Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200199

Prediction of Cancer and Suggestion of Therapies

Bhavani M [a,1], Pavithra V [b] and Monesh R [b]
[a] *Assistant Professor, CSE Dept, Rajalakshmi Engineering College, Chennai*
[b] *UG Scholar, CSE Dept, Rajalakshmi Engineering College, Chennai, India*

Abstract. Cancer is becoming one among the common diseases in day to today life, determining cancer in an earlier stage is still problematic. Identification of genetic and environmental factors is necessary to predict the type of cancer. The idea is to develop a cancer prediction system that predict lung and oral cancer depending on the symptoms. The gathered data is pre-processed and the data mining algorithm such as decision tree, logistic regression, Random Forest and Support Vector machines are used to measure the performance. The attribute selection algorithms are used to obtain the mandatory attributes. The main aim of this study is to do a comparative analysis using different algorithms for cancer prediction and suggestion of therapy.

Keywords. Cancer, logistic regression, Decision Tree, Random Forest, Support Vector, Data Mining

1. Introduction

Cancer is one of the dreadful diseases discovered in the world. It is the uncontrolled growth of the abnormal cell in the body by the rupturing of DNA [17]. When the DNA breaks and damages then it is not killed by our antibodies it grows as the abnormal cells and it is also known as malignant cells or tumors. In cancer, the cell division occurs multiple times. When a normal cell divides into hundreds and thousands and it kills the normal cells and leads to organ failure or death.

Data mining is the process of exploring and analyzing large amounts of data to gain meaningful trends and patterns. It is all about discovering previously unknown relationships among the dataset. Data mining is also known as Knowledge Discovery in Database which involves data cleaning, data integration, data transformation, data selection, data mining, pattern evaluation, and knowledge presentation.

The data mining technique associates the use of data analysis tools to detect previously unknown and valid patterns in large data set. In this study, to classify the data decision tree algorithm and logistic regression is used. A decision tree represents a tree like structure that includes a root node, branch nodes, and leaf nodes. The decision tree generate frequent patterns in the data. Attribute selection algorithms were used for data reduction to remove the unwanted attributes from the data. Logistic

[1] Bhavani M , Assistant Professor, CSE Dept, Rajalakshmi Engineering College, Chennai, India
E-mail: bhavani.m@rajalakshmi.edu.in

regression is a statistical analysis method used to solve classification problems. Data mining methods are implemented together to predict the existence of lung or oral cancer and the therapies for that cancer type based on the symptoms given by the user.

2. Review Of Literature

Ramachandran et al[1] developed a system that predicts various types of cancer such as lung, breast, blood, oral, cervix and stomach cancers using classification technology and clustering the cancer and non-cancer patients using k-means algorithm.

K. Arutchelvan et al[2] proposed a cancer prediction system based on data mining techniques. It determines the risk level based on the predicted value and also suggests the clinical and lab tests for the predicted cancer.

Deepika Verma and Dr. Nidhi Mishra[3] used a WEKA tool for analysis and prediction using five algorithms such as Naive Bayes, J48, MLP, SMO and REP Tree. They analyzed the performances of the five algorithms for the Diabetics and breast cancer and found that naive Bayes and SMO algorithms gave 72.7% accuracy and 76.8% accuracy on the breast cancer and diabetes dataset.

K. Sivakami[4] presented paperwork on analyzing breast cancer using DT-SVM Hybrid Model. Three classification techniques are compared and shown that Decision Tree-Support Vector Machines has the highest accuracy than Naive based classifiers, Instance-based learning and Sequential Minimal Optimization.

Krishnaiah, Dr. N. Subash Chandra et al[5] created a prototype that extracts the hidden knowledge from the database using classification techniques. The most effective model in predicting lung cancer appears to be Naive Bayes than Artificial Neural Network, Decision Tree and Rule-based algorithms.Shweta Kharya[6] have discussed various data mining approaches for breast cancer diagnosis and prognosis. He has suggested that among the various techniques the decision tree has given the highest accuracy of 93.62%.

Neelam Singh et al[7] has developed a system that uses the data collected from different centers to cluster the relevant and non-relevant data to cancer. He considered 20 risk factors for cancer assessment such as age, genetic Risk, environment, mental trauma, smoking, food habit, physical activity, obesity, tobacco, hypertension, heart disease, excessive alcohol and chronic lung diseases.N.V. Ramana Murty et al[8] have made a study to analyze lung cancer prediction using various classification algorithms. He took 32 instances and 57 attributes dataset and compared the results of various methods.

S. Muthuselvan, DR.K. Soma Sundaram et al[9] collected the blood test datasets for implementing data mining. The obtained data are preprocessed, then implanted using different data mining algorithms. From the executed algorithms J48 algorithm was best as a result of the correctly classified instances and low mean absolute error.

Subrato Bharathi, Mohammad Atikur Rahman and Prajoy Podder[10] have predicted the cancer using five classifiers that includes Naive Bayes, Random Forest, Logistic Regression and Multilayer Perceptron. They used 286 instances to predict and analyze breast cancer. They also have given some measures such as Kappa statistic, Root mean squared error, Mean absolute error, F-measure, ROC Area, Relative absolute error and Precision- Recall in which they observed K-nearest

neighbors classifier has highest percentage of ROC Area and Multilayer Perceptron has second-highest percentage of ROC Area.

B. Padmapriya and T. Velmurugan[11] used three most popular classification algorithms such as the J48 algorithm, CART Algorithm, and AD tree. They found that the performance of classification algorithms in analyzing Breast Cancer data through analyzing the mammogram images. The accuracy of taken algorithms was measured by various measures like kappa statistics, specificity and sensitivity.

Ankita Tyagi, Ritika Mehra and Aditya Saxena[12] suggested different machine learning techniques and diagnoses for the prevention of thyroid. Machine learning algorithms such as K- Nearest Neighbors, SVM, Decision Trees were used to predict the risk on a patient's chance of getting thyroid disease.

S M Halawani et al[13] suggested that probabilistic clustering behaved well than hierarchical clustering algorithms. The data points were clustered into a cluster, due to an irrelevant choice of distance measures.

Ada et al[14] attempted to identify the lung tumors from the cancer images and a tool was developed to check the normal and abnormal lungs to predict survival rate and years the abnormal patient can live so that patient's lives could be saved.Zakaria Suliman Zubi et al[15] used data mining techniques such as classification of lung cancers in X-ray aiming at identifying the characteristics that denote the group to which each case belongs and neural networks for detection.

3. Proposed System

The information for this study is collected from various cancer datasets that is combined together. The data is validated and preprocessed. The data consists of 34 attributes like age, gender, alcohol usage, obesity, smoking, cancer type, treatment, etc. Data validation ensures that the data is complete that is it does not contain any blank or NULL values. It ensures that the data is accurate during analysis. Data pre-processing involves converting raw data into an understandable format. The collected data might contain missing values that may cause inconsistency. In order to gain good results data need to be preprocessed that improves the efficiency of the data. The Attribute selection algorithms such as cfs Subset Eval, Principal Component, One R attribute Eval, Gain Ratio Attribute Eval, and Info Gain Attribute Eval were used to obtain the mandatory attributes in which principal component was found effective. The attributes were reduced from 34 attributes to 20 attributes. Data visualization is done on certain columns such as age, gender, treatment, etc. using scatter plot, Bar charts. In this module, the data mining algorithms logistic regression, Decision tree, Random Forest and support vector machines are applied to measure the performance. Precision, F-Measure, Recall and ROC area are calculated to find the performance measures. These comparative study is used to develop a cancer prediction system.

Table 1. Table of the compared algorithms

Algorithms	Precision	Recall	F-Measure	ROC Measure
Decision tree	99.2	98.9	99	99.3
Logistic regression	98	98.1	98	99.5
Support Vector	75.5	78.4	80.6	84.8
Random Forest	89.6	89.7	89.6	99.1

Table 1 shows the comparison of different algorithms with their performance measures.

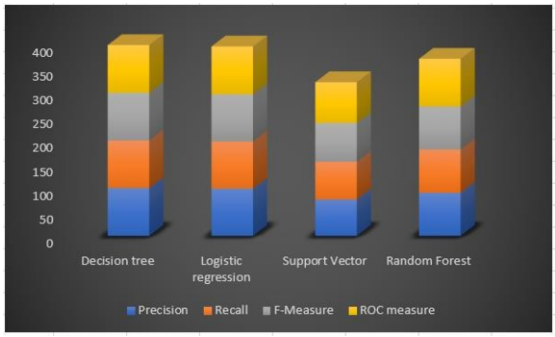

Figure 1. Chart of the algorithms compared

3.1. Advantage

1. The patients can check whether he/she has cancer without clinical tests which saves cost and time of the patients.

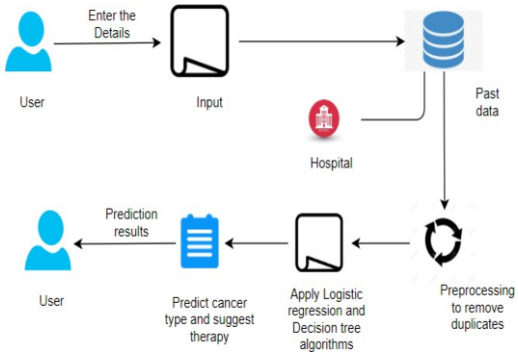

Figure 2. Architecture of the proposed work

The input is obtained from the user. The obtained data is compared with past data. The dataset is preprocessed and data mining algorithms logistic regression and decision tree are used. The result is provided to the user which shows the cancer type and the therapy suggested.

References

[1] P.Ramachandran, T.Bhuvaneswari and N.Girija, Early Detection and Prevention of Cancer using Data Mining Techniques .International Journal of Computer Applications, Volume 97– No.13, July 2014.

[2] K.Arutselvan and Dr.R.Periyasamy, Cancer Prediction Systems using datamining Techniques. International Research Journal of Engineering and Technology(IRJET) Volume: 02 Issue: 08 | Nov-2015.

[3] Dr. Nidhi Mishra and Deepika VermaAnalysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques.proceedings of the international conference on Intelligent sustainable systems(ICISS) 2017.

[4] K.sivakami, Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model.International journal of scientific engineering and applied science(IJSEAS)- August(2015).

[5] Dr.N.Subashchandra, Dr.G.Narsimha and V.Krishnaiah, .Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. International journal of computer science and information technologies(2013).

[6] Sewta kharya, Using datamining techniques for diagnosis and prognosis of cancer disease.International journal of computer science, engineering and information technology(IJCSEIT), April(2012).

[7] Santosh kumar singh and Neelam singh, Early Detection of Cancer Using Data Mining.International journal of applied mathematical sciences Volume 9(2016).

[8] Prof. M.S. Prasad Babu and N.V. Ramana Murty .A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis.International journal of computational intelligence research(2017).

[9] Dr.Prabasheela, S.Muthuselvan and Dr. K. Soma Sunadarm, Prediction of breast cancer using claasification Rule mining techniques in Blood test datasets.International conference on information communication and embedded system(ICICES) 2016.

[10] Prajoy podder, Mohammad atikur rahman and Subrato Baharathi, Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA.Fourth International conference on Electrical enginnering and Information technology(2018).

[11] B.Padmapriya and T.Velmurugan, "Classification Algorithm Based Analysis of Breast Cancer Data", International journal of data mining techniques and applications June(2016).

[12] Ankita tyagi, Ritika Mehra and Aditya saxena, .Interactive Thyroid Disease Prediction System Using Machine Leraning Technique,Fifth IEEE Internatinal conference on parallel, Distributed and Grid computing,Dec(2018).

[13] S M Halawani .A study of digital mammograms by using clustering algorithms. Journal of scientific and industrial resesrch, Sep(2012).

[14] Rajneet kaur and Ada, .Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient.International journal of computer science and mobile computing, April(2012).

[15] Zakaria Suliman zubi .Improves Treatment Programs of Lung Cancer using Data Mining Techniques Rajit Nair and Amit Bhagat, Feature Selection Method To Improve The Accuracy of Classification Algorithm. International journal of innovative technology and exploring engineering (IJITEE), April (2019).

[16] V.D.Ambeth Kumar and M.Ramakrishan (2013), Temple and Maternity Ward Security using FPRS. in the month of May for the Journal of Electrical Engineering & Technology (JEET) ,Vol. 8, No. 3, PP: 633-637.