Skin lesion classification with ensembles of deep convolutional neural networks,'' *J. Biomed. Informat.*, vol. 86, pp. 25_32, Oct. 2018.

[12] Chip M. Lynch, Behnaz Abdollahi, Joshua D. Fuqua and Alexandra R. deCarlo, “Prediction of lung cancer patient survival via supervised machine learning classification techniques”, PMC,

[13] Lynch, Chip M., et al. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." International journal of medical informatics 108 (2017): 1-8. [13] S. Taheri and M. Mammadov, ``Learning the naive Bayes classifier with optimization models,'' *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, pp. 787_795, Dec. 2013.