A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms

Radhika P R
Department of Computer Science and Engineering
Amrita VishwaVidyapeetham,Amritapuri,India
radhikapr17@gmail.com

Rakhi.A.S.Nair
Department of Computer Science and Engineering
Amrita VishwaVidyapeetham,Amritapuri,India
rakhisopanam@gmail.com

Veena G
Department of Computer Science and Applications
AmritaVishwaVidyapeetham,Amritapuri,India
veenag@am.amrita.edu

***Abstract*—The growth of cancerous cells in lungs is called lung cancer. The mortality rate of both men and women has expanded due to the increasing rate of incidence of cancer. Lung cancer is a disease where cells in the lungs multiply uncontrollably. Lung cancer cannot be prevented but its risk can be reduced. So detection of lung cancer at the earliest is crucial for the survival rate of patients. The number of chain-smokers is directly proportional to the number of people affected with lung cancer. The lung cancer prediction was analysed using classification algorithms such as Naive Bayes, SVM, Decision tree and Logistic Regression. The keyobjective 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**

INTRODUCTION

Lung cancer is the principal cause for cancer-related death. Lung cancer can initiate in the windpipe, main airway or lungs. It is caused by unchecked growth and spread of some cells from the lungs. People with lung disease such as emphysema and previous chest problems have more chance to be diagnosed with lung cancer. Overusage of tobacco, cigarettes and beedis, are the major risk factor that leads to lung cancer in Indian men; however, among Indian women, smoking is not so common, which indicate that there are other factors which lead to lung cancer. Other risk factors include exposure to radon gas, air-pollutions and chemicals in the workplace. A cancer that starts in lung is primary lung cancer whereas those which starts in lung and spread to other parts of body is secondary lung cancer. Size of tumour and how far it has spread determines the stage of cancer. An early stage cancer is a small cancer that is diagnosed in lung and advanced cancer is the one that has spread into surrounding tissue or other part of body . A better understanding of risk factors can help to prevent lung cancer disease.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 .

AN OVERVIEW OF STUDY

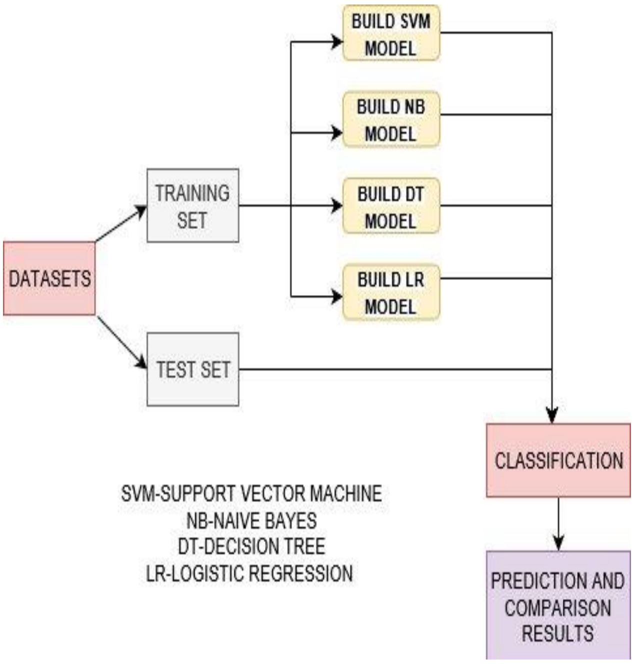

Fig.1. Overall Architecture

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

978-1-5386-1507-2/18/$31.00 ©2018 IEEE

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.

RELATED PAST WORKS

Machine learning takes AI software a step further as it enables intelligent learning to occur within the component based on previous work it did or extrapolations made 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:

[1]Deals with the prediction of post-operative life expectancy in lung cancer patients using predictive data mining algorithms to compare algorithms such as Decision Tree, Naive Bayes and Artificial neural network. A stratified 10-fold cross-validation comparative analysis was conducted on the above algorithms and accuracy was calculated for each classifier.

[2]Paper deals with comparative study of classification algorithm for detection of Brain Tumour. Using volumetric and location features overall accuracy rate was calculated based on 2 classification classes such as logistic regression and Quadratic Discriminant and 3 classification classes such as Linear SVM, Coarse Gaussian SVM, Cosine KNN and Complex and median tree.

[3]In this paper, different results are produced for each classifier on the lung cancer dataset obtained.The classifiers such as KNN,SVM,NN and Logistic Regression were implemented and corresponding accuracy rates were obtained. Support Vector Machine has the highest accuracy with 99.3%.The proposed method was applied to medical dataset which helped doctors to make more correct decision.

[4]Various segmentation algorithms were discussedwhich includes Naïve Bayes, Hidden Markov Model etc. Proper explanation is given about how and why various segmentation algorithms are used in detection of Lung tumour.

[5]Explained about how to create a basic flowchart for an algorithm which is used to detect brain tumour. Discussed about two types of data mining techniques and there classification methods.
1.Statistical methods- Naïve Bayes, SVM
2.Data compression methods-Decision tree, Neural Network
3.Discussed about various datasets.

- BRATS Dataset
- OASIS Dataset
- NBTR Dataset

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.

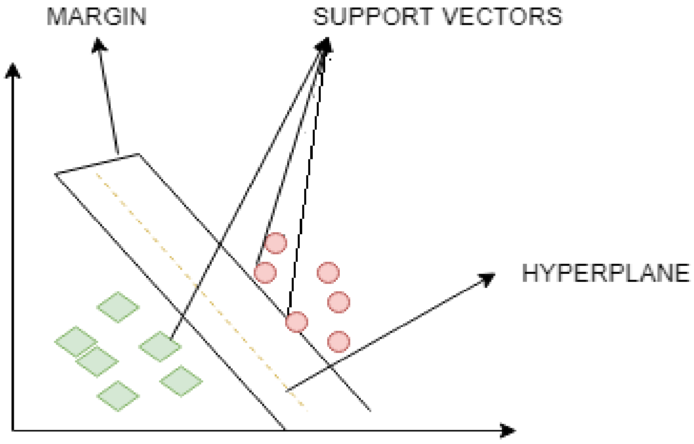

Fig.2. General Structure of SVM

B. Decision Tree

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) .Initially,entropy of parent is calculated. Then Information gain is calculated by subtracting weighted sum of entropy of children from entropy of parent. 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 symptom from the set $S = \{s_1,s_2,s_3\ldots s_j\}$ where S specify conditional attributes, $v_{i,k}$denotes the values of each branch i.e. the h-th range for i-th symptom and leaves which present decisions $D=\{d_1,d_2,\ldots d_k\}$ and their binary values, $w_{dk}=\{0,1\}$. By writing down each path from the root to the leaves ,a set of association rules was created by converting the decision tree .
Fig.3describe Decision tree as a set of association rule.

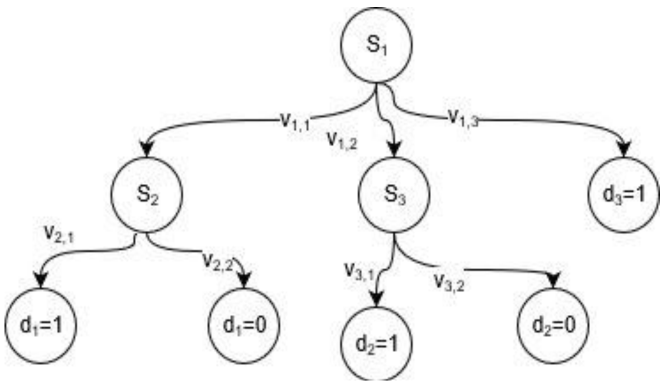

Fig.3. GENERAL VIEW OF DECISION TREE

The set of association rules for the above given tree are:

$$(S_1, v_{1,1}) \wedge (S_2, v_{2,1}) \Rightarrow (d_1 = 1)$$
$$(S_1, v_{1,3}) \wedge (S_2, v_{2,2}) \Rightarrow (d_1 = 0)$$
$$(S_1, v_{1,2}) \wedge (S_3, v_{3,3}) \Rightarrow (d_2 = 1)$$
$$(S_1, v_{1,2}) \wedge (S_3, v_{3,2}) \Rightarrow (d_2 = 0)$$
$$(S_1, v_{1,3}) \Rightarrow (d_3 = 1) \quad (1)$$

C. *Naive Bayes*

Naïve Bayes is mostly used in the area of Data Mining and Machine Learning.Takingadvantage of statistical methods SVM classification process is done. We used following equation to calculate the probabilities.

$$\frac{P(C1/X)}{P(C2/X)} > 1$$

$$\frac{P(C1/X)}{P(C2/X)} = \frac{\frac{P(X/C1).P(C1)}{P(X)}}{\frac{P(X/C2).P(C2)}{P(X)}} \quad (2)$$

Initially, inorder to decide which class the instance belongs to, probabilistic value is calculated. The final class label is the class with highest probability value. In figure 4 shown below ,there is a new incoming X and each of C1, C2, C3 labels represents the classes. According to probability values given in the figure, class C1 has the highest probability value and therefore the incoming X belongs to class C1.

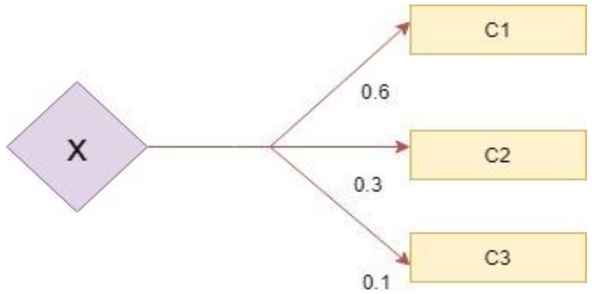

Fig .4.Example X data and Probability-Class Relations

D .*Logistic Regression*

Logistic Regression (LR), a popular mathematical modelling procedure used in the analysis of epidemiologic datasets, especially area of machine learning.

Logistic Regression method can be run in these steps:
1.Calculateusing logistic function.
2.Learn the coefficients for a logistic regression model.
3.Finally, make predictions using a logistic regression model.
The logistic function is given below:

$$f(x) = \frac{L}{1 + e^{-K(x-x_0)}} \quad (3)$$

E= Euler's number
x0=Middle x-value of sigmoid function
L= The maximum value of curve
K= Abruptness of curve.
Input values (x) to estimate an output value (y); logistic regression equation is used.
The logistic regression model is given in equation as:

$$y = \frac{e^{bo+b1*x}}{1 + e^{bo+b1*x}} \quad (4)$$

Logistic regression parameters are estimated by maximizing logarithmic likelihood function using training data. Fig 5 shows the example of a logistic regression to distinguish two classes (orange- yellow images).

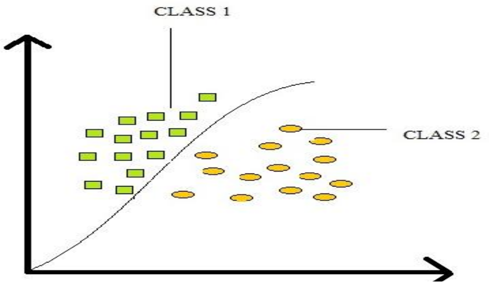

Fig.5. Logistic Regression to Distinguish two classes

EXPERIMENTAL RESULTS AND EVALUATION PERFORMANCE

The Datasets used in this study are taken from the UCI Machine Learning Repository and data.world.

UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/lung+cancer
In this dataset: Number of Instances: 32
,Number of Attributes: 57 (1 class attribute, 56 predictive)
Attribute Information: attribute 1 is the class label

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

In this dataset: Number of Instances:1000

Number of Attributes:25(1 class attribute,24 predictive)

Attribute Information: attribute 25 is the class label

.

Attribute Description of data.world is given below:

Proper classification of lung cancer detection is done using the effective utilisation of attributes in which the attributes represents the symptoms .The attribute such asAge, Gender, Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic Risk, Chronic Lung Disease, Balanced Diet, Obesity, Smoking, passive smoker, chest pain, coughing of blood, Fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of finger nails, Frequent Cold, Dry Cough, Snoring are taken into consideration for predicting lung cancer. 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:

Table 1:Lung Cancer Dataset:UCI Machine Learning Repository

No of Folds	**ML Algorithm**	**Accuracy(%)**
7	Logistic Regression	96.9
7	Decision Tree	85.71

Table 2:Lung Cancer Dataset:data.world

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

Table 1 shows that the performance of Logistic Regression exceeds the performance of Decision Tree, whereas Table 2 shows that the performance of SVM exceeds all other classification algorithms including Logistic Regression. So we can conclude that SVM has the highest accuracy rate among all other classification algorithms for these particular datasets.

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 sometimes a 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 a 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 are presented. The predictive performance of classifiers are compared quantitatively. In the performance chart, different results are produced for each classifier on the lung cancer dataset. 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 lung cancer detection can be done using this technique. Therefore, there is less mistakes. Finally, by adding extra pre-processing the accuracy rate can be enhanced.

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” ,Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 20167918045, pp. 661-666,2016