Lung cancer Prediction and Classification Using Machine Learning Techniques

Abstract

In every nation, lung cancer is one of the major causes of death for both men and women. Because of its poor prognosis, lung cancer has a high mortality rate. With the use of picture recognition and data analytics, the computing sector is entirely automating it, and the medical sector is doing the same. In order to detect lung cancer at an early stage and perhaps save many lives, this research aims to examine the accuracy ratio of two classifiers, Random Forest (RF) and Naive Bayes (NB). Basically, the informational indexes utilized as a part of this examination are taken from UCI datasets for patients affected by lung cancer. This paper's main focus is on the execution analysis of the classification algorithm's accuracy. The experimental results show that RF gives the best result with 97% and NB with 94%.

Introduction

When compared to other cancers, lung cancer is the second most frequent, affecting one in five men and one in nine women. Unfortunately, during the past few years, while lung cancer incidence has slowly decreased in males, it has been rapidly increasing in women. In 1940, just seven women out of every 100,000 acquired the illness; now, that number is 42. The cause of smoking is smoking, according to all the data. The number of cigarettes you smoke each day affects how long it takes to get cancer, according to one expert in the subject. However, research shows that giving up smoking does reduce the risk.

Small cell lung cancer (SCLC), sometimes known as oat cell cancer due to the cells' resemblance to oat grains, and non-small cell lung cancer are the two main kinds of lung cancer (NSCLC). The type of tumor found determines the disease's severity and available treatments. The importance of early discovery and timely treatment—typically surgery to remove the tumor—is underscored by the fact that many kinds of lung cancer develop and spread quickly and that the lungs are essential organs.

The prevalence of illnesses including cholera, chikungunya, and cancer is rising along with the population's pace of fast growth. Cancer is increasingly the leading cause of mortality for all of them. Since the human body contains billions of cells, cancer may develop practically everywhere. Human cells typically develop and divide to create new cells as needed by the body. New cells replace old ones when they die as a result of aging or injury. However, this systematic mechanism disintegrates when cancer cells grow. Old or damaged cells survive when they should die and new cells emerge when they are not required as cells become more and more abnormal. These excess cells might continue to develop indefinitely and can result in growths known as tumor.

There are two sorts of tumors: benign and malignant. A benign tumor is a growth of cells that cannot travel to other parts of the body, but a malignant tumor may. This process of an infection spreading to other parts of the body is known as metastasis. There are many different types of cancer, including colon, lung, and leukemia.

Since the early 19th century, lung cancer incidence has considerably grown. Lung cancer can be caused by a number of factors, including smoking, asbestos exposure, radon gas exposure, and passive smoking. Compared to SCLC, non-small cell lung cancer is more prevalent and tends to progress and spread more

slowly. Nearly identical to smoking, SCLC develops larger tumors more quickly and has the potential to spread broadly across the body. They frequently begin in the bronchi around the chest's center. The overall number of cigarettes smoked has an impact on the death rate from lung cancer. [1] Signs and symptoms of lung cancer include:

- dyspnea (shortness of breath with activity),
- hemoptysis (coughing up blood),
- chronic coughing or change in regular coughing pattern,
- wheezing,
- chest pain or pain in the abdomen,
- cachexia (weight loss, fatigue, and loss of appetite),
- dysphonia (hoarse voice),
- clubbing of the fingernails(uncommon),
- dysphasia(difficulty swallowing),
- Pain in shoulder ,chest , arm,
- Bronchitis or pneumonia,
- Decline in Health and unexplained weight loss.

Tobacco usage is a major cause of mortality and morbidity. Typically, lung cancer grows within the bronchial tree's wall or epithelium. However, any region of the respiratory system can be impacted and it can begin anywhere in the lungs.

The majority of lung cancer patients are between the ages of 55 and 65, and the disease frequently takes many years to manifest [2].

There are two primary forms of lung cancer. They are also known as oat cell cancer and small cell lung cancer, respectively. Every form of lung cancer is treated differently and develops and spreads in a unique way. Mixed small cell/large cell cancer is a term used to describe a cancer that possesses traits from both categories.

Lung cancer symptoms:

The following are the generic lung cancer symptoms [3].

i. A cough that does not go away and gets worse over time
ii. Coughing up blood (heamoptysis) or bloody mucus.
iii. Chest, shoulder, or back pain that doesn't go away and often is made worse by deep Hoarseness
iv. Weight loss and loss of appetite
v. Increase in volume of sputum
vi. Wheezing
vii. Shortness of breath

viii. Repeated respiratory infections, such as bronchitis or pneumonia
ix. Ix. Repeated problems with pneumonia or bronchitis
x. x. Fatigue and weakness
xi. xi. New onset of wheezing
xii. xii. Swelling of the neck and face
xiii. xiii. Clubbing of the fingers and toes. The nails appear to bulge out more than normal.
xiv. Xiv.Paraneoplastic syndromes which are caused by biologically active substances that are secreted by the tumor.
xv. xv. Fever xvi. Hoarseness of voice
xvi. Xvii.Puffiness of face
xvii. Xviii.Loss of appetite
xviii. Xix.Nausea and vomiting

Lung cancer risk factors: (Figure – 1**)**

a. Smoking:

i. Beedi

ii. Cigarette

iii. Hukka

b. Second-hand smoke

c. High dose of ionizing radiation

d. Radon exposure

e. Occupational exposure to mustard gas chloromethyl ether, inorganic arsenic, chromium, nickel, vinyl chloride, radon asbestos

f. Air pollution

g. Insufficient consumption of fruits & vegetables

h. Suffering with other types of malignancy.

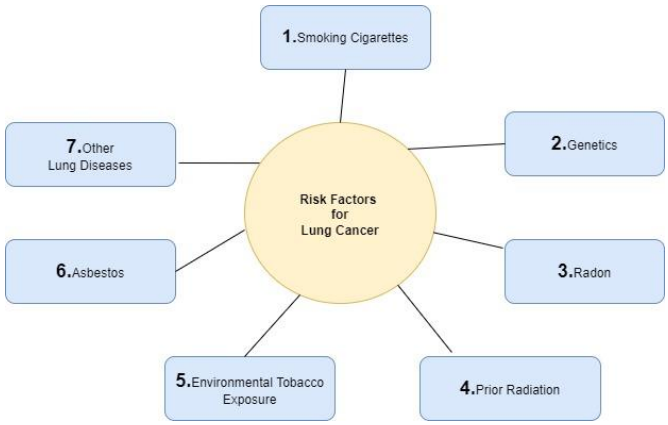

Figure – 1: Risk factors for Lung Cancer

Related work

Krishnaiah, V., G. Narsimha et al. [1] proposed a method to predict Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. In this study, Naive Bayes, ODANB, NCC2, Data Mining, Classification applied for train and test this system.

Yongqian Qiang et al. [2] developed a system to predict the Diagnostic Rules of Peripheral Lung Cancer Preliminary Study Based on Data Mining Technique.

Murat Karabhatak et al. [3] invented a system to create an expert system for detection of breast cancer based on association rules and neural network.

Y. Xie *et al. [4] developed a system to detect* Early lung cancer diagnostic biomarker discovery using machine learning. In this study, they collected total 110 lung cancer patients and 43 healthy individuals data to train and test their method. In this study, six machine learning techniques of K-nearest neighbor (KNN), Naïve Bayes, AdaBoost, Support Vector Machine (SVM), Random Forest, and Neural Network with 10-cross fold technique were used for the early lung tumor prediction based on the metabolomic biomarkers features.

D. M. Abdullah et al. [7] *proposed a system to detect* Lung cancer Prediction and Classification based on Correlation Selection method using machine learning techniques. They used UCI dataset to train and test their method. In this study, using SVM authors obtained 95.56% accuracy. While using KNN and CNN they obtained 88.40% and 92.11% accuracy, respectively.

J. Pati et al. [10] invented a system to detect Gene Expression Analysis for Early Lung Cancer Prediction using Machine Learning techniques and an Eco-Genomics approach. They used Kent Ridge Bio-Medical Dataset Repository. In this study, using SMO authors obtained 91.6667% accuracy. While using Multi-Layer Perceptron and Random Sub Space they obtained 86.6667% and 68.3333% accuracy, respectively.

C. H. Huang *et al. [4] proposed a system to detect Lung Cancer* Using a Chemical Sensor Array and a Machine Learning Technique. They used lung cancer cases and non-tumour controls between 2016 and 2018 for train and test their method. In this study, using SVM internal validation authors obtained 92.7% accuracy. While using LDA internal validation, LDA external validation and SVM external validation they obtained 90.2%, 85.4% and 85.4% accuracy, respectively.

Methodology

Proposed method:

In this study, there have used different type of machine learning algorithms. In this time, Random Forest (RF) provides best result in this study. That is 97%.

Data set:

Survey Lung Cancer is a familiar and commonly used data set for the prediction of Lung Cancer. (survey lung cancer | Kaggle). This data set consists of 310 rows and 16 columns. The attributes included in the column are gender, age, smocking, yellow_fingers, anxiety, peer_pressure, chronic disease, fatigue, allergy, wheezing, alcohol, coughing, shortness, shallowing, chest pain, lung_cancer.

Data Visualization:

Data visualization helps to understand the data better by putting it in a visual form. In this phase, data are represented in the form of bar chart.

References

[1] Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." International Journal of Computer Science and Information Technologies 4.1 (2013): 39-45.

[2] Yongqian Qiang, Youmin Guo, Xue Li, Qiuping Wang, Hao Chen, & Duwu Cuic 2007 .The Diagnostic Rules of Peripheral Lung cancer Preliminary study based on Data Mining Technique. Journal of Nanjing Medical University, 21(3):190-195

[3] Murat Karabhatak, M.Cevdet Ince 2008. Expert system for detection of breast cancer based on association rules and neural network. Journal: Expert systems with Applications.

[4] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, Jan. 2021, doi: 10.1016/j.tranon.2020.100907.

[5] G. A. P. Singh and P. K. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6863–6877, Oct. 2019, doi: 10.1007/s00521-018-3518-x.

[6] Dayananda Sagar College of Engineering, Institute of Electrical and Electronics Engineers. Bangalore Section, and Institute of Electrical and Electronics Engineers, *2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2020) : conference proceedings : 5-7 March, 2020*.

[7] C. H. Huang *et al.*, "A study of diagnostic accuracy using a chemical sensor array and a machine learning technique to detect lung cancer," *Sensors (Switzerland)*, vol. 18, no. 9, Sep. 2018, doi: 10.3390/s18092845.

[8] N. Banerjee, "Prediction Lung Cancer-In Machine Learning Perspective," 2020.

[9] M. Imran Faisal, S. Bashir, Z. Sikandar Khan, and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer."

[10] D. M. Abdullah, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques", doi: 10.48161/Issn.2709-8206.

[11] "AN EXTENSIVE REVIEW ON LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUES," *Journal of critical reviews*, vol. 7, no. 14, Jul. 2020, doi: 10.31838/jcr.07.14.68.

[12] Q. Gu *et al.*, "Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer," *European Journal of Radiology*, vol. 118, pp. 32–37, Sep. 2019, doi: 10.1016/j.ejrad.2019.06.025.

[13] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2019, doi: 10.1109/ACCESS.2018.2886604.

[14] J. M. Luna *et al.*, "Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning," *Radiotherapy and Oncology*, vol. 133, pp. 106–112, Apr. 2019, doi: 10.1016/j.radonc.2019.01.003.