International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-8, June 2020

Prognosis of Cancer and Proposition of Therapeutics

M. Bhavani, Sherine Glory, V. Pavithra, R. Monesh

Abstract*: Cancer is becoming one of the common diseases in day today life, identifying it in a prior stage is still difficult. Identification of environmental and genetic factors is necessary to predict the cancer. We developed a cancer prediction system to predict lung and oral cancer based 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 system is to predict the type of cancer and the suggested therapy using random forest algorithm.

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

I. 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. 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 patterns and trends. It is all about discovering previously unknown relationships among the dataset. Data mining is further known as Knowledge Discovery in Database . The data mining technology comprises the data analysis tools to recognize the previously unknown, 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 structure that includes a root node, branch nodes, and leaf nodes. Attribute selection algorithms were used for data reduction to remove the unwanted attributes from the data. Logistic regression is a statistical analysis method used to solve classification problems.

Revised Manuscript Received on June 30, 2020.
***** Correspondence Author
Ms. M. Bhavani*, Assistant Professor, Department of Computer Science & Engineering, Rajalakshmi Engineering College, Chennai, India.
Ms. Sherine Glory, Assistant Professor, Department of Computer Science & Engineering, Rajalakshmi Engineering College, Chennai, India.
Ms. Pavithra, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India.
Mr. R. Monesh, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India.

It is a predicting modelling algorithm used mostly for classification problem. Random forest consist of large number of decision tree. Each decision tree gives a predicted outcome. Based on the voting obtained, the outcome of random forest is predicted. 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.

II. REVIEW OF LITERATURE

Ramachandran et al[1] developed a system that predicts various types of cancer like lung, oral, cervix, etc., using classification technology and clustering the cancer and non-cancer patients using k-means algorithm. Those clusters were further divided into many clusters and then a prediction system was developed to predict the risk levels of the cancer which was used for prognosis.

K. Arutchelvan et al[2] based on data mining techniques proposed a cancer prediction system. It determines the risk level based on the predicted value and also suggests the clinical and lab tests. The predicted system determines the risk of lung, skin and breast cancer by examining number of factors like genetic and non-genetic.

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 naïve Bayes and SMO algorithms gave 72.7% accuracy and 76.8%acuuracy on the breast cancer and diabetics dataset. Irina Ionita and Liviu Ionita[4] using the KNIME Analytics platform and WEKA data mining software built and tested the classification models. The Romanian data were collected and the authors analyzed and compared those data using four classification models. The study referred to two common thyroid dysfunctions in which they found decision tree gives the best classification rate.

V. 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, chronic lung diseases etc.

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 was preprocessed, then implanted using different algorithms and from those algorithms J48 algorithm was best as a result of the correctly classified instances and low mean absolute error.

Shilpi Shandilya and Chaitali Chandankhede[10] have presented a paper on the research study which uses the online and offline data to classify the cancer. The study is based on the fact that data mining technique can be used to fetch the symptoms from the large data available and with the help of those data fetched, a prediction system can be developed to classify the tumor into benign or malignant.

B. Padmapriya and T.Velmurugan[11] used the three most popular Classification algorithms such as the J48 Algorithm, CART Algorithm, and AD tree. They found the performance of breast cancer data through the mammogram images through classification methods. 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 the diagnoses for the prevention of thyroid using machine learning techniques. 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 showed better performance than hierarchical clustering algorithms. The data points were clustered into one cluster, due to an irrelevant choice of distance measures.

Ibrahim Akduman et al[14] have worked to investigate the design consideration for microwave breast scanner which is a preliminary design of low cost and low dielectric materials for matching the breast and antennas

Zakaria Suliman Zubi et al[15] used data mining techniques for classification of lung cancers through X-ray which aims at determining the characteristics that denote the group to which each case belongs and neural networks for detection.

III. PROPOSED SYSTEM

The developed cancer prediction system is preprocessed and then the algorithms are applied to yield the results. The information for this study is collected from various cancer datasets 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 is a method for checking the accuracy and quality of data. Data

validation ensures that the given data does not contain any blank or NULL values. It ensures that the data is accurate during analysis. Data preprocessing involves converting raw data into an understandable format. The collected data might contain some missing values which lead to inconsistency. In order to gain better results data need to be preprocessed that improves the efficiency of the data. The Attribute selection algorithms such as cfs SubsetEval, Principal Component, OneR attributeEval, GainRatioAttribute Eval, and InfoGainAttributeEval 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, Recall, F-Measure and ROC area are calculated to find the performance measures. The cancer and the treatment are predicted by the random forest algorithm which provides the best accuracy. A GUI is presented to the user where they enter the required details and the system provides the user, the type of cancer and the suggested therapy based on the given symptoms.

Table 1. Table of the compared algorithms.

Algorithms	Precision	Recall	F-Measure	Roc Area
Random Forest	99.2	98.9	99	99.5
Decision Tree	98	98.1	98	99.5
Support Vector	75.5	78.4	80.6	84.8
Logistic Regression	89.6	89.7	89.6	99.1

In this table the four algorithms decision tree, logistic regression, support vector and random forest are compared with different performance measures.

Fig 1 . Chart of the algorithms compared.

Random Forest:

Random Forest is one the supervised machine learning algorithms which is used for regression and classification problems. It is a predicting modelling algorithm used mostly for classification problem. Random forest consist of large number of decision tree.

Each decision tree gives a predicted outcome. Based on the voting obtained, the outcome of random forest is predicted.

Algorithm: Pseudo code for the random forest algorithm

To generate *c* classifiers:
for *i=1* to *c* do
  Randomly sample the training data *D* with replacement to produce D_i
  Create a root node N_i containing D_i
  Call BuildTree(N_i)
end for
BuildTree(*N*):
if N contains instances of only one class then
  return
else
  Randomly select x% of the possible splitting features in *N*
  Select the feature *F* with the highest information gain to split on
  Create f child nodes of N , $N_1, \ldots, N_f$, where *F* has *f* possible values($F_1, \ldots, F_f$)
  for i=1 to f do
   set the contents of N_i to D_i is all instances in N that match
   F_i
   Call BuildTree(N_i)
  end for
end if

Fig 2. Pseudocode of random forest algorithm.

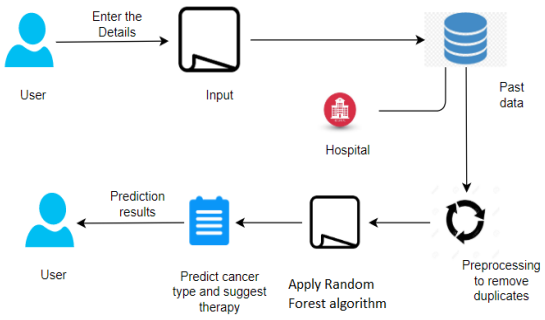

Fig 2. Architecture of the proposed work.

The input is obtained from the user. The obtained data is compared with past data. The dataset is preprocessed and data mining algorithms such as logistic regression, decision tree, support vector machines and random forest are used to measure the accuracy. The result is provided to the user based on the algorithm that showed best accuracy. The user is given the cancer type and the suggested therapy for the input provided.

IV. RESULTS AND DISCUSSION

Fig 3. Sample output for lung cancer

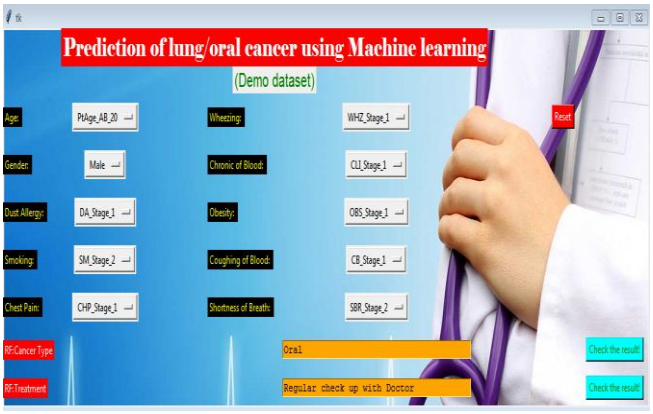

Fig 4. Sample output for oral cancer

V. CONCLUSION

Cancer is a fatal disease and detecting it in prior stage is mandatory. In this work, we have developed a cancer prediction system to predict the lung and oral cancer. The analysis is made using four data mining algorithms in which random forest gave the best accuracy based on the performance measures. The result is compared with prior patient records. The cancer type and the treatment is given to the user based on the input given by them.

REFERENCES

1. P.Ramachandran, N.Girija and T.Bhuvaneswari, " 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 | Nov2015.
3. Dr. Nidhi Mishra and Deepika Verma, "Analysis and prediction of breast cancer and Diabetics disease datasets using data mining classification Techniques", proceedings of the international conference on intelligent sustainable systems(ICISS) 2017.
4. Irina Ionita and Liviu Ionita, "Prediction of thyroid disease using data mining techniques", BRAIN. Broad research in artificial intelligence and neuroscience, August 2016.
5. Dr. N. Subhash Chandra, 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. Sweta Kharya, "Using data mining 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 lung cancer disease detection and diagnosis", International journal of computational intelligence research(2017).
9. Dr.Prabasheela, S.Muthuselvan and Dr.K.Somasundaram, "Prediction of breast cancer using classification rule mining techniques in blood test datasets", International conference on information communication and embedded system(ICICES) 2016.
10. Shilpi Shandilya and Chaitali Chandankhede, "Survey on recent cancer classification systems for cancer diagnosis", IEEE WiSPNET 2017 Conference.
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 learning technique", Fifth IEEE International 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 research, Sep(2012).
14. Ibrahim, Mehmet and Hulya, “A new multi-static system for microwave breast cancer imaging: Preliminary design”,IEEE 2018.
15. Zakaria Suliman zubi “Improves treatment programs of lung cancer using data mining techniques”, Journal of software engineering and applications, February 2014.
16. 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).

AUTHORS PROFILE

Ms. M.Bhavani is an Assistant Professor in the Department of Computer Science & Engineering at Rajalakshmi Engineering College, Chennai. She received her Bachelor’s in Information Technology and Masters in Computer Science & Engineering. Her main area of interests includes Data Mining and Machine Learning.

Ms. Sherine Glory is an Assistant Professor in the Department of Computer Science & Engineering at Rajalakshmi Engineering College, Chennai. She received her Bachelor’s in Information Technology and Masters in Computer Science & Engineering. She is an Active Member in Institution of Engineers and India Society of Technical Education. Her main area of interests includes Web Mining and Machine Learning.

Ms. Pavithra is pursuing bachelor’s degree in the department of computer science and engineering at Rajalakshmi Engineering College, Chennai.

Mr. R. Monesh is pursuing bachelor’s degree in the department of computer science and engineering at Rajalakshmi Engineering College, Chennai.