International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms

Sakshi Jaiswal
Information Technology
Government College of Engineering
Amravati, India
jaiswalsakshi891@gmail.com

Rujwi Deshmukh
Information Technology
Government College of Engineering
Amravati, India
rujwideshmukh08@gmail.com

Nikhil Meshram
Information Technology
Government College of Engineering
Amravati, India
meshramnc@gmail.com

Aditya Deshpande
Information Technology
Government College of Engineering
Amravati, India
adi1234desh16@gmail.com

Prof. Bhushan Wakode
Information Technology
Government College of Engineering
Amravati, India
bhushan.wakode@gmail.com

Abstract: The diagnosis and detection of lung cancer has been conducted using various methods of data analysis and classification. Because lung carcinoma has no known cause, circumvention becomes impossible, making early detection of tumors in lungs the best way to treat lung cancer. It is called lung cancer when cancerous cells grow in lungs.
Increasing cancer rates have resulted in a higher mortality rate for both men and women. It is impossible to prevent lung cancer, but its risk can be reduced. In order to improve the survival rate of lung cancer patients, it is crucial to detect it at the earliest stage. There is a direct relationship between the number of chain smokers and lung cancer patients. There have been several attempts to develop predictive models to guide clinicians in the management of indeterminate lung nodules discovered incidentally or during screening. The use of such systems could lead to less variability in nodule classification, better decision making, and a reduction in the number of benign nodules that are not needed to be followed up. The lung cancer prediction was analysed using classification algorithms such as Naive Bayes, SVM, Decision tree and Logistic Regression. The key objective of this paper is the early diagnosis of lung cancer by examining the performance of classification algorithms.

Keywords—
DecisionTree;LogisticRegression;LungCancer Prediction,;NaïveBayes;Support Vector Machine

I. INTRODUCTION

Lung cancer is the principal cause for cancer-related death. The windpipe, lungs, and main airways can all develop lung cancer. The probability of being diagnosed with lung cancer is higher among people with lung diseases such as emphysema. Smoking is a major risk factor for lung cancer among Indian men; however, cigarettes are less prevalent among Indian women, which implies that other factors are to blame.There is also a risk associated with workplace exposure to chemicals, air pollution, and radon gas. As a primary lung cancer, primary lung cancer originates in the lung. As a secondary lung cancer, it originates in the lung then spreads to other body parts. A cancer's stage depends on the size and extent of the tumour. A small cancer found in the lung is considered an early stage, whereas a cancer that has spread to surrounding tissues or other body parts is considered an advanced stage. We can prevent lung cancer disease with a better understanding of risk factors.

The key is early detection using machines learning techniques, and if we can improve the diagnosis process and the quality of radiology reports with this, then we will be making a very big step towards improving early detection. The Lung Cancer datasets used for this study are taken from UCI Machine Learning Repository and Data World. First, the given datasets are divided into training and test data by using k-fold cross validation technique. Then using the classification algorithms such as SVM, Logistic Regression, Naïve Bayes and Decision Tree, respective classification models are implemented using the given training data. The classification models are created using training data and the

corresponding models are evaluated using test data to get the accuracy of the models. Finally, we compared the accuracy rates of each and every classification models that we implemented and arrived at a conclusion.

With the help of latest technology, like mobile tech, data science, machine learning algorithms etc. significant improvement can be made to existing system. The rest of the paper is arranged as follows – Section 2 which presents literature review, Section 3 describes the technology used, Section 4 shows the methodology employed for the machine learning task and the final section displays the results.

II. RELATED RORK

There are following are some works done related to same field-

By allowing intelligent learning within a program, machine learning takes AI software to a whole new level. The computer can learn from previous work that it performed or extrapolate from data. The software performs sophisticated decision making processes as it goes along and learns from previous activities. A brief description of the research papers based on Lung Cancer detection using different Machine learning algorithms are explained below.

A study that compares algorithms such as decision trees, naive bayes and artificial neural networks to predict post-operative life expectancy in lung cancer patients. Each algorithm has been analysed using a stratified 10-fold cross-validation approach and accuracy has been calculated for each classifier.

In this paper, we compare classification algorithms for detecting brain tumours. The overall accuracy rate was calculated by using volumetric and location features as well as 2 classifications, such as linear regression and quadratic discriminant, and 3 classifications, such as linear SVM, coarse Gaussian SVM, cosine KNN, and complex and median trees.

The accuracy rates of different classifiers with regards to lung cancer were investigated in this study. The classifiers KNN, SVM, NN, and Logistic Regression were all tested and their results were obtained. Based on a medical dataset, the proposed method proved to be 99.3% accurate. This method helped doctors make better decisions with the proposed method.

III. AN OVERVIEW OF STUDY

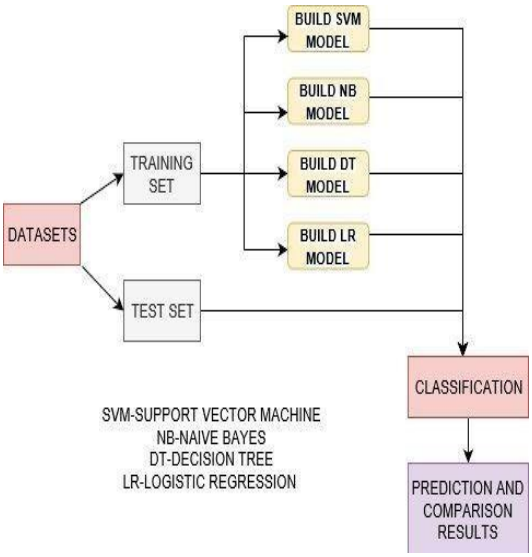


Fig.1. Overall Architecture

The Lung Cancer datasets used for this study are takenfrom UCI Machine Learning Repository and Data World. First, the given datasets are divided into

training and test data by using k-fold cross validation technique. Then using the classification algorithm such as SVM ,Logistic Regression, Naïve Bayes and Decision Tree, respective classification models are implemented using the given training data. The classification models are created using training data and the corresponding models are evaluated using test datato get the accuracy of the models. Finally, we compared the accuracy rates of each and every classificationmodels that we implemented and arrived at a conclusion.

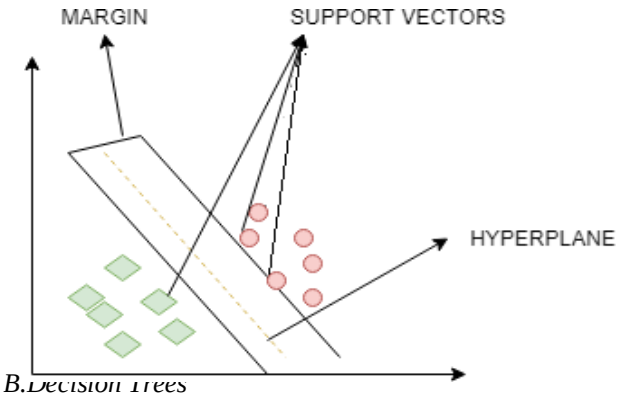
.

IV. CLASSIFICATION ALGORITHMS

A. Support Vector Machine

SVM is a supervised learning method that analyse data which is used for classification analysis. For non-linearly separable datasets, SVM is more suitable since it reduces the misclassification rate.In SVM given a data, the objective is to find the minimum distanced point from the classes and trying to find the maximized distance. Fig. 2 shows the structure of SVM.Here,green and pink images represent two different classes which is separated by a hyperplane. Also the margin and support vectors are properly labelled below.

Fig.2. General Structure of SVM



B.Decision Trees

Decision tree uses supervised learning technique to build a model which is in the form of a tree data structure(set of nodes arranged in hierarchical fashion) .Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.Initially,entropy of parent is calculated. Then Information gain is calculated by subtracting weighted sum of entropy of children from entropy ofparent. The one with highest Information gain is considered as the root node and the process goes on until the classification is done.Given a new test data,the tree is used to predict the result. In decision tree,each node specify a particular.

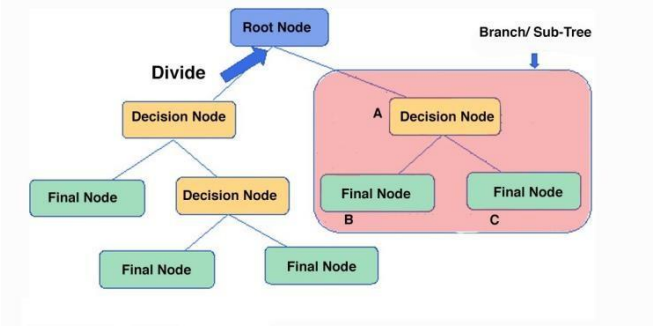


Fig.3 General Structure of Decision tree.

C. Naive Bayes

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

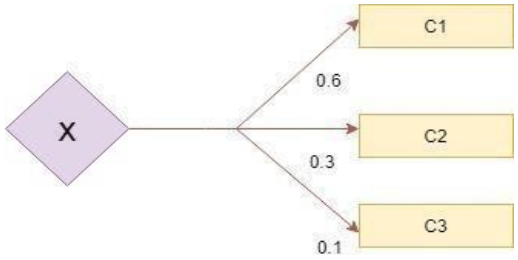
Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events and P(B) $\neq 0$.

Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.



Initialization involves calculating probabilistic values in order to determine which class the instance belongs to.The final class label is determined by the probability value that is the highest.The final class label is determined by the probability value that is the highest.A class C1 will have the highest probability value in the figure, and therefore the incoming X is from this class.

D.Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.Fig 5 shows the example of a logistic regression to distinguish two classes (orange-yellow images)

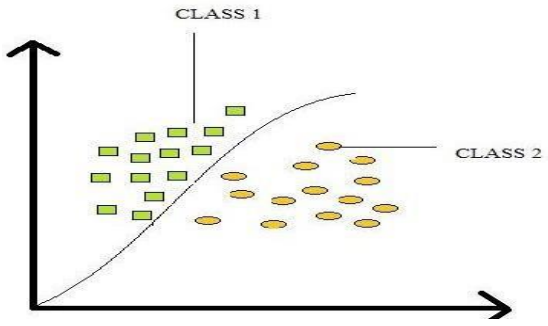


Fig.5. Logistic Regression to Distinguish two classes

E.KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

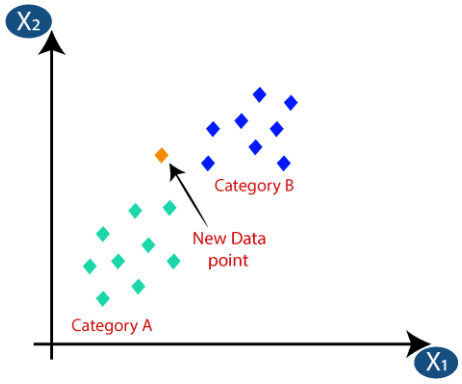


Fig 6. KNN to Distinguish new data point

EXPERIMENTAL RESULTS AND EVALUATION PERFORMANCE

In order to detect lung cancer effectively, attributes representing the symptoms must be effectively utilized. For the prediction of lung cancer, attributes such as age, gender, air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoking, chest pain, blood coughing, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of fingernails, frequent colds, dry coughs, and snoring are considered. The severity value ‘2’ in label indicates a malignant tumour ,‘1’-benign tumour and ‘0’ indicates a heathy person with no tumour. The following classification algorithms were used to detect lung

cancer and corresponding accuracy rates were obtained as follows:

ML Algorithm	Accuracy (%)
Logistic Regression	94.59
Decision Tree	81.08
Naïve Bayes	87.78
SVM	89.18
KNN	88.28

Confusion Matrix

A confusion matrix was used to evaluate the accuracy of each classifier. Experimental results show that the best classification missions are obtained using the five attributes from the SVM classifier. 95.56% prediction rate and 92.11% CNN correct answer rate. On the other hand, the estimated percentage of KNN is the lowest, 88.40 percent.

CONCLUSION

In earlier times, the doctor has to do multiple tests in order to detect whether a given patient has lung cancer or not . But this was a very time consuming process. In a diagnosis sometimesa patient has to undergo unnecessary check-ups or different tests to identify the disease of lung cancer.To minimize the process time and unnecessary check-ups there needs to be preliminary test in which both the patient and the doctor will be notified with the possibilities of lung cancer. Nowadays the machine learning algorithms plays animportant role in the prediction and classification of medical data. Logistic Regression, SVM, decision tree and Naïve Bayes are the machine learning algorithms used for this comparative study. A comparative analysis of accuracy rates of each classifier arepresented. The predictive performance of classifiers arecompared quantitatively. In the performance chart, different results are produced for each classifier on the lung cancerdataset.Looking at the correct classification (CA) and other metrics; the best result is given by the support vector machine algorithm. SVM algorithm used high dimension to classify the observation so it's performance is the best. More accurate lungcancer detection can be done using this technique. Therefore, there is less mistakes. Finally, by adding extra pre-processing the accuracy rate can be enhanced

ACKNOWLEDGEMENT

Prof. B.V.Wakode, Professor of Department of Information Technology, Government College of Engineering, Amravati, has been a tremendous source of support and guidance to the authors throughout this project.

REFERENCES

[1] 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

[2] Survey of Intelligent Methods for Brain Tumor Detection-IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1,

September 2014

[3] 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.

[4] 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

[5] Survey of Intelligent Methods for Brain Tumor Detection-IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 1,

September 2014

[6] 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

[7] 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

[8] 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.

[9] 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

[10] 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.

[11] Sathyan H, Panicker,J.V,,"Lung Nodule Classification Using Deep ConvNets on CT Images",9th International Conference on Computing,

Communication and Networking Technologies, ICCCNT 2018

[12] Isaac,J., Harikumar, S.,"Logistic regression within DBMS" ,Proceedingsof the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 20167918045, pp. 661-666,2016