WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Lung Cancer Prediction Using Machine Learning Algorithm by Clinical Data

Gurugubelli Hari Narayana[1*], T.V.S.Divakar[2]

[1,2]Department of Electronics and Communication Engineering, GMR Institute of Technology, Rajam, 532127, Andhra Pradesh, India.*

A B S T R A C T

One of the most controllable risk factors for cardiovascular disease (CVD) is smoking, which has a substantial and intricate effect on cardiovascular health (CVH). Early diagnosis and intervention are essential because cardiovascular disease CVD remains the world's leading cause of illness and death. In this endeavour, predicting smoking behaviour is crucial.This study explores the complex relationship between smoking and CVD emphasizing the detrimental effects of smoking on cardiovascular health. MissingNaN values in the bio-signal data pertaining to smoking behaviour are eliminated in the first stage of analysis because they only appear in the target columnsmokingand may otherwise introduce bias into the predictions. The dataset is then standardized using robust scaling which makes sure that every feature has the same scale. Lastlythe most pertinent variables for precisely predicting smoking status are found using the Random Forest feature selection method.By assisting physicians and researchers in early lung cancer detection risk factor identification, and more accurate patient outcome prediction than conventional techniques.Machine learning plays a major role in lung cancer prediction. It accomplishes this by identifying patterns in vast amounts of medical data including genetic information, pictures and patient medical records and then applying those patterns to classifications or predictions.

Keywords: *Lung cancer,Machine learning,Artificial Neural Network, Algorithm*

1. Objectives

- Early detection of possible lung cancer cases increases the likelihood of a successful course of treatment and survival before outward symptoms show up.
- To make wellinformed predictions, use patient data such as CT scan images, genetic information, smoking habits, age, gender, exposure to air pollution, etc.
- Give algorithms the ability to learn from sizable datasets and identify intricate connections between risk factors such as medical history, smoking and air quality.
- Support or enhance medical experts diagnosis by providing consistent, data-backed predictions.
- Create models that predict lung cancer with high accuracy, precision, and recall such as SVM, Random Forest, CNN, ANN, etc.

2. Introduction

Nowadays cancer is one of the deadliest illnesses. It can take many different forms including thyroid, kidney, lung, and blood cancers. A class of diseases known as cancer is defined by the unchecked development and dissemination of aberrant cells. Body cells typically undergo controlled growth after divide and death. Changes mutations to a cell's DNA which contains instructions for how the cell should function, are the cause of cancer a genetic disease. Cell division and growth may be impacted by these genetic alterations.

Lung cancer is a type of cancer that starts as an uncontrolled growth of abnormal cells in the tissues of the lungs often in the cells that line the air passages and it is also known as lung carcinoma. Lung Cancer is serious health issue and it is the leading cause of cancer deaths in worldwide.

There are mainly two types of lung cancer they are

- NSCLC
- SCLC

NSCLC: The full form of NSCLC is Non-Small Cell Lung Cancer. It is the most common type of lung cancer. It accounts for over 80% of lung cancer cases. Common types include adenocarcinoma and squamous cell carcinoma and sarcomatoid carcinoma are two less common types of NSCLC.

SCLC: The full form of SCLC is Small Cell Lung Cancer. It grows more quickly and is harder than NSCLC. It is often found as a relatively small lung tumor that is already spreads to other parts of your body. Specific types if SCLC include small cell carcinoma it is also called oat cell carcinoma and combined small cell carcinoma.

Nowadays medical applications heavily rely on machine learning. It offers insights that enhance patient care speed up research and simplify operations by analyzing enormous volumes of data. Applications of machine learning in medicine include the analysis of CT, MR, and X-ray scans. A patient's likelihood of contracting a particular disease how they will probably react to a particular treatment and any possible adverse effects can all be predicted by machine learning models.

Machine learning plays a critical and revolutionary role in earlier cancer prediction. Because early detection greatly improves survival rates and lessens the need for aggressive treatment it is essential.

3. Literature Survey

Smoking poses a major risk to cardiovascular health (CVH) and is one of the primary causes of many diseases and fatalities globally [1]. It entails breathing in and out the smoke produced by burning tobacco, which contains thousands of dangerous chemicals.Actuallythere are almost 4,000 different substances found in tobacco smoke many of which are toxic and seriously harmful to human health [2].

Nicotine, tar, carbon dioxide and carbon monoxide are among the toxic gases found in tobacco smoke. These chemicals are released into the air when tobacco burns harming the respiratory system both immediately and over time as well as the major organsepithelial tissues [3]. Smoking is well known to be a significant risk factor for poor cardiovascular health (CVH) raising the chance of developing diseases like stroke, coronary artery disease, and peripheral arterial disease. Based on variables such as age, blood pressure, cholesterol, diabetes, and smoking habits, the Framingham Risk Score (FRS) evaluates the risk of cardiovascular disease (CVD) [4].

Even though the risks associated with smoking are well known it is still difficult to predict a person's smoking behaviour. Promising techniques for locating hidden smokers and comprehending their effects on CVH are provided by recent developments in Deep Learning (DL) and predictive analytics. According to studies, heavy smokers who stop smoking have a much lower CVD risk score than those who keep smoking [5].

The purpose of this study is to investigate whether DL models can forecast smoking behaviour in concealed smokers and evaluate the effect they have on CVH. Personalized treatment plans, prevention tactics and early diagnosis can all be improved with accurate prediction. The significance of incorporating machine learning models in smokingrelated CVD risk prediction has been emphasized by earlier research such as that conducted by D'Agostino et al. (2013) Wang et al. (2023) and Martinez et al. (2019) [6–8].

4. Methodology

A.SVM(Support Vector Machine):

Support Vector Machines, or SVMs for short, are supervised machine learning algorithms that are primarily used for classification tasks but are also occasionally used for regression tasks.In an n-dimensional space. SVM looks for the optimal boundary referred to as a hyperplane that divides various classes of data points.A vector such as height, weight is used to represent each data point.The hyperplane that best separates the classes is determined by the SVM algorithm.It accomplishes this by optimizing the margin or the separation between the closest data points referred to as support vectors and the hyperplane.

B. Random Forest Classifier:

The supervised machine learning algorithm Random Forest is primarily utilized for tasks involving regression and classification. Because it is accurate, reliable and simple to use even with complex or messy data.It is one of the most widely used and potent algorithms. In essence Random Forest is a forest or collection of numerous decision trees. Each tree makes its own prediction and then the forest combines all those predictions to make a final decision usually by voting for classification or averaging for regression. Thusit's referred to as random since it adds unpredictability to

i. Choosing random data subsets (rows)

ii. Selecting arbitrary feature (column) subsets for every tree

C. K-Nearest Neighbors (KNN):

Although it is most frequently used for classification tasksthe supervised learning algorithm KNN (K-Nearest Neighbors) can also be used for regression.The classification of a data point is determined by the classification of its neighbors.

Select the number of neighbors (K) which is typically a small positive number such as 3, 5 or 7.Calculate the distanceusually Euclidean distancebetween the new data point and every other point in the training dataset.Choose the K nearest neighbors.

D. Artificial Neural Network (ANN):

An Artificial Neural Network (ANN) is a machine learning model inspired by the structure and working of the human brain.It is made up of layers of interconnected neurons are mathematical functionsthat can learn patterns and relationships in data.

It contains three layers:

i. Input layer

ii. Hidden layers

iii. Output layer

E. Decision Tree:

A decision tree is a supervised learning algorithm that can be applied to tasks involving regression and classification. Like a series of yesor no questions it divides data into branches according to feature values until it reaches a final decision a leaf node.

F. Logistic Regression:

Despite its name, Logistic Regression is a classification algorithm rather than a regression technique.

It is employed to forecast categorical results including:

Ex: Yes/No,0 /1

G. Stacking Classifier:

An ensemble model that combines several distinct machine learning algorithms to create a more robust final model is called a stacking classifier short for stacked generalization.

Divide the training set.

i. Train multiple base models e.g. Random Forest, KNN, SVM on the same data.

ii. On the validation set each base model generates predictions.

iii. For the metalearner these predictions become novel features.

iv. They are combined by the metalearner like Logistic Regression.

v. Base models produce outputs during prediction and the metamodel provides the final forecast.

H. Voting Classifier:

In machine learning a voting classifier is an ensemble model that combines several different models such as SVM, Decision Tree, Logistic Regression, etc. To produce a final prediction based on the average for regression or majority votefor classificationof their outputs.It is a component of ensemble learning which uses the advantages of several models to enhance prediction performance.

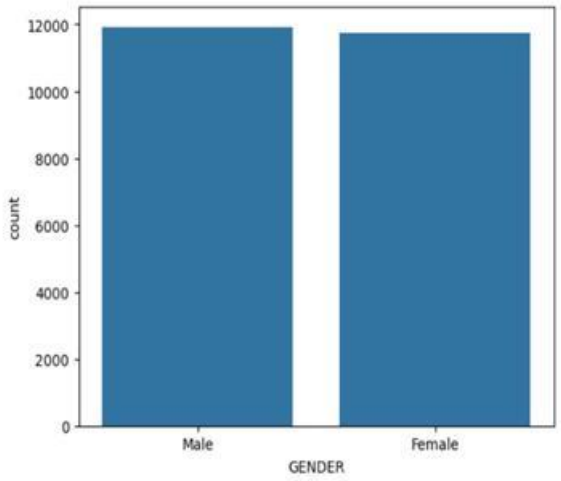
5. Results

(i)Age of the people

In the data set there are 23638 people are there with different age. Here we can see the different aged people we can take the people age in graph in between 30 to 80 years old.

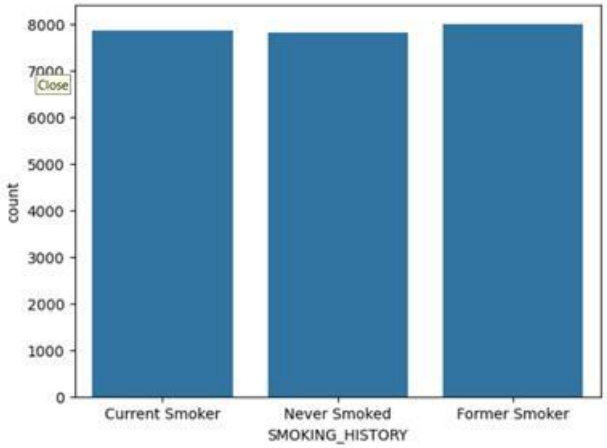
(ii) Gender

In the data set there are 23638 patients are there.



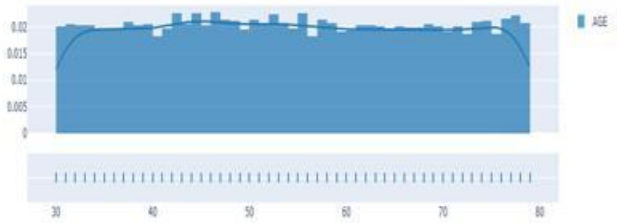
There are nearly 11,500 above male patients are there and above 11,000 female patients are there.

(iii) Smoking _History



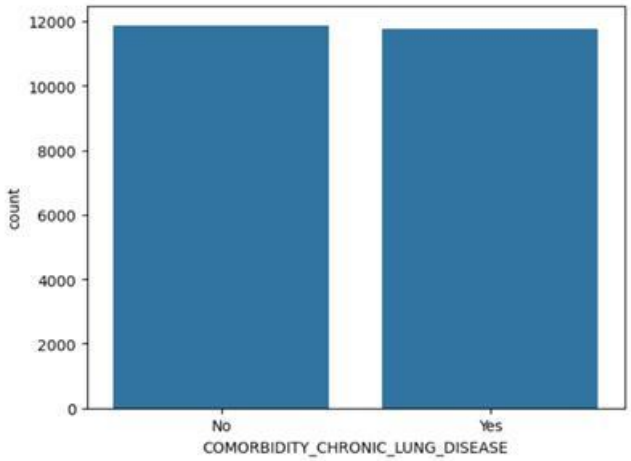
In the process of smoking history test there are nearly 7800 patients are current smokers and never smoked persons are nearly 7500 are there. Thethird one is former smoker are 7900 there. They are previous smoker now they didn't have this habit.

(iv) Stage



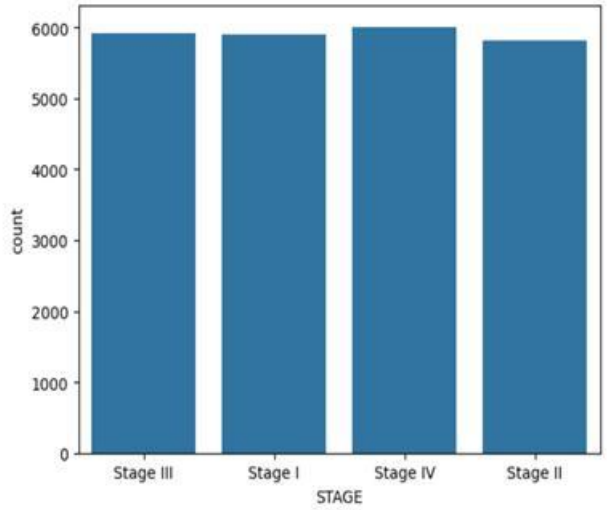
The above data indicates which stage the cancer patient are there. The people who are there in the stage 1 & 2 safe and they can quire with in some months. Who is there in the stage 3 they half of the chances. But people who is there in the stage4 difficult to live and they didn't have chance to survive.

(v) Lung Cancer



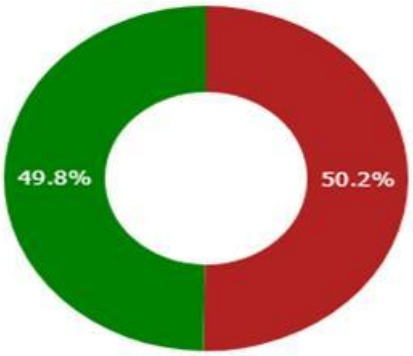
In the give data set the patients who struggling with lung cancer disease there are nearly 11,000 above people are struggling with lung cancer and 11.990 people are doesn't have any lung cancer.

(vi) Analysis of chance of Brain Stroke based on Each Parameter:



a. Pie chart for lung cancer patients

The below pie chart shows the data of the patients who are struggling with lung cancer. It shows split of 50.2%and 49.8% most likely refers to the prevalence of cancer caused by smoking specifically for particular region and green colour indicates yes and red colour indicates no.



6. Conclusion

This study emphasizes how important machine learning is for forecasting lung cancer and associated health hazards. One of the most important and preventable risk factors for lung cancer and cardiovascular disease (CVD) is still smoking. This study shows that by analysing medical and behavioural datait is possible to predict lung cancer accurately and efficiently using machine learning algorithms like SVM, Random Forest, KNN, ANN, Decision

Tree, Logistic Regression, Stacking and Voting Classifiers. The results collectively demonstrate that machine learning models are useful instruments for lung cancer risk assessment, diagnosisand early detectionassisting medical practitioners in reaching quicker and more accurate clinical judgments. The accuracy of early intervention and prediction techniques can be significantly increased by combining medical data with sophisticated computational models. This will ultimately save more lives and lessen the worldwide burden of cancer.

References

1.Y. Zhao, E. P. Wood, N. Mirin, S. H. Cook, and R. Chunara, ‘‘Social determinants in machine learning cardiovascular disease prediction models: A systematic review,’’ *Amer. J. Preventive Med.*, vol. 61, no. 4, pp. 596–605, Oct. 2021

2.C. Wang, H. Zhu, and C. Rao, ‘‘Machine learning-based decision-making mechanism for risk assessment of cardiovascular disease,’’ *Comput. Model. Eng. Sci.*, vol. 138, no. 1, pp. 691–718, 2024

3.M. Hu, R. Benson, A. T. Chen, S.-H. Zhu, and M. Conway, ‘‘Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing,’’ *Drug Alcohol Dependence*, vol. 228, Nov. 2021, Art. no. 109016

4.H.-I. Liu, M.-W. Chen, W.-C. Kao, Y.-W. Yeh, and C.-X. Yang, ‘‘GSAPAhybridGRUandselfattentionbasedmodelfordualmedicalNLPtasks,’’ in *Proc. 14th Int. Conf. Knowl. Smart Technol. (KST)*, Jan. 2022, pp. 80–85

5.A. Karlsson, A. Ellonen, H. Irjala, V. Väliaho, and K. Mattila, “Deep Learning-Based Analysis of Smoking Cessation and Cancer Mortality from Medical Records,” IEEE J. Biomed. Health Inform., vol. 25, no. 9, pp. 3512–3521, Sept. 2021.

6.J. Li, Y. Tao, and T. Cai, “Predicting Lung Cancers Using Epidemiological Data: A Generative-Discriminative Framework,” IEEE/CAA J. Automatica Sinica, vol. 8, no. 5, pp. 1067–1078, May 2021.

7.M. Ammar, N. Javaid, N. Alrajeh, M. Shafiq, and M. Aslam, “A Novel Blending Approach for Smoking Status Prediction in Hidden Smokers to Reduce Cardiovascular Disease Risk,” IEEE Access, vol. 12, pp. 162956–162974, 2024. doi: 10.1109/ACCESS.2024.3480310.

8.S. Doppalapudi, R. G. Qiu, and Y. Badr, “Lung Cancer Survival Period Prediction and Understanding: Deep Learning Approaches,” Int. J. Med. Informat., vol. 148, Apr. 2021, Art. no. 104371.

9.C. Wang, H. Zhu, and C. Rao, “Machine Learning-Based Decision-Making Mechanism for Risk Assessment of Cardiovascular Disease,” Comput. Model. Eng. Sci., vol. 138, no. 1, pp. 691–718, 2024.