iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6323

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

Survey on Lung Cancer Detection using Machine Learning

Hemant Jaiman[1], Dr. Kuldeep Sharma[2], Sujatha K[3]
[1]*Mtech Data Science Student,* [2]*Associate Professor,* [3]*Assistant Professor, Department of Computer Science and Engineering, Jain Deemed to be University, Bengaluru, India*

Abstract: ***Lung cancer is considered as the development of cancerous cells in the lungs. Mortality rates for both men and women have increased due to increasing cancer incidence. Lung cancer is an illness in which cells uncontrollably multiply in lungs. Lung cancer cannot be prevented but can reduce its risk. So earliest detection of lung cancer is crucial to patients' survival rate. The number of chainsmokers is directly proportional to the number of people who have affected by lung cancer. The prediction of lung cancer is analysed using various machine learning classification algorithms such as Naive Bayes, SVM, Tree of Decision and Logistic Regression. The key aim of this paper is to diagnose lung cancer early by examining the performance of classification algorithms.***

Keywords: Artificial Intelligence. Machine Learning, Deep Learning, Lung cancer.

I. INTRODUCTION

Lung cancer constitutes large proportion of death rates among cancer patients. Lung cancer may initiate in windpipe, main airway, or lungs. It is caused by unregulated growth and spread of some lung cells. People with pulmonary illness such as Emphysema and previous chest issues have a higher risk of being diagnosed with lung cancer. Overuse of tobacco, cigarettes and beedis is the main risk factor that leads to lung cancer in Indian men; however, smoking is not so common among Indian women, which suggests other factors leading to lung cancer. Other risk factors include exposure air pollution, radon gas and chemicals I n the workstation. A cancer that begins in the lung is primary cancer of the lung whereas those that begin in the lung and spread to other parts of the body are secondary cancer of the lungs. The stage of cancer is measured by tumour size and how far it has spread. An early stage cancer is a small cancer diagnosed in the lung, and advanced cancer has spread into the surrounding tissue or other body parts. A better understanding of risk factors can help prevent pulmonary cancer. Early detection using machine learning techniques is the key to improving the rate of survival and if we can make the diagnostic process more efficient and effective for radiologists using this, it'll be a key step towards the goal of improved early detection.

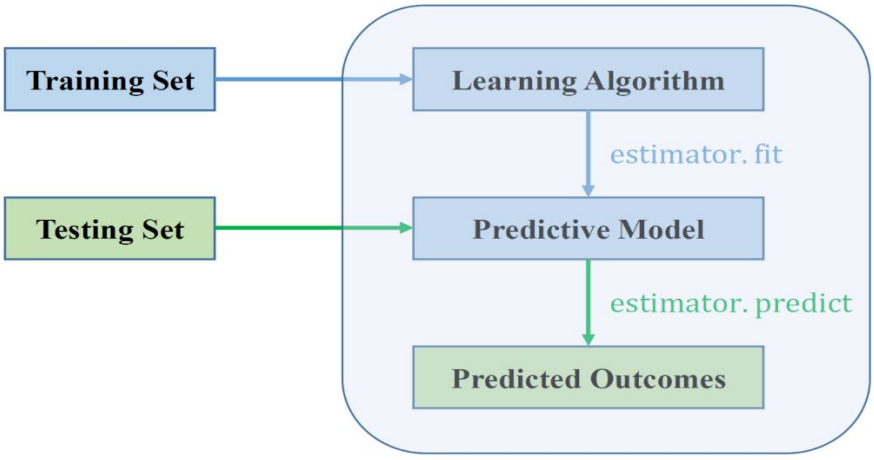


Fig 1.overall architecture

The data set of Lung cancer in this paper is from UCI Machine Learning Database. First, by using k-fold cross validation technique, the given datasets are split into training and test data. Then use the classification algorithms such as SVM, Logistic Regression, Naïve Bayes and Decision Tree to implement the respective classification models using the training data provided. Classification models are created using training data, and the respective models are evaluated using test data to obtain model accuracy. Lastly, we compared the accuracy rates of each and every classification model we implemented and we concluded.

II. RELATED PREVIOUS WORKS

Machine learning takes AI software a step further as it allows intelligent learning to take place within the component based on previous work done or data extrapolation. The software performs sophisticated decision-making processes as it progresses and learns from previous activities. A brief description of research papers based on detection of Lung Cancer using various machine learning algorithms is explained below:

A. Compare algorithms such as Decision Tree, Naive Bayes and Artificial Neural Network with the prediction of post-operative life expectancy in lung cancer patients using predictive data mining algorithms. On the above algorithms a stratified 10-fold comparative cross-validation analysis was performed and the accuracy was determined for each classifier.

B. The paper deals with a comparative study of the Brain Tumour detection classification algorithm. The overall accuracy rate was calculated using volumetric and location features based on 2 classification classes such as logistic regression and Quadratic Discriminant, and 3 classifications such as Linear SVM, Coarse Gaussian SVM, Cosine KNN and Complex and Median Tree.

C. For each lung cancer classifier obtained, different results are produced in this paper. Classifiers such as KNN, SVM, NN and Logistic Regression have been implemented and appropriate accuracy rates have been obtained. Support Vector Machine has the highest accuracy at 99.3 percent. The proposed method has been applied to the medical dataset, which has helped doctors make better decisions.

D. Several segmentation algorithms like Naïve Bayes, Secret Markov Model etc. were discussed. Proper explanation is given of how and why different segmentation algorithms are used to detect Lung tumour.

E. Explained how to create a basic flowchart for an algorithm used for brain tumour detection. Discussed two types of techniques of data mining and methods of classification

Statistical methods- Naïve Bayes, SVM
Data compression methods-Decision tree, Neural Network
Discussed about various datasets.
BRATS Dataset
OASIS Dataset
NBTR Dataset

III.CLASSIFICATION ALGORITHMS

A. Support Machine Vectors

SVM is a supervised method of learning that analyzes data which I used to evaluate classification. For non-linear divisibility. SVM is more suitable for datasets, as it reduces Level of misclassification. The goal in SVM given the data is to Find the least distant point from the classes and try to find the distance maximized. As explained in below figure the red dots belongs to class 1 and green stars belongs to class 2 and both are separated by a hyper plane.

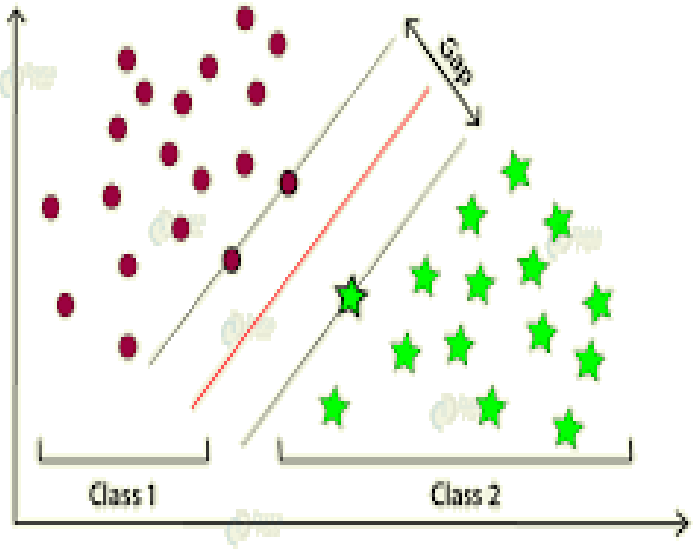


Fig.2. Basic Structure of SVM

B. Decision Tree

Decision tree employs supervised learning techniques build a model which is in the form of a tree data Structure (set of in hierarchical organized nodes Fashion) .Initially, parent entropy is measured. The gain of information is then calculated by the subtraction Weighted entropy of children parent. The one with the highest benefit in knowledge is considered the root node and is continuing the process Until the Classification is completed. To predict the result, the tree is used. In the decision tree, each node specifies a specific symptom from the set S = {s1,s2,s3 ... sj} where S specifies conditional attributes, vi, k denotes the values of each branch, i.e. the i-th range of I symptoms and leaves with D={d1,d2, ... dk} and binary values, wdk={0,1}. By writing down each path from the root to the leaves, by converting the decision tree, a set of association rules was created. Fig.3 describes Decision Tree as an association set rule.

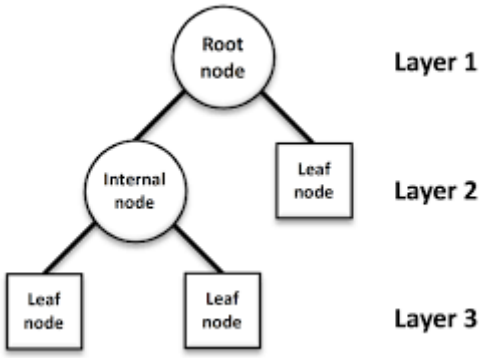


Fig3. Basic decision tree

C. Naive Bayes

A classifier Naive Bayes is a model of probabilistic machine learning which is used for classification tasks. This classifier is very quick and easy to implement but their biggest disadvantage is that they have to be independent of the predictors. In most cases of real life, the predictors are dependent, this impedes the classifier's efficiency. One of the easiest ways to select the most likely hypothesis, given the data we have that we can use as our prior knowledge of the issue. Bayes' Theorem offers a way of estimating the probability of a hypothesis given our prior knowledge.

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

Fig .4. Naive bayes formula

D. Logistic Regression

By fitting a linear equation to observed data, linear regression attempts to model the relationship between two variables. One variable is considered an explanatory variable, and the other variable is considered as dependent.

The linear regression has an equation in the form of Y = a + bX, where X is the independent feature variable and Y is the dependent variable which is also known as label. The slope of the line is b, and a is the intercept on y axis when x=0.

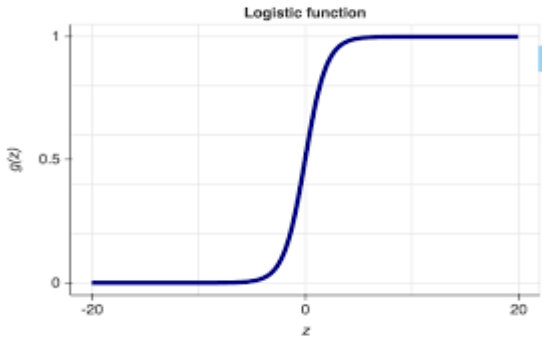


Fig.5. logistic regression view

IV. EXPERIMENTAL RESULTS AND EVALUATION PERFORMANCE

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads—the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

We used the dataset from data.world for the study.

data.world: https://data.world/cancerdatahp/lung-cancer data.

This dataset contains: Number of examples: 1000 and Number of columns: 25(1 class (label) column, 24 predictive/feature column).

Columns Description of data.world: Proper classification of the detection of lung cancer is made using efficient use of attributes under which attributes are represents the symptoms. The attribute of age , sex, Occupational air pollution, drug consumption, dust allergy Genetic risks, chronic lung disease, a healthy diet, Obesity, smoking, passive smoker, chest pressure, coughing Pain, weight loss, exhaustion, shortness of breath, wheezing, Swallowing trouble, finger nail clubbing, Frequent Cold, Dry cough, Snoring is taken into account to predict Pulmonary cancer. In the mark, the severity value '2' indicates a Malignant tumor,'1'-benign tumor and '0' Healthy without tumor.

The following machine learning classification models were used to detect lung cancer and corresponding accuracy rates were noted as below:

ML Algorithm	Accuracy(%)
Logistic Regression	66.7
Decision Tree	90
Naïve Bayes	87.87
SVM	99.2

Table 1: Lung Cancer on different machine learning algorithms

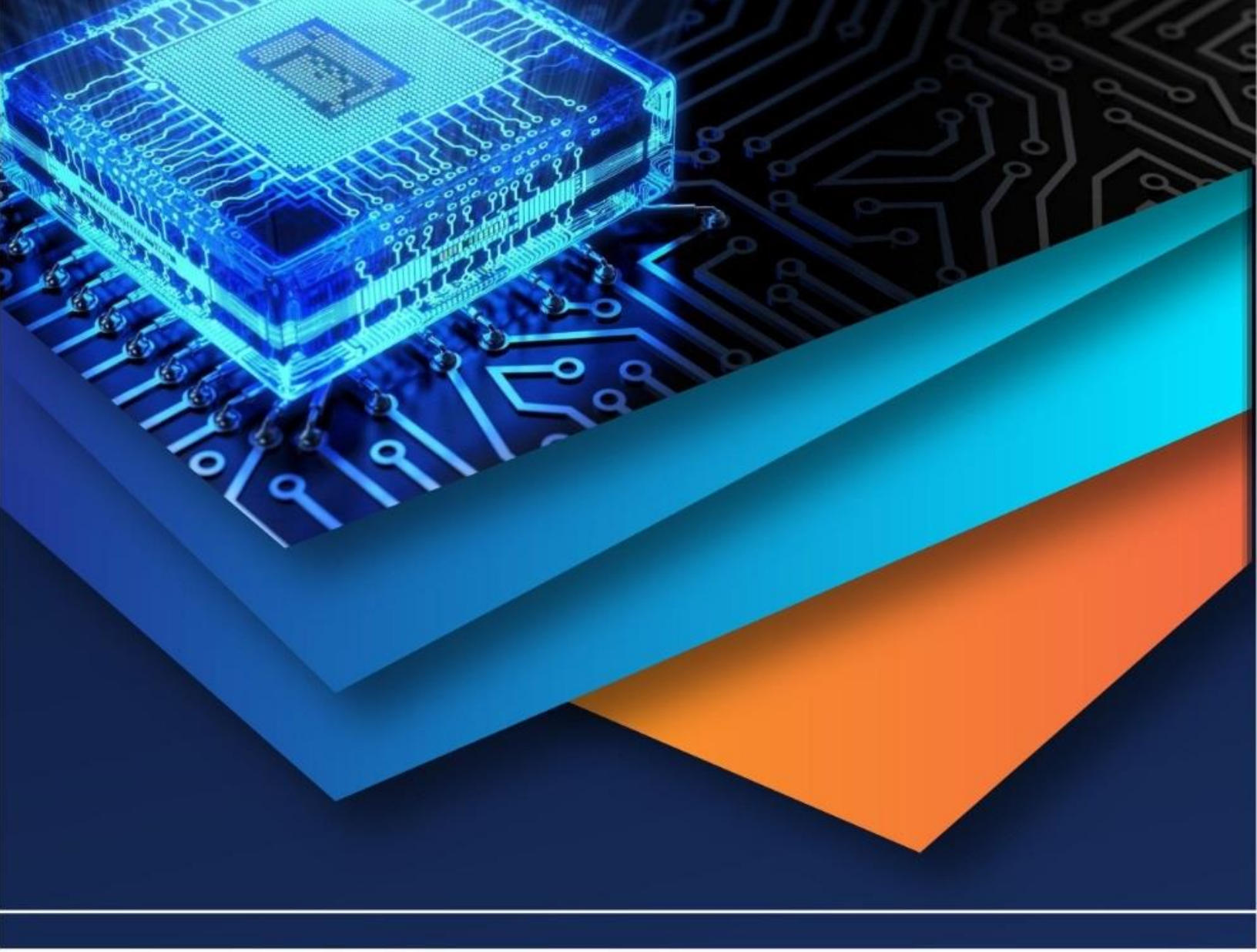
Above Table shows that the efficiency of SVM exceeds all other machine learning classification algorithms including Logistic Regression. So we can conclude that with SVM we obtain the highest accuracy rate compare to all other classification algorithms for this particular datasets.

V. CONCLUSIONS

The doctor has to do several tests in earlier periods to determine whether patient is sufferer of lung cancer or not. But that was a process that was very time consuming. Sometimes in a diagnosis a patient has to undergo needless check-ups or multiple tests to classify the lung cancer disease. There needs to be a preliminary test in which both the patient and the doctor are notified of the possibilities of lung cancer to minimize process time and unnecessary check-up. The machine learning algorithms currently play an important role in predicting and classifying medical data. The machine learning algorithms used for this comparative study are Logistic Regression, SVM, decision tree, and Naïve Bayes. A comparative analysis is presented of the precision rates of each classifier. The predictive performance of the classifiers is quantitatively compared. Various results for each classifier on the lung cancer dataset are produced in the performance chart. Looking at proper classification (CA) and other metrics; the support vector machine algorithm gives the best result. The SVM algorithm used high dimensions to identify the result, so that it is the best output. Using this technique, more precisely detecting lung cancer can be done. Thus, there are less errors. Finally, the precision can be improved by adding extra pre-processing.

REFERENCES

[1] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[2] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.
[3] KwetisheJoroDanjuma, " Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients" Department of Computer Science, ModibboAdama University of Technology, Yola, Adamawa State, Nigeria
[4] [2] Survey of Intelligent Methods for Brain Tumor Detection-IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1, September 2014
[5] Zehra Karhan1, Taner Tunç2, "Lung Cancer Detection and Classification with Classification Algorithms" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 22788727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 71-77.
[6] Ada, RajneetKaur, " A Study of Detection of Lung Cancer Using Data Mining Classification Techniques " International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013
[7] Survey of Intelligent Methods for Brain Tumor Detection-IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1, September 2014
[8] Lung Cancer detection and Classification by using Machine Learning & Multinomial Bayesian-IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) e-ISSN: 2278-2834,p- ISSN: 2278- 8735.Volume 9, Issue 1, Ver. III (Jan. 2014), PP 69-75
[9] K. V. Bawane , A. V. Shinde"Diagnosis Support System for Lung Cancer Detection Using Artificial Intelligence"-International Journal of Innovative Research in Computer and Communication Engineering,Vol. 6, Issue 1, January 2018
[10] H.R.H Al-Absi, B. B. Samir, K. B. Shaban and S. Sulaiman,"Computer aided diagonosis system based on machine learning techniques for lung cancer",2012 International Conference on Computer and Information Science(ICCIS),Kuala Lumpeu, 2012, pp. 295-300.
[11] Sukhjinder .Kaur "ComparativeStudy Review on Lung Cancer Detection Using Neural Network and Clustering Algorithm", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 4, Issue 2, February 2015
[12] D. Vinitha, Dr.Deepa Gupta, and Khare, S., "Exploration of Machine Learning Techniques for Cardiovascular Disease", Applied Medical Informatics, vol. 36, pp. 23–32, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY