e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:05/May-2023 Impact Factor- 7.868 www.irjmets.com

LUNG CANCER PREDICTION USING MACHINE LEARNING

Amrutha K Shaji*[1], Anjali S*[2], Sanjay J*[3], Prof. Blessy Rapheal*[4]

*[1,2,3]Student, Department Of Electronics And Communication Engineering, RIST,
Palakkad, Kerala, India.

*[4]Professor, Department Of Electronics And Communication Engineering
RIST, Palakkad,Kerala, India.

DOI : https://www.doi.org/10.56726/IRJMETS37797

ABSTRACT

Lung cancer is a widespread disease that poses a significant challenge for radiologists to diagnose accurately. The early detection of lung cancer symptoms makes it possible to treat the disease effectively. With the latest advancements in computational intelligence, it is feasible to create a sustainable prototype model for treating lung cancer. However, researchers have been utilizing machine learning algorithms to develop smart computer-aided systems that can assist radiologists in making more precise diagnoses. The dataset is trained using various techniques, including SVC, K-Nearest Neighbour, Decision Tree, Logistic Regression, XgBoost, Gradient boosting and Random Forest. These models are implemented using python programming.

Keywords: Lung Cancer,Logistic Regression, Classification.

I. INTRODUCTION

In the past few years the Lung Cancer became the major health diseases in human body. It is quite difficult to diagnose Lung Cancer in early stage of it, which may leads to increase the risk factor of survival of patients. Correspondingly the treatment on Lung cancer depends upon the how early this disease can be diagnose so that treatment can control on increasing (in stage) and spreading of Lung Cancer in other part of body. It is quite possible to control Lung Cancer disease by giving proper treatment, there are various treatments are available in the field of Medical Science such as Surgery, Chemotherapy and radiography as it is depend on the stage of disease, health of patient, and some other factors. The rate of survival is only 14 % of patient for five years. Lung Cancer develops in respiratory epithelium of bronchial tree of lungs. Its death rate scores among all cancer deaths, and is also the top killer towards male and female cancer death. There have been nearly 1.8 million fresh cases of lung cancer annually (13 percent among all cancers), 1.6 million deaths worldwide (19.4 percent among all cancers). Lung cancer is a proliferation of expanding and developing irregular cells into a cancer.Of the other forms of cancer, the death rate of lung cancer is the greatest. Cigarette smoke induces an approximate 85 percent of cases of lung cancer in males and 75 percent in females. Lung cancer is amongst the most terrible illnesses in the developing countries, with a death rate of 19.4 percent. Lung cancer is among the most dangerous cancer worldwide, with lowest success rate following diagnosis, with a steady rise in casualty count per year. Advantages of Fuzzy logic in the earlier predictions will lead to result oriented analysis. Survival of lung cancer as a result of diagnosis is directly related to its progress. Yet individuals have a greater success rate it will be found in the early stages of life. Cancer cells are distributed in blood from the lungs, the lymph fluid that covers the lung tissues. The lymph passes into lymph vessels that discharge through lymph nodes in the lungs and chest region. Examination and treatment of lung disease has become one of the biggest obstacles that humanity faces in recent years. Early tumor diagnosis will reliably promote its survival of vast numbers of life around the world. It is very rare to detect the lung cancer before the age 45 years but generally Lung Cancer may be detected in the age 55 to 70.

In our project, we trained 7 various algorithms[2]and compared them all.Results from several models are evaluated.Apply algorithms to a dataset and compare the results of each method once we have determined which is optimal for our problem. We can then concentrate on those algorithms and alter their parameters to better the outcome. This study helps readers decide which model to use based on accuracy by estimating models using the algorithm with the highest accuracy.

II. METHODOLOGY

The Lung cancer prediction methodology section details the system's development and implementation at various phases. This section aims to explain in detail the Data framing, Data analysis, Model prepartion, Model training and verification of output.

Data collection -Data set used in the Lung cancer prediction comprises of 391 cases downloaded from Kaggle . It consists of 15 variables (Age, Gender, Anxiety, Coughing, Smoking, Wheezing, Yellow fingers, Peer pressure, Chronic dieases, Fatigue, Alcohol consuming, Shortness of breathe, Allergy, Swallowing difficulty, Chest pain) and 3 severity classes(low, average, critical).

Data framing-Data framing in python involves organizing and Manipulating data in atabular format using a data structure called data frame.Dataframes are a crucial component popluar data manipulation libraries like pandas,which are extensively utilized for tasks related to data analysis and manipulation in python.the basic concepts of data framing in python using pandas are importing libraries,creating data frame, inspecting data, data manipulation,data cleaning and data visualization

Data Analysis-Python includes a variety of modules and tools for data analysis, including statistical methodologies, machine learning algorithms, and other approaches. You may create models that let you forecast the future and obtain insightful knowledge from your data by utilising well-known frameworks like TensorFlow and Scikit-learn.

Model Preparation-In order to prepare a model for lung cancer prediction, data must be gathered, preprocessed, relevant characteristics selected, a suitable machine learning method selected, the model trained, and the model's performance assessed. Patients with and without lung cancer should be represented in the data used to train the model. Select and identify the characteristics that are most important for predicting lung cancer. After analyzing various model with same parameter, . In the analysis of epidemiological datasets, particularly in the field of machine learning, the widely used mathematical modelling technique known as logistic regression (LR) is applied.Logistic regression gives more accurate prediction.

Evaluation-An web application that uses machine learning for lung cancer prediction. With 391 cases, it has been trained. The performance of the model was more precise. With the aid of such a model, it will be able to predict whether the person have lung cancer or not.

III. MODELING AND ANALYSIS

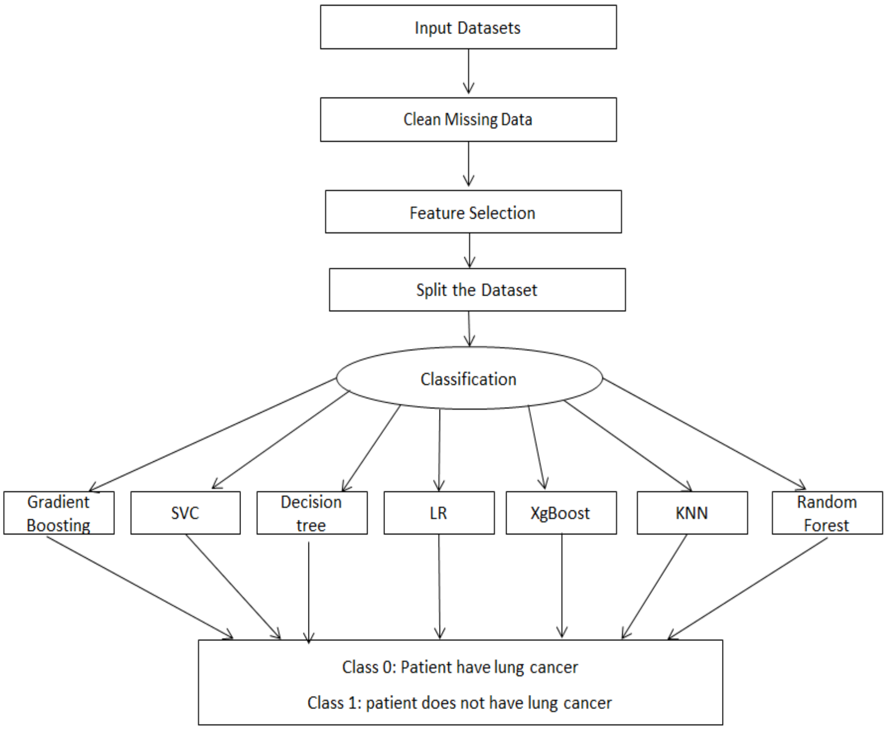
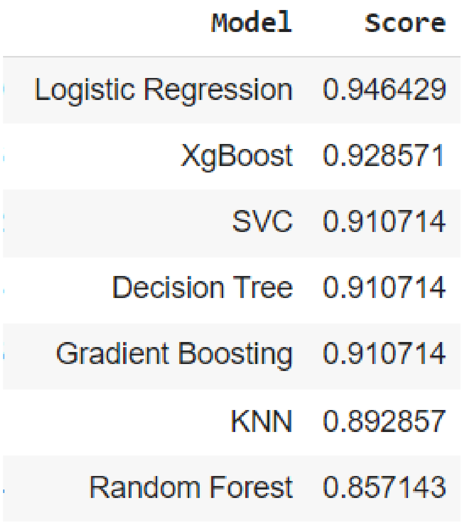


Figure 1: Flow chart for lung cancer prediction

Lung cancer detection is crucial for taking preventative action. To achieve this, a platform or web application can be developed to collect user input and predict the likelihood of a person having lung cancer. The first step is to capture the dataset and clean it using data cleaning techniques to address missing values. Next, feature selection is applied to the normalized dataset, which is then divided into two segments: 80% train data and 20% test data.



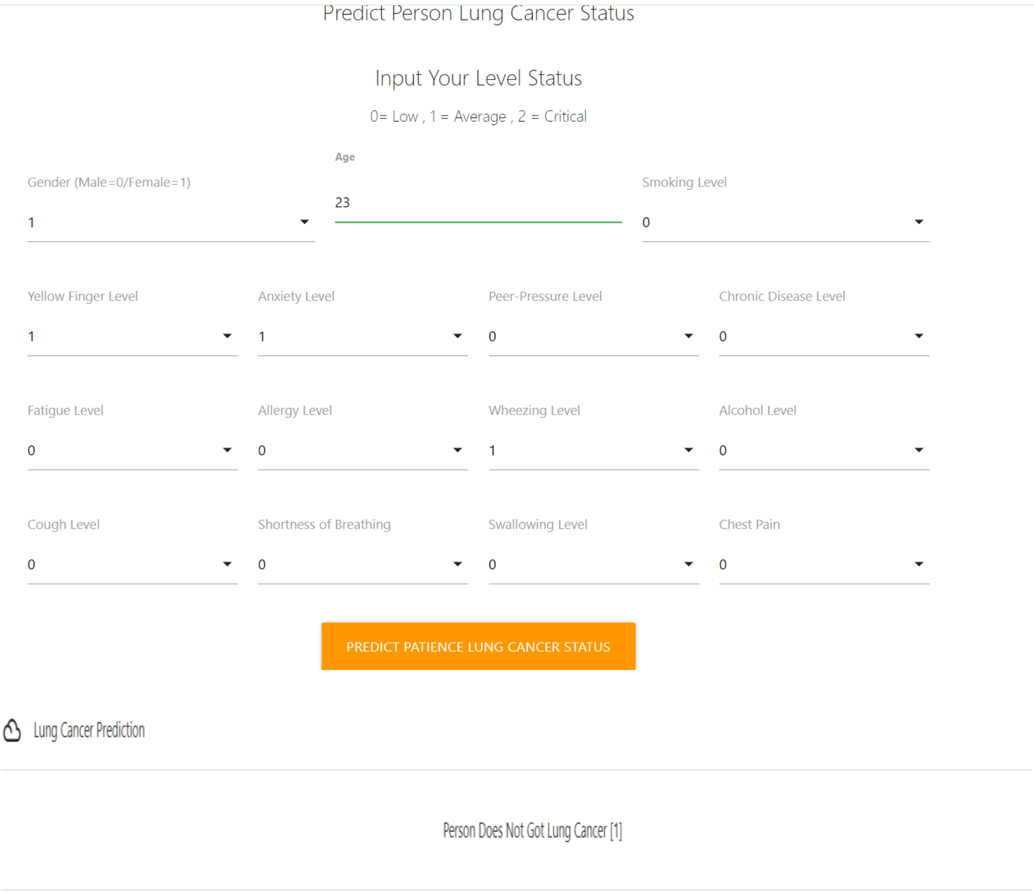
Model	Score
Logistic Regression	0.946429
XgBoost	0.928571
SVC	0.910714
Decision Tree	0.910714
Gradient Boosting	0.910714
KNN	0.892857
Random Forest	0.857143

Figure 2: Accuracy of various models

Various classification algorithms are applied to these segments, and the results show that logistic regression provides the highest accuracy of 0.94 compared to other methods such as Xgboost (0.92), SVC (0.91), decision tree (0.91), gradient boosting (0.91), KNN (0.89), and random forest (0.85). Further improvements can be made by utilizing the logistic regression classifier, which has the highest accuracy among the tested algorithms. The system heavily depends on modeling and analysis to identify the main causes of lung cancer and predict its occurrence, which in turn helps individuals to take necessary safety precautions. This is achieved through the utilization of machine learning techniques to model and analyze the symptoms of individuals.

IV. RESULTS AND DISCUSSION

A web application has been created using Flask, where users can input data for predicting whether they have cancer or not. The machine learning model used in this application is logistic regression, which has demonstrated the highest accuracy. Based on the input data provided by the user, the model will predict the likelihood of the person having cancer and assign a severity scale of either 0 or 1.Figure 2 and 3 shows the homepage of our web application, which displays various user-inputted parameters such as age, smoking habits, peer pressure, and wheezing. Once the user fills in these parameters, they can click on the Predict button, which transfers the data to the backend. Here, our chosen machine learning algorithm, logistic regression, processes the input data. Based on the analysis, the output will indicate whether the patient is diagnosed with lung cancer or not.



Predict Person Lung Cancer Status

Input Your Level Status

0= Low , 1 = Average , 2 = Critical

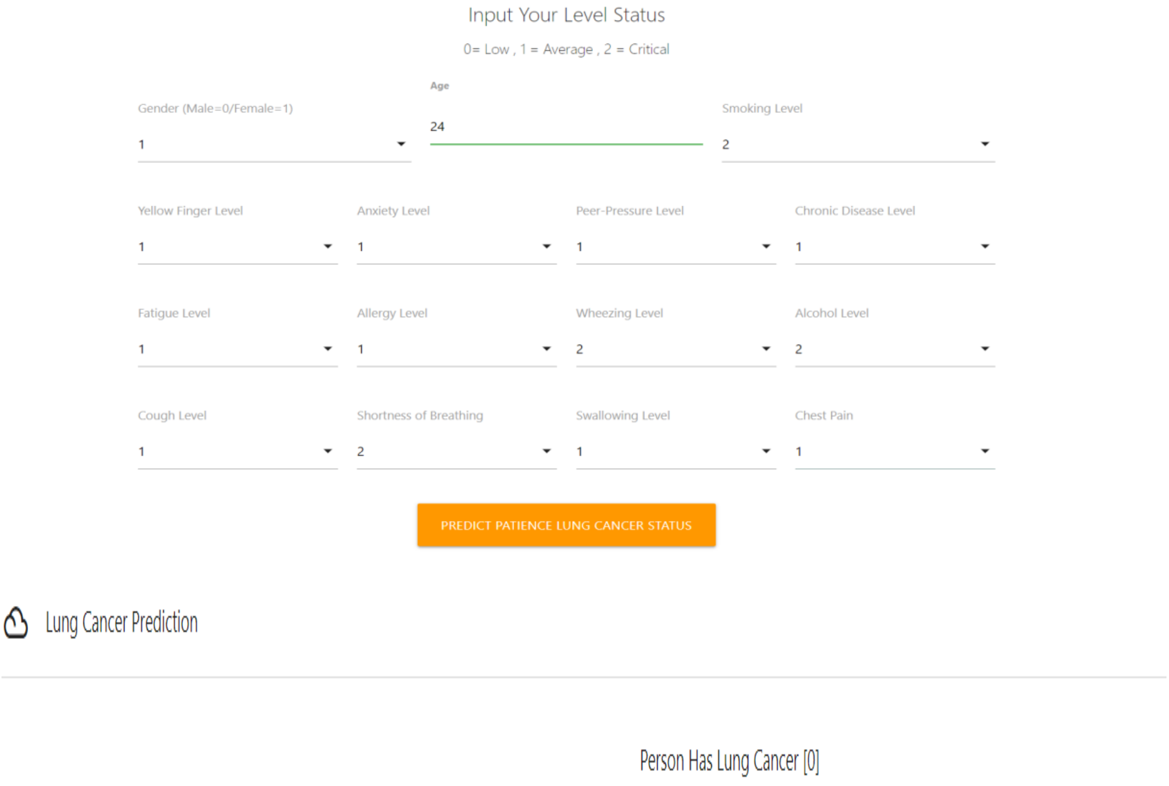
Gender (Male=0/Female=1)	Age	Smoking Level
1	23	0

Yellow Finger Level	Anxiety Level	Peer-Pressure Level	Chronic Disease Level
1	1	0	0
Fatigue Level	Allergy Level	Wheezing Level	Alcohol Level
0	0	1	0
Cough Level	Shortness of Breathing	Swallowing Level	Chest Pain
0	0	0	0

Lung Cancer Prediction

Person Does Not Got Lung Cancer [1]

Figure 2: prediction status of a person 1



Input Your Level Status

0= Low , 1 = Average , 2 = Critical

Gender (Male=0/Female=1)	Age	Smoking Level
1	24	2

Yellow Finger Level	Anxiety Level	Peer-Pressure Level	Chronic Disease Level
1	1	1	1
Fatigue Level	Allergy Level	Wheezing Level	Alcohol Level
1	1	2	2
Cough Level	Shortness of Breathing	Swallowing Level	Chest Pain
1	2	1	1

Lung Cancer Prediction

Person Has Lung Cancer [0]

Figure 3: prediction status of person 2

V. CONCLUSION

Lung cancer is one of the primary and frequent causes of cancer deaths worldwide, both in terms of incidence and short-term effects. The recent increase in disease detection and shortcomings in efficient treatment are the primary causes of the rise in mortality from it. Therefore, early detection is essential to prevent this disease from taking lives. Modern machine learning techniques, such as logistic regression, can be used to predict the lung cancer survivability rate. The proposed approach can identify lung cancer in its early stages, increasing patient survival rates. Early detection is crucial in improving patient outcomes and reducing mortality rates. To predict the survival rate of lung cancer, advanced machine learning techniques such as logistic regression can be utilized. This system can accurately predict lung cancer in its early stages, which is critical in improving patient survival rates.In our model the prediction system predicts that 0 indicates the person have lung cancer and 1 indicates that person doesnot have lung cancer.

ACKNOWLEDGEMENTS

The Lung cancer prediction System would not have been possible without the assistance and direction of a number of individuals and institutions. We would like to thank the management of Rajadhani Institute of Science and Technology, Palakkad, for providing us with the resources and facilities needed to complete this project.We would like to express our heartfelt gratitude to Dr.Ramani K, the institution's Principal, for her unwavering support and encouragement throughout the project. We are particularly grateful to Ms.Sithara krshnan, Head of the Department, for her invaluable guidance and assistance in the project's implementation. Our sincere thanks go to Ms. Blessy Rapheal, our project guide, for her constant support and guidance in the development of the lung cancer prediction system. Her expertise in the field of machine learning was invaluable in the successful completion of this project. Finally, we would like to express our gratitude to all those who have helped us in any way to complete this project successfully. Their support and encouragement have been invaluable in the development of the lung cancer prediction system.

VI. REFERENCES

[1] Raghavendra, Patil G E., Sinchana, C G., Tejashwini, P., et al. 2020. Lung Cancer Prediction System Using Logistic Regression Approach. International Research Journal of Modernization in Engineering Technology and Science, 656-660

[2] Kasthuri, M., & Jency, M. R. (2020). Lung Cancer Prediction Using Machine Learning Algorithms on Big Data: Survey. International Journal of Computer Science and Mobile Computing, 9(10), 73-77.

[3] Hazra, A., Bera, N., & Mandal, A. (2017). Predicting lung cancer survivability using SVM and Logistic Regression Algorithms. International Journal of Computer Applications, 174(2), 19-24.

[4] Staceyinrobert. 2017. Survey Lung Cancer. [Online] Available at:
https://data.world/sta427ceyin/survey-lung-cancer

[5] Nurlaila Dwi, Dadan Kusnandar, Evy Sulistianingsih. 2013. Perbandingan Metode Maximum Likelihood Estimation (MLE) dan Metode Bayes dalam Pendugaan Parameter Distribusi Eksponensial. [Online] Available at: https://core.ac.uk/download/pdf/326807809.pdf