American Journal of Electronics & Communication

Shouvik Ghosh et. al., American Journal of Electronics & Communication, Vol. IV (3), 6-10

Lung Cancer Prediction Using Supervised Machine Learning

Shouvik Ghosh[1], Satavisha Sur[2] and Debanjana Ghosh[3]

Department of Electronics and Communication
University of Engineering and Management, Kolkata

[1]ghoshshouvik500@gmail.com
[2]satavishasur7@gmail.com
[3]debanjana.ghosh@uem.edu.in

Abstract: 'Cancer' in today's generation has become a prime factor of concern as because if we compare our past decade with the present, the statistical studies say the death rate has increased by 21% since 1990. So here we have brought up a model which detects lung cancer with high precise accuracy. 'Lung Cancer' are cells that develop in the lungs, due to certain exterior factors. As cancer cells grow rapidly, symptoms are caught at the last stage at maximum situations. So, to put this as our top priority we have presented a Lung Cancer detection model using classification techniques such as (KNN Algorithm, Decision Tree and Confusion Matrix). The ultimate objective of this paper is the early diagnosis of 'Lung Cancer', in order to increase the chances of survival.

Keywords - Lung Cancer, Classification Algorithm, KNN Algorithm, Decision Tree, diagnosis.

I. Introduction

Lung Cancer has taken a topmost position in the rankings of cancer. Too much consumption of Tobacco and Intake of regular Alcohol are some of the common reasons which may lead to lung cancer. The main reason behind this is the 'lack of awareness '. The lack of Knowledge is the major drawback specially in a country called India. The key to improve the survival rate is early detection using Machine Learning Techniques and if we can make the diagnosis process more efficient and effective for radiologists by using this, then it will be a key step towards the goal of improved early detection. Machine Learning takes AI to the next level as it enables intelligent learning to occur within the component based on previous work it did or extrapolations made from data. In the year 2018,

U.S.A brand-new cancer cells instances of 1,682,210 and fatalities of 896,690. As a result of the impact of the Lung Cancer cells or otherwise acknowledgement of lung cancer cells at the very early action the survival price is a lot less than various other cancer cells. The dataset which we used for training our model is taken from Boston UCI Repository ML Lung – Cancer Dataset.

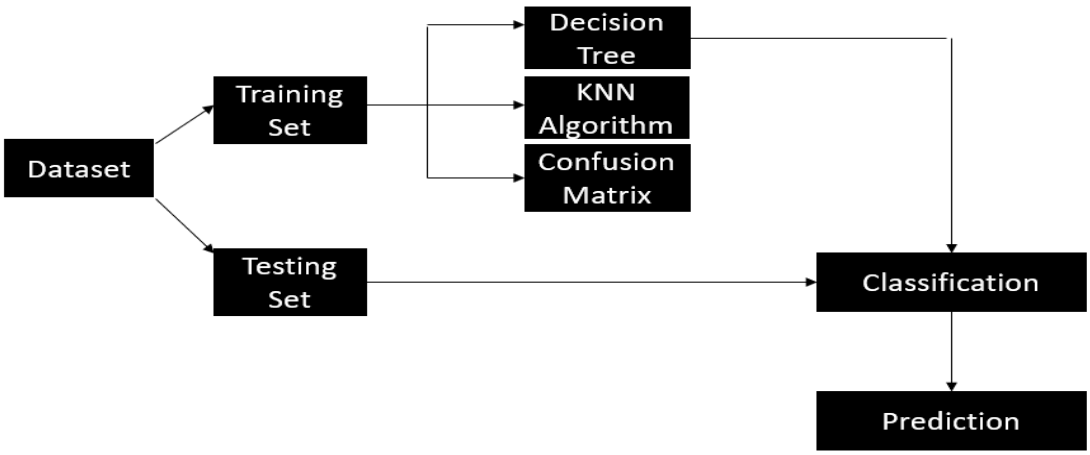
II. Literature Survey

Machine learning helps learning from previous data or work and can be used to train a model to predict diseases such as lung cancer by training the model with sets of previous data about lung cancer. A short and precise description about different research papers on lung cancer detection using ML are given below.

[1] This paper deals with reduction of risk of lung cancer by predicting the cancer early. Classification algorithms such as Naïve Bayes, SVM, Decision tree and Logistic Regression have been used for the prediction of lung cancer. A k-fold cross validation technique has been used for examining these classification algorithms. [2] This paper deals with the development of an ML model for prediction of a disease based on the symptoms of the patient. Decision tree classifier is used in this model to detect a disease by receiving the symptom data of the patient. Additionally, data from the patient's EHR is gathered in order to use NLTK to summarize the prescription and test results. [3] The authors of this paper have designed a system that can efficiently discover the rules to predict the risk of heart disease of the patients based on the parameters given about their health. The performance of the system is then evaluated by using classification accuracy to predict heart disease risk level more accurately. [4] The author of this paper states that the healthcare industry gathers a large amount of heart disease data and discovers hidden information for effective decision making for the prediction of heart diseases. Data mining techniques are used for analyzing data and for identifying relationships. This paper explores the working of various decision tree algorithms in classifying and predicting the disease. [5] This paper deals with the development of an Alzheimer disease prediction model which would be used to assist medical professionals in predicting the status of the disease when the medical data about the patients are provided. This model uses the decision tree algorithm on the sample data that has five important attributes namely gender, age, genetic causes, brain injury and vascular disease. [6] In this paper the author talks about the comparison of the reliability of association rule and decision tree

for disease prediction. The paper says experiments have shown that decision trees can find simple rules. Their reliability is somewhat low and refers to a small number of patients, but by comparison the association is reliable and often refer to larger set of patients. [7] This paper deals with the investigation of the accuracy level of various ML algorithms that deal with the prediction of lung cancer. The accuracy level of different models is evaluated, and the limitations and drawbacks are listed out. The author talks about the need for the development of a better model for prediction because none of the accuracy of the previous models have reached near 100%. [8] This paper deals with the development of a heart disease prediction model by using J48 decision tree for classifying heart diseases based on the clinical features against unpruned, pruned and pruned with reduced error pruning approach, which gives better accuracy. [9] This paper discusses the development of a model that can detect mental health problems in a wide range of patients. The mental health monitoring system will be able to measure stress based on various physical parameters such as heart rate, Spo2, body temperature and pressure, and will also have GPS sensors to track and rescue distressed patients in an emergency.

III. Block Diagram



This is the ultimate block diagram which shows the workflow of our Machine Learning Model.

- Initially, the dataset has been taken along with performing Data-Preprocessing Techniques of removal of outliers and Data-Cleaning.
- The Dataset is further divided into Training and Testing datasets.
- Using the Decision- Tree, KNN Algorithm along with their Confusion Matrix, the training data is fed into the two classifiers and the model is trained.
- The model is therefore evaluated with the testing data.
- Finally, the prediction results are taken.

IV. Classification Algorithms

A. Decision Tree:

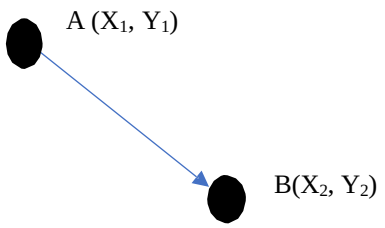
Decision Tree is a classification algorithm which is of three types, mainly Binary, Multi class and Multi Label. We have used Binary classification of '1'and '0'. Where '1' -> The person has a high probability of suffering from Lung Cancer. And '0' says that the person has low chances of Lung Cancer.

The root node is the node which the decision tree classifier wants to predict. They are calculated by two methods: 1. Gini Indexing 2. IG [Information Gain] & Entropy. The root node is the node with the highest IG. The following comes the Leaf node, which has no further classifications. The classifier has decided whether it's a yes or a no. The leaf node has no further branches.

B. KNN Algorithm:

KNN stands for K -Nearest Neighbor. This algorithm is used to predict an unknown data point belonging to a specific class. It is calculated by the Euclidean distance between any data point of any class and the unknown data point. The working principle goes as follows: The unknown data point is placed on the graph. Now we take randomly 5 – 6 datapoints close-by as the nearest neighbors of the unknown data

point. Now in order to calculate the radius we find the *Euclidean Distance* (default distance measuring unit in KNN Algorithm). We calculate it by the following methodology:



Therefore, the Euclidean Distance between the two points is:

$$\sqrt{(X1 - X2)^2 - (Y1 - Y2)^2}$$

Now repeating the steps for other points and now, the shortest distance of all points will determine which class does the unknown point determine.

V. Workflow diagram

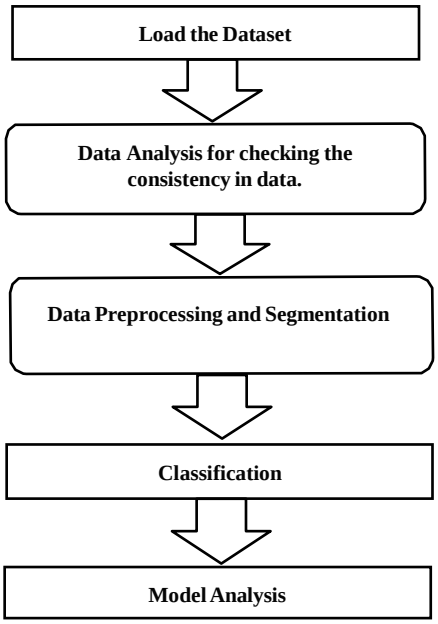


Fig. 1

VI. Model Results

The below graphs show the resultant output after the model gets trained with all the classifier.

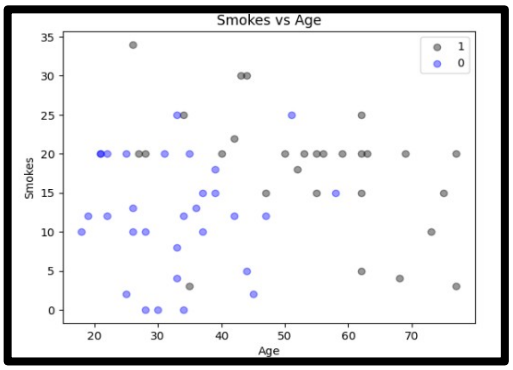


Fig. 2

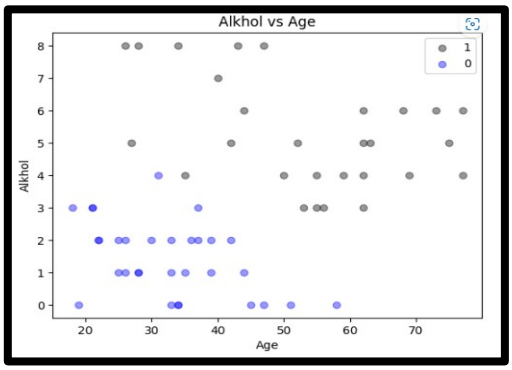


Fig. 3

The graph in Fig. shows the smokes vs age graph which tells us the chances of having Lung Cancer if an individual smokes as per the quantity of cigarettes per day according to the age. The parameters are; : The person has low chances of Lung Cancer, whereas : The person has high chances of Lung Cancer. The graph in Fig. points out the probability of Lung Cancer of individuals according to the Alchohol consumption quantity. The parameters are; : The person has low chances of Lung Cancer, whereas : The person has high chances of Lung Cancer.

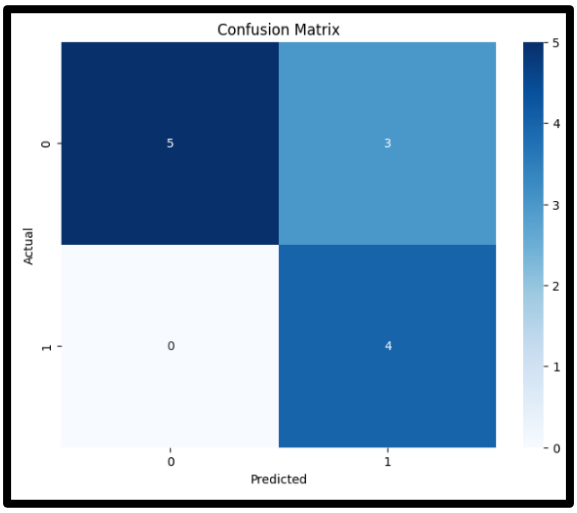


Fig. 4 Confusion Matrix for Decision Tree Classifier

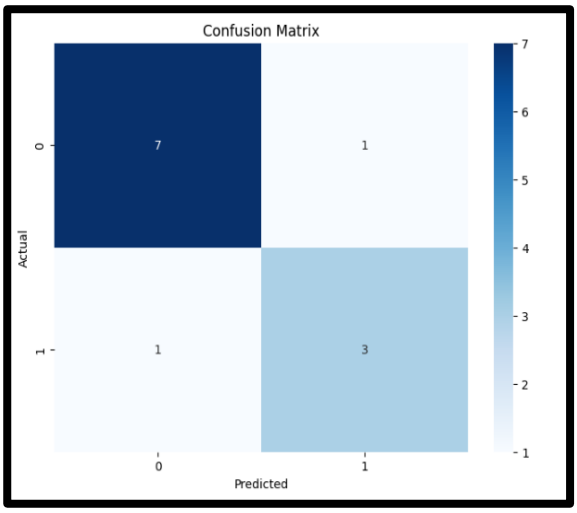


Fig. 5 Confusion Matrix for KNN Algorithm

Classifier	Accuracy
Decision Tree	92.03%
KNN	83.34%

Table 1 : Comparison of Accuracy

Classifier	Sense	Spec	Pres
Decision Tree	1.0	0.625	0.571
KNN	0.75	0.875	0.75

Table 2 : Comparison of Sensitivity, Specificity, Precision

even today early symptoms of cancer is not predictable. So with the modern techniques and application of AI, ML and DL we have created a model with a goal of early diagnosis of Lung Cancer.

VII. Problem Faced

Although our model, which has been trained using an open source dataset can predict lung cancer with an accuracy above 80%, availability of hospital test dataset of real patients would have improved the performance and accuracy of our model.

VIII. Future Scope

We can implement deep learning for image classification of CT scan, MRI and Ultrasound test of patients. We will use original data of the patients for training our ML model. We can train our ML model to detect other diseases apart from lung cancer. We can create a website for easier approach by the end users.

IX. Conclusion

We have finally succeeded in creating a ML Model which would predict precisely whether a person has Lung Cancer or not. Earlier back in history when there were no modern techniques and AI, ML, DL was not in action, it was difficult for doctors to predict manually, and

X. References

[1] “A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms”, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amrita Puri, India, IEEE 2018.

[2] “Symptoms Based Disease Prediction Using Decision Tree and Electronic Health Record Analysis”, European Journal of Molecular & Clinical Medicine, ISSN 2515-8260 Volume 7, Issue 4, 2020.

[3] “Efficient heart disease prediction system using decision tree”, International Conference on Computing, Communication and Automation (ICCCA 2015).

[4] “Heart Disease Prediction Using Classification with Different Decision Tree Techniques”, International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November 2014.

[5] “Using decision tree classification to assist in the prediction of Alzheimer’s disease”, 2014 6th International Conference on CSIT.

[6] “Comparing Association Rules and Decision Trees for Disease Prediction”, Carlos Ordonez, University of Houston, Houston, TX, USA-November 2006.

[7] “An extensive review on lung cancer detection using machine learning techniques”, Article in Journal Critical Reviews - May 2020.

[8] “A heart disease prediction model using decision tree”, IOSR Journal of Computer Engineering (IOSR- JCE), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 6 (Jul. - Aug. 2013), PP 83-86.

[9] “IoT based smart system to detect mental health emergencies: A proposed model”, American Journal of Science and Engineering volume-2, Issue-4 - March 2022.