IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Carcinoma Prediction System Using Data Mining Algorithms And Machine Learning Techniques

MALATHI S[1],BALAMURUGAN M[2] and GOPAL K[3]

[1](ME Computer Science engineering/The Kavery Engineering College/India)

[2](Head of the Department/Computer Science Engineering/The Kavery Engineering College/India)

[3](Assistant Professor of Computer Science Engineering/The Kavery Enginering College/India)

Abstract -- The main objective of this paper is to predict the possible level of carcinoma. Today Cancer is the worst disease that causes a lot of death. Because in most of the cases, it is incurable. But it is not so the case, if it is detected at earlier stage. So the earlier diagnosis is necessary. However there are so many steps and efforts are needed to predict the occurrence of carcinoma. Data mining is the process of extracting the required information from the huge volume of datasets using various algorithms. In medical field, it leads to lot of researches. Here, we use Knn algorithms, Support vector Machine algorithms and decision tree algorithms that predicts the lung cancer based on simple parameters without much efforts.

keywords—Carcinoma, Data Mining, Support vector machine ,naive bayes,Decision tree

I. INTRODUCTION

Carcinoma is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. This growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. It is the number one cause of cancer deaths in both men and women in the U.S. and worldwide.The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath, chest pains etc. Treatment and long-term outcomes depend on the type of cancer, the stage (degree of spread), and the person's overall health. Most cases are not curable[2].

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns and relationships [3]. It is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems [4].In healthcare, data mining techniques have been widely applied indifferent applications including: modelling health outcomes and predicting patient outcomes, evaluation of treatment effectiveness, hospital ranking and infection control[3].

In this paper, we build several models to predict the level of lung cancer. The goal is to better understand which factors contribute to complications of this disease. The models are built by applying data mining techniques .The algorithms used here are Decision tree, Naïve bayes, Support vector machine algorithm .

II. LITERATURE REVIEW

Data mining has been used in various fields for detection, prediction and diagnosis. Isra Al-Turaiki, Mona Alshahrani, Tahani Almutairi has applied data mining algorithms to predict the stability and the possibility of recovery from MERS-CoV infections. They applied Naive Bayes and decision tree algorithms for prediction. For Recovery model Naive Bayes gave higher accuracy, whereas for Stability model Decision Tree is better in accuracy. The accuracy of the models is between 53.6% and 71.58% [3]. P . Bhuvaneswari , Dr. A. Brintha Therese had used Knn algorithm and genetic algorithm for the detection of lung cancer. In that paper, knn gave the highest accuracy of 90% in MATLAB[11].In Ada and Rajneet Kaur paper ,neural network and Support Vector Machine were applied to detect the lung cancer and classified as normal and abnormal based on health conditions. The detection of lung cancer in earlier stages will help the physicians in diagnosis[12].According to V.Krishnaiah , Dr.G.Narsimha and Dr.N.Subhash Chandra ,the most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. Decision Trees results are easier to read and interpret [5].

Vidya R, Latha V and Venkatesan S had applied Decision tree algorithm, neural network and naivebayes to predict the lung cancer based on Smoking and Non Smoking parameters. They proved that Naivebayes provides higher

accuracy of 83.4 [6].Supreet Kaur and Amanjot Kaur Grewal examined the potential use of classification based such as BFO, SUPPORT VECTOR MACHINE and Neural Network to massive volume of healthcare data [9].

III. METHODOLOGY:

Dataset Description:

The dataset for lung cancer was obtained from the data world. It contains 24 data attributes are Air pollution,Gender,Age,Dust allergy, Alcohol use, Genetic research, Obesity, Balanced Diet, Chronic lung disease Chest pain, Blood on Coughing, Fatigue, Weight Loss, Passive smoker, Shortness of breath, Wheezing, Smoking ,clubbing nails, Frequent cold, Swallowing difficulties, Dry cough, Snoring . It contains nominal data from 1-10 based on intense of that attributes. The class label is Level which contains the classification as high, medium and low based on these attributes. Out of 1000 instances, majority of the rows are provided as training data and the remaining is given as test data[1].

IV.DATA MINING

Data mining algorithms(both classification and clustering) are used in various domains. Here we apply Naive Bayes, Decision tree and Support Vector Machine algorithms to detect the level of lung cancer. The algorithms are

A. Naive Bayes: Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. It For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. Naive Bayes is a simple classification algorithm based on Thomas Bayes' conditional probability theorem. Everyone is aware that this algorithm is naive because it assumes that measurement features are independent of one another and contribute equally to the outcome.A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the colour, roundness, and diameter features [7].

B. Decision tree: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated. A decision tree consists of three types of nodes:

1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles

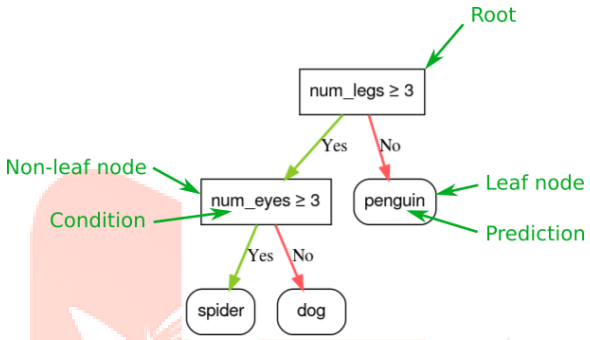

Figure 1.The Decision tree model

One of the best method of machine learining .

C. Support Vector Machine: Support Vector Machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data. An Support Vector Machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. Support Vector Machine can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [10].

D. K-Nearest Neighbor(KNN) Algorithm:
K-Nearest Neighbors (KNN) is a simple way to classify things by looking at what's nearby. Imagine a streaming service wants to predict if a new user is likely to cancel their subscription (churn) based on their age. They checks the ages of its existing users and whether they churned or stayed. If most of the "K" closest users in age of new user canceled their subscription KNN will predict the new user might churn too. The key idea is that users with similar

ages tend to have similar behaviors and KNN uses this closeness to make decisions.

V. EXPERIMENTAL RESULTS

The R tool is used for classification in our paper. It is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility.

The naive bayes algorithm provides 98% accuracy and its kappa value is 97% for true classified values. But it is reduced in false values. The obtained accuracy and kappa values are 89% and 83% respectively. It provides a stable model for detecting lung cancer. Now we discuss about decision tree algorithm. It provides a better accuracy of 95%, but it is less when compared with Naive Bayes. The fig. 1 is the obtained output. 'Coughing' is chosen as root attribute. Among the 24 attributes, these 5 attributes such as coughing, Air pollution, wheezing plays vital role in detecting lung cancer. It is observed that the accuracy rate is higher when the

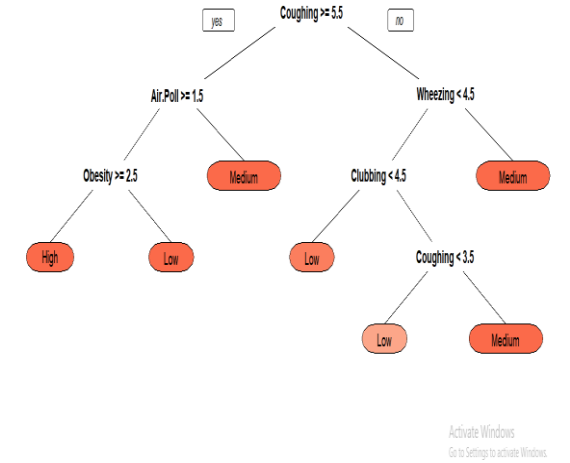

Figure 2.The Decision tree

Level is high and it is low when it is medium. The sensitivity value is also higher when the level is low. But the specificity value is more for the level classified as high. The kappa value obtained is 93% for decision tree algorithm.

The Support Vector Machine Algorithm provides the best prediction model .The accuracy obtained is 100% which is the highest one for Support Vector Machine non linear algorithm and it is 98% for linear Support Vector Machine algorithm. The objective of the SVM algorithm is to find a hyperplane that, to the best degree possible, separates data points of one class from those of another class. "Best" is defined as the hyperplane with the largest margin between the two classes, represented by plus versus
minus in the figure below..

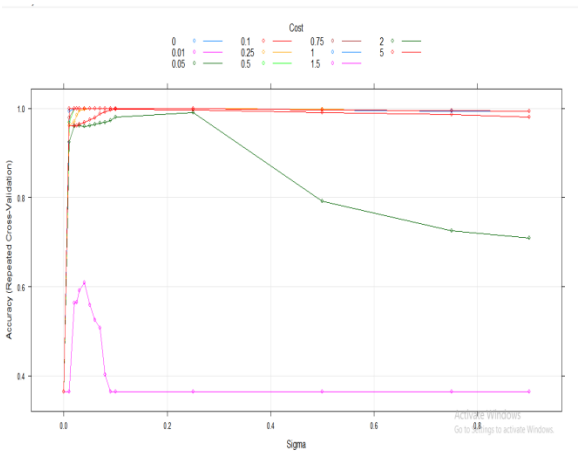

Figure 3. Support Vector Machine

It also provides the highest sensitivity and specificity values .The output graph with grid values of support vector machine is shown in fig. 2. The final sigma and c values used for model are 0.04 and 0.5 respectively and it gives highest accuracy and efficiency.

By organizing data into vectors, machine learning models can perform various operations on the data, such as clustering, classification, and regression. Vectors enable algorithms to leverage mathematical operations, such as calculating distances and similarities, to make predictions and learn patterns within the data.

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane.

VI. CONCLUSION

In this paper several data mining algorithms were applied for the detection of lung cancer. Naive bayes, decision tree and Support Vector Machine data mining algorithms are used to build the prediction models. Naive bayes provides 98% accuracy where as decision tree gives 95% accuracy which is the least among all. And finally Support Vector Machine provides the best accuracy of 100%.So therefore Support Vector Machine is the best prediction model to detect the carcinoma.

VII. REFERENCES

1. Lung Cancer dataset website .URL https://data.world/cancerdatahp/lung-cancer-data
2. Lung Cancer Definition website URL:https://en.wikipedia.org/wiki/Lung_cancer
3. Isra Al-Turaiki, Mona Alshahrani, Tahani Almutairi. Building predictive models for MERS-CoVinfections using data mining techniques. Journal of Infection and Public Health (2016) **9**, 744—748
4. DataMining definition Wikipedia URL:https://en.wikipedia.org/wiki/Data_mining
5. V.Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques . V. Krishnaiah et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1), 2013, 39 - 45
6. Vidya R, Latha V and Venkatesan S. Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques. Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com
7. Naive Bayes definition Website URL:https://en.wikipedia.org/wiki/Naive_Bayes_classifier
8. Decision Tree definition website URL:https://en.wikipedia.org/wiki/Decision_tree
9. Supreet Kaur, Amanjot Kaur Grewal. A REVIEW PAPER ON DATA MINING CLASSIFICATION TECHNIQUES FOR DETECTION OF LUNG CANCER. International Research Journal of Engineering and Technology(IRJET). Volume: 03 Issue: 11 | Nov- 2016.www.irjet.net . e-ISSN: 2395 -0056, p-ISSN: 2395-0072
10. P . Bhuvaneswari , Dr. A. Brintha Therese . Detection of Cancer in Lung With K-NN Classification Using Genetic Algorithm. 2nd International Conference on Nanomaterials and Technologies (CNT 2014). Procedia Materials Science 10 (2015) 433 – 440
11. R tool definition URL: https://www.r-project.org/about.html