

# Data Mining Techniques to Identify Lung Cancer Prediction

<sup>1</sup>Dharnasi Prasad, <sup>2</sup>Dr. G. Apparao Naidu

<sup>1</sup>Dept of CSE, Shri Venkateshwara University, Gjraulta, Uttar Pradesh, India

<sup>2</sup>Dept of CSE, JB. Institute of Engineering and Technology, Moinabad, Hyderabad, India

## Abstract

The major cause for death in human beings is because of cancer. Lung cancer is one of the most common and serious types of cancer that severely harms the human body. In order to cure the cancer early cancer detection is required. If lung cancer is diagnosed at early stages many lives will be saved. The other name for lung cancer is lung carcinoma, an uncontrolled malignant tumor distinguished by undisciplined cell growth in lung cells. There are many people suffering from this kind of cancer and confining to death. If this is left untreated, this may grow later than lung by metastasis into other parts of body. Many of the cancers starts from lungs, called as primary lung carcinoma. There are two types of small cell lung carcinoma (SCLC), non small cell lung carcinoma(NSCLC). The main reason for lung cancer is smoking of cigarette. There are many researches targeting on exact approaches for treating cancer. To predict the survival rate for NSCLC patients data mining techniques can be used with selection of algorithms. The algorithms used to detect the lung cancer are Support vector machine (SVM), Decision tree, k-Nearest neighbour, Random forest, Logistic regression. In this paper By implementing 2 different datasets and various packages and libraries in python, it is compared and on implementation found suitable algorithms have more accuracy on certain data sets for optimum prediction rate of lung cancer.

## Keywords

SCLC, NSCLC, SVM, Decision Tree, Logistic Regression, Random Forest, KNN Classifier.

## I. Introduction

Lung cancer is the genuine purpose behind disease destruction on the planet. The symptoms of lung malignant growth come into light at the last orchestrate. So it is incredibly hard to distinguishing its beginning stage. Hence, the passing rate is high for lung malignant growth in relationship with each and every other kind of disease. The infection has various phases where it undertakes from the minor tissue and escalates all through the various regions of the lungs by a mechanism called metastasis. It is the uncurbed growth of unwanted cells in the lungs [1]. It is assessed that around 12,203 individuals had lung threatening development in 2016, 7130 folks and 5073 females, suffering from lung sickness in 2016 were 8839. Lung cancer asserts a larger number of lives every year than prostate, ovarian and bosom cancers joined.

Individuals who smoke have the pinnacle danger of lung cancer, however lung cancer can likewise happen in individuals who have never smoked. Endurance rate for lung cancer is assessed to be 15% following 5 years of analysis. Information mining procedures can be valuable to gauge danger of mortality because of lung cancer dependent on symptomatic and treatment properties. The threat of lung cancer increments with the period of time and the quantity of cigarettes you've smoked. In the event that you quit smoking, even in the wake of smoking for a long time, you can radically lessen your odds of creating lung cancer.

As indicated by a survey made by world wellbeing association that consistently more than 7.6 million people suffered the lung malignant growth. It is estimated around 17 million cases of lung cancer throughout world in 2030. In year of 2005, around 1,362,825 new malignant growth cases are typical and around 5,71,590 patients suffer in view of lung disease in United States. It was evaluated that there will be 162,921 people from lung disease, suffer out of which 30% have malignancy attack. If predicted in advance can make a doctor aware of the situations with he can take necessary precaution to save the patient from initial to advance stage of cancer. Our strategy for finding the conceivable Lung cancer patients depends on the precise investigation of side effects and hazard factors. Non-clinical side effects and hazard components are a portion of the conventional pointers of the cancer ailments. The prediction of cancer tumor can be possible by utilizing data mining techniques. Use of algorithms like Naive Bayes, Apriori, clustering algorithms produces results about by which we can anticipate or caution the patient about the tumour and the dangers related with it. Here in the survey we find various classifiers of support vector machine (SVM), k-Nearest Neighbours, KNN, Logistic Regression, Decision Tree, Random Forest for predicting and comparing on a existing dataset.

## II. Literature Survey

Literature review on lung cancer methods of Machine Learning predicting missing information is very common. In a given therapeutic space, AI models ought to have good execution in any event, when missing information happens. Non-small Cell Lung Cancer (NSCLC) is a main demise malady in most of the global regions. Xueyan Mei [1]. Predicting Five-year Overall Survival in Patients with Non-small Cell Lung Cancer by ReliefF Algorithm and Random Forests: Mohammed [2] proposed a technique called feature selection which is the key task in which there will be multiple raw data contains many features selected based on their importance.

## Data Mining Techniques to Identify Lung Cancer Prediction

In [3] it is found that Relief and random forest generating best prediction.

Numerous studies are concentrating on careful ways to deal with treating the illness. The five-year by and large endurance rate for NSCLC patients is commonly anticipated by conventional relapse models with little examples and information size. In [4] AI apparatuses is presented which includes choice calculations and arbitrary timberlands classifier to anticipate the five-year in general endurance rate on a huge database. The consequences of this investigation [5] show Lung cancer patients who get radiotherapy as a major aspect of their treatment are exposed to danger radiation-actuated lung in-jury known as radiation pneumonitis (RP). RP is a possibly deadly symptom leading to risky treatment. Subsequently, new techniques are expected to manage doctors to recommend focused on treatment measurement to patients at high danger of RP [6]. A few prescient models dependent on customary factual strategies and AI systems have been accounted for, nonetheless,

no direction to variety in execution has not been given till date. Accordingly, in this work, it is thought about a few broadly utilized arrangement calculations in the AI field are utilized to recognize diverse hazard gatherings of RP[7] The presentation of these arrangement calculations is assessed related to a few element choice systems and the effect of the component determination on execution is future assessed[8].

Early Detection of Lung Cancer utilizing SVM Classifier [9] in Biomedical Image Processing used Deep learning for detection of small cell lung cancer using CT Scanning. It is likewise recommended that CT scan [10] picture is utilized as information picture, is handled and beginning period lung disease is recognized utilizing a SVM (support vector machine) calculation as a classifier in the grouping stage to improve exactness, affectability and explicitness [11-14].

### III. Methodologies

In this paper 5 Machine learning algorithms based algorithms are implemented and compared for a common data set. Implementation results shows the accuracy of each algorithm for prediction. The algorithms implemented are as follows.

- Decision tree.
- K-Nearest Neighbour.
- Logistic Regression.
- Random Forest.
- Support Vector Machine.

#### A. Decision tree

A decision tree is a resolve help tool that uses a tree-like model of decisions and their potential results, including chance occasion results, asset expenses, and utility. It is one approach to show a calculation that just contains restrictive control access.

Decision trees are normally utilized in operation research, explicitly in decision examination, to help intent a system well

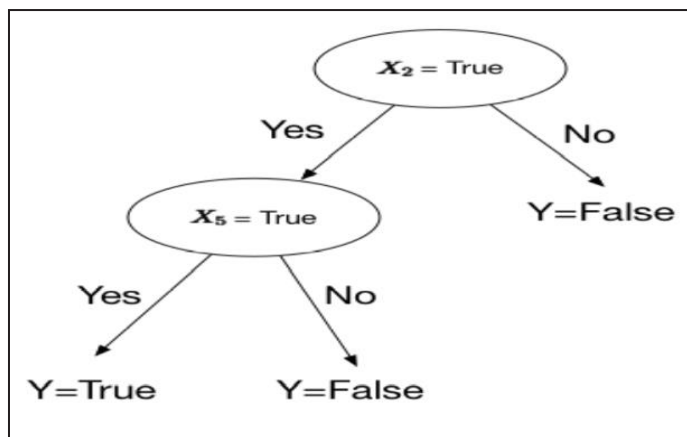


Fig. 1: Indicates Example Of Decision Tree

#### B. K-Nearest Neighbour

The k-nearest neighbours estimation (k-NN) is a non-parametric methodology used for gathering and relapse. In the two cases, the data involves the k closest getting ready models in the segment space. The output depends upon whether k-NN is used for request or backslide.

#### C. Logistic Regression

In statistics, the strategic model (logistic model) is utilized to display the probability of a specific class or occasion existing, for example, pass/fail, win/lose, alive/dead.

Numerically, a binary logistic model has a dependent variable with two possible qualities, for example, pass/fail which is spoken to by a pointer variable, where the two qualities are marked "0" and "1".

#### D. Random Forest

Random Forest is a classifier that advances from choice trees. It comprises of numerous choice trees. To order another case, every choice tree gives an arrangement to enter information; irregular Random Forest gathers the characterizations and picks the most casted a ballot forecast as the outcome. The contribution of each tree is inspected information from the first dataset. What's more, a subset of highlights is haphazardly chosen from the discretionary highlights to develop the tree at every hub. Each tree is developed without pruning. Basically, irregular random forest empowers numerous feeble or pitifully related classifiers to frame a solid classifier.

#### E. Support Vector Machine

In ML, support-vector machines (SVMs, in addition support-vector networks) are regulated learning models with related learning algorithms that analyze data set used for arrangement and regression analysis. Given a set of training examples, each set apart as having a position with both of two groupings, a SVM preparing calculation constructs a model that assigns out new guides to one class or the other, making it a non-probabilistic binary classifier.

### IV. Experimental Results

By using data sets from UCI different ML techniques are applied (decision tree, KNN classifier, logistic regression, random forest, SVM classifier) to predict the accuracy.

Table 1: Comparison Results of Accuracy of ML Techniques

ML Techniques	Training set Accuracy	Testing set accuracy
K Nearest Neighbour	76.52	66.43
Logistic Regression	92.72	90.20
Decision Tree	10	95.1
SVM Classifier	0	63
Random Forest	10	95.1

Discussions Of Out Put Of Dataset 1:

The below shown screenshots are the output of the accuracy of prediction of lung cancer by different machine learning techniques by using the lung cancer data. The Fig 1.2 represents the output of data analysis of the cancer data which we have taken among which how many have the cancer and how many did not have the cancer.

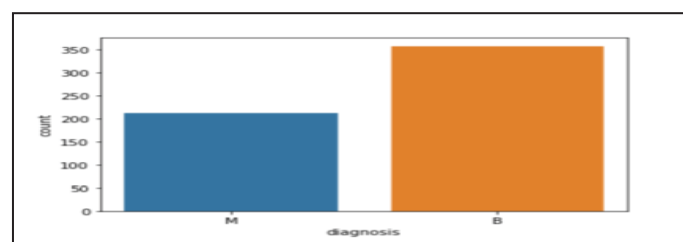


Fig. 2: Data Analysis of Diagnosis Count

Fig. 3 shows the accuracy of each technique implemented in Python. It shows K-NN classifier prediction as 93.45% on the training data and 91.2% correctly among test data. Correspondingly all other 4 methods accuracy obtained is shown in fig. 3.

```

Accuracy of K-NN classifier on training set: 0.7652582159624414
Accuracy of K-NN classifier on test set: 0.6643356643356644
logistic regression Training set score: 0.9272300469483568
logistic regression Test set score: 0.9020979020979021
dtree Accuracy on training set: 1.000
dtree Accuracy on test set: 0.951

random forest Accuracy on training set: 1.000
random forest Accuracy on test set: 0.958
svc Accuracy on training set: 1.00
svc Accuracy on test set: 0.63
    
```

Fig. 3: Accuracy for Each Technique.

Fig. 4 Shows the accuracy of each technique in bar graph blue represents, training set and orange represents testing set.

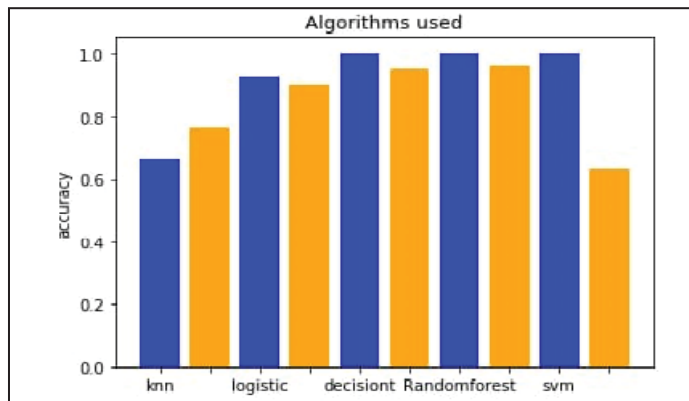


Fig. 4: Comparison Results

**Discussion of Out put of Dataset 2:**

Table 2: Comparison Results of Accuracy of ML Techniques

ML techniques	Training set Accuracy	Testing set accuracy
K nearest neighbour	75	37
Logistic regression	100	25
Decision tree	10	12
Svm classifie	0	38
Random forest	83	25

**Lung Cancer Prediction using Data Mining Techniques**

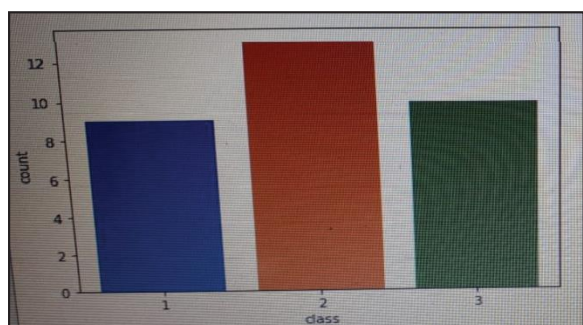


Fig. 5: Class Data Set Count

The fig. 5 shows screenshots of the output on the accuracy of prediction of lung cancer by different machine learning techniques by using the lung cancer data. The fig. 5 represents the output of data analysis of the cancer data which is implemented and compared.

```

Accuracy of K-NN classifier on training set: 0.75
Accuracy of K-NN classifier on test set: 0.375
c:\users\sahana\appdata\local\programs\python\python37\lib\site-packages\sklearn\linear_
Specify a solver to silence this warning.
FutureWarning)
c:\users\sahana\appdata\local\programs\python\python37\lib\site-packages\sklearn\linear_
Specify the multi_class option to silence this warning.
"this warning.", FutureWarning)
logistic regression Training set score: 1.0
logistic regression Test set score: 0.25
dtree Accuracy on training set: 1.000
dtree Accuracy on test set: 0.125
random forest Accuracy on training set: 1.000
random forest Accuracy on test set: 0.250
svc Accuracy on training set: 0.83
svc Accuracy on test set: 0.38
c:\users\sahana\appdata\local\programs\python\python37\lib\site-packages\sklearn\svm\base
version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or '
"avoid this warning.", FutureWarning)
    
```

Fig. 6: Accuracy of each techniques using Data set 2

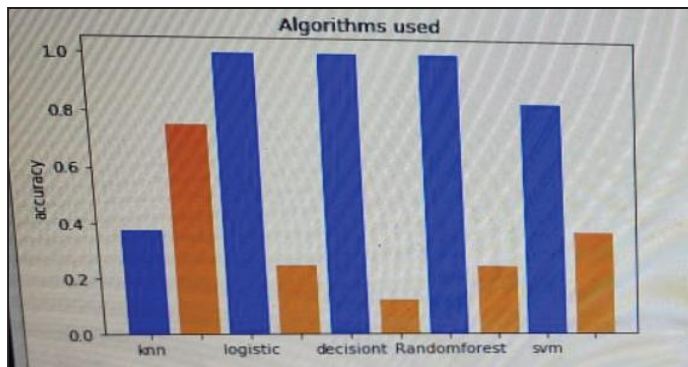


Fig. 7: Comparison of Results for Data set 2

Fig. 7 shows the accuracy of each technique in bar graph blue represents, training set and orange represents testing set.

**V. Conclusion**

By obtaining the results of two data sets with the help of ML techniques, It is clear that the rate of prediction performed is better in KNN and Logistic Regression compared with decision tree, random forest, and SVM .It is also concluded that accuracy depends on the training data set methods before introducing them to testing.

**References**

- [1] Xueyan Mei, Predicting Five-year Overall Survival in Patients with Non-small Cell Lung Cancer by ReliefF Algorithm and Random Forests, Feb 2017.
- [2] Mohammad Ismail K.Naga Lakshmi, Y. Kishore Reddy, M. Kireeti, T.Swathi” Design and Implementation of Student Chat Bot using AIML and LSA” International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-6,
- [3] Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm”, march 2017.
- [4] Mohammad Ismail K.Naga Lakshmi, Y. Kishore Reddy, M. Kireeti, T.Swathi” Design and Implementation of Student Chat Bot using AIML and LSA” International Journal of Innovative Technology and Exploring Engineering (IJITEE)

Volume-8 Issue-6, ISSN: 2278-3075 PP 1742-1746 April 2019

- [5] K.Srinivas,Dr.Mohammed Ismail.B “Testcase Prioritization With Special Emphasis On Automation Testing Using Hybrid Framework” Journal of Theoretical and Applied Information Technology Vol. 96. No 13 4180-4190 July 2018
- [6] Jane Alam, Sabrina Alam, Alamgir Hossan, “Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier”, 2018.
- [7] Mohammed Ismail B , K. Bhanu Prakash, M. Nagabhushana Rao” Collaborative Filtering-Based Recommendation of Online Social Voting” International Journal of Engineering and Technology “ Volume 7 issue 3 1504-1507 July 2018
- [8] Sarah Soltaninejad, Mohsen Keshani, Farshad Tajeripour, “Lung Nodule Detection by KNN Classifier and Active Contour Modelling and 3D Visualization”, April 2012.
- [9] Mohammad Ismail, V. Harsha Vardhan, V. Aditya Mounika, K. Surya Padmini “An Effective Heart Disease Prediction Method Using Artificial Neural Network “International Journal of Innovative Technology and Exploring Engineering’ at Volume-8 Issue-8, pp 1529-1532 June 2019.
- [10] Mohammed Ismail. B, B. Eswara Reddy, T. Bhaskara Reddy “Cuckoo Inspired Fast Search Algorithm for Fractal Image Encoding” Elsevier Journal of King Saud University Computer and Information Sciences volume 30 issue 4, ISSN: 1319-1578 Pages 462–469 DOI 10.1016/j.jksuci.2016.11.00 Oct 2018
- [11] Kumar S.A, Vidyullatha P "A comparative analysis of parallel and distributed FSM approaches on large-scale graph data" International Journal of Recent Technology and Engineering, Volume 7, Issue 6, April 2019, Pages 103-109
- [12] Pellakuri Vidyullatha, Rajeswara Rao D, "Training and development of artificial neural network models: Single layer feedforward and multi layer feedforward neural network", journal of Theoretical and Applied Information Technology Volume 84, Issue 2, 20 February 2016, Pages 150-156
- [13] Mareedu Lakshmi Vihari, K Amarendra, Navvula Anusha Recognition of Zeroday Exploit, International Journal of Engineering & Advanced Technology (IJEAT), ISSN: 2249-8958, Volume: 08, Issue: 04, pp.1875-1877, April (2019).



Mr. Dharnasi Prasad, well known Author and excellent teacher Received B.Tech(CSIT) and M.Tech (Software Engineering) from Jawaharlal Nehru Technological University and as well as pursuing my Ph.D in Machine Learning and Artificial Intelligence from Shri Venkateswara University is working as Lecturer in Computer Science and Engineering Department as well as Department of BAIS (Business Administration and Information

System (Sawla Campus)) from Arba Minch University, He is an active member of ISTE. he has 17 years of teaching experience in various engineering colleges. To his credit no of publications both national and international conferences /journals . His area of Interest includes Data Base Management, Compiler Design, mobile computing, computer organization, Neural networks and fuzzy logic, Software Engineering, Computer Network, Windows Programming, Object Oriented SAD, Data Communication and Computer network as well as Computer network and System Administration and other advances in computer Applications.



Dr. G. Apparao Naidu, well known Author, and excellent faculty in Computer Science and Engineering, presently working as Professor of CSE and Dean Academics at J.B. Institute of Engineering & Technology, Moinabad, Hyderabad. Under his guidance 8 scholars' research work is in progress. He has submitted one patent and two book chapters with international publishers. His areas of

interests are Data mining, Information security, Image processing, Software Engineering as well as Machine Learning and Artificial Intelligence. Published more than 60 papers in various International Journals and Conferences. He was Conducted 12 Workshops and Attended 32 Workshops. He was Guided 90 Projects for UG and PG students. He got 4.27 Lakhs AICTE grant for FDP. He is acting as a Member BOG, BOS and Member Secretary Academic Council for JBIET.