An Approach for the Cancer Prediction using Data Mining Techniques

G . Vivek Vardhan
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai , India

Katta Harendra
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai , India

P . Tharun
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai , India

C . Sabarinathan
Assistant Professor (OG)
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai , India

Abstract - Cancer is one of the most identified disease among people and it's one of the major reasons for increase in mortality rate. There are different types of cancers which are present such as lung cancer, breast cancer, blood cancer, etc. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. Data mining techniques contribute a lot in the development of such system. There are various algorithms and methodologies used for automated screening of cancer by segmenting and classifying cancer cells into different categories. For grouping different types of cancers we are using clustering techniques of data mining in which different types of cancers are clustered into different groups according to which the risks of each cancer is predicted group wise. For predicting the levels of risks we are using Decision tree algorithm and CNN algorithm. This study presents various data mining and machine learning algorithms integrated together to predict the types of cancer along with the overall risk levels of cancer according to the symptoms given by he users.

Keywords— Cancer Prediction , CNN Algorithm , Decision Key Algorithm

I. INTRODUCTION

According to WHO (2002) cancer was responsible for the deaths of millions of people worldwide with an unprecedented 50 percent rise for developing countries and 70 percent of overall cancer deaths. Developing nations have just 5 per cent of global funds for cancer prevention, according to previous studies, and very little human and material resources are available in these countries as well.The American Cancer Society (2008) describes cancer as a general term for a wide number of diseases that may affect any part of the body; malignant tumors and neoplasms are other names. For example, Breast cancer is a type of cancer which affects the breast tissue which is most commonly from the inner lining of milk ducts or the modules that supply the ducts with milk. Breast cancer is caused by a number of factors called risk factors; they are classified as modifiable or non – modifiable factors. Various researchers also suggested that smoking tobacco appears to increase the risk of cancer which is higher depending on how long the person has been smoking. Long term smokers have an increased risk of about 35% to 50%. The risk of cancer increases with an increased diet especially for those with fat diet, alcohol intake and obesity. Radiation exposure also increases the chances of cancer risk. Also, exposure to pesticides, chemicals and organic solvents are believed to increase cancer risks.
According to some researchers genetics is also believed to be the cause of 5% to 10% of cancer cases with those with none, one or two affected relatives with cancer respectively. Those with first degree relative with the disease face double the risk than a normal person. Classification is a data mining technique in which population or data points are divided into number of groups such that data point in one group are more similar to data points in the same group but dissimilar to data points in other groups. This study aims at using data mining techniques to classify cancer risks using data sets of patients' information which contains the risk factors and the cancer classes (unlikely, likely and benign). The decision trees and CNN classification of cancer was also performed.

II. LITERATURE SURVEY

Title : Recommendation of Attributes for Heart Disease Prediction using Correlation Measure.

Description : In general, filter and wrapper methods are being used for feature selection for predicting heart diseases. In filter methods where feature selection is independent of the prediction algorithm , different statistical factors such as information Gain, Chi-square test, Fisher Score, Correlation, LDA (Linear Discriminant Analysis) and ANOVA (Analysis of Variance) are used for finding relevancy. As wrapper methods are computationally very expensive, filter methods are frequently used in practice . Hence we made an investigation on research works which employ filter methods. From literature, several research works have used the thirteen attributes, namely, age, sex, chest pain type(cp), resting blood pressure (trestbps), serum cholesterol(chol), fasting blood pressure (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak exercise ST segment(slope), number of major vessel colored by flourosopy (ca) and thalassemia (thal) for prediction of heart diseases.

Title : Prediction of Breast Cancer Using Ensemble Learning .

Description : Data mining techniques to model the breast cancer data using decision trees to predict the presence of cancer. Data collected contained 699 instances (patient records) with 10 attributes and the output class as either benign or malignant. Input used contained sample code number, clump thickness, cell size and shape uniformity, cell growth and other results physical examination. The results of the supervised learning algorithm applied showed that the random tree algorithm had the highest accuracy of 100% and error rate of 0 while CART had the lowest accuracy with a value of 92.99% but naïve bayes' had the an accuracy of 97.42% with an error rate of 0.0258. The analysis involved the use of three random 500 records form the pre-processed data of 1183 and was used as training data and the lowest error rate achieved was 0.599. During the testing phase, the C4.5 classification rules were applied to a test sample and the algorithm showed had an accuracy of 92.2%, sensitivity of 46.66% and a specificity of 97.4%. Future enhancement of the work will require the improvisation of the C4.5 algorithm to improve classification rate to achieve greater accuracy.

Title : A Cancer Survival Prediction Method Based on Graph Convolutional Network.

Description : Both diagnostic and prognostic breast cancer data. The classification procedure adopted by them for diagnostic data is called Multi Surface Method – Tree (MSM – T) that uses a linear programming model to iteratively place a series of separating planes in the feature space of the examples. If the two sets of points are linearly separable, the first plane will be placed between them. If the sets are not linearly separable, MSM – T will construct a plane which minimizes the average distance of misclassified points to the plane, thus nearly minimizing the number of misclassified points. The procedure is recursively repeated. Moreover they have approached the prognostic data using Recurrence Surface Approximation (RSA) that uses linear programming to determine a linear combination of the input features which accurately predicts the Time – To – Recur (TTR) for a recurrent breast cancer case. The training separation and the prediction accuracy with the MSM – T approach was 97.3% and 97 % respectively whereas the RSA approach was able to give accurate prediction only for each individual patient. Their drawback was the inherent linearity of the predictive models.

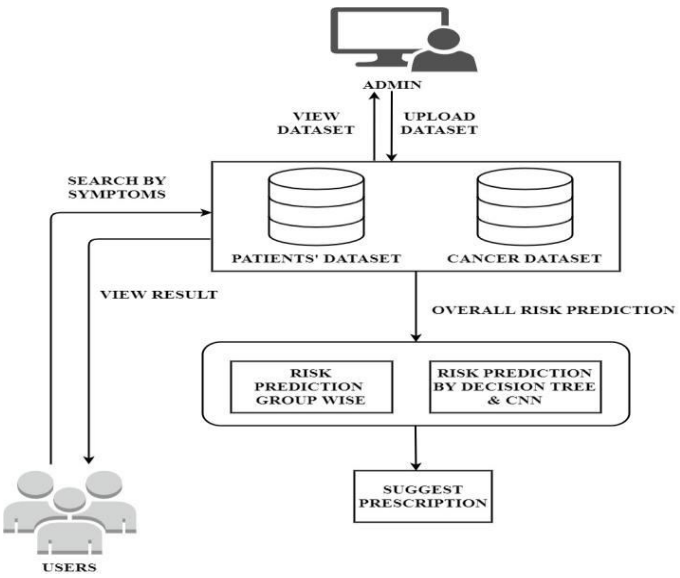
III. EXISTING SYSTEM

There are various ways to detect various types of cancer including Mammography, Magnetic Resonance Imaging (MRI) Scans, Computed Tomography (CT) Scans, Ultrasound, and Nuclear Imaging. Though, none of these aforementioned techniques gives a completely correct prediction of cancer. Tissue – based diagnosis is mainly done with a staining methodology. In This procedure elements of tissues are coloured by some staining element, usually hematoxylin and eosin (H&E). Cell structures, types, and other foreign elements are stained accordingly, and are easily visible under high resolution. Pathologists then examine the slide of stained tissues under a microscope or using high – resolution images taken from the camera.For detection of tumours, a histopathology test is essential. It is an old method used to predict invasive cancer cells from H&E stained tissues.

IV. PROPOSED SYSTEM

This paper introduces and assesses a data mining and machine learning techniques for automating the cancer prediction using the symptoms given by the user. We have described different Deep Neural Network architectures, such as Convolution Neural Networks (CNN). This used the labelled (benign/malignant) input from the dataset uploaded by the admin. After that he will collect all the cancer patient details and group the different types of cancer into different clusters using clustering algorithm of data mining. For example, the patients having lung cancer are grouped into one cluster and the patients having blood cancer are grouped into another cluster. According to which the risk level of the cancer is predicted. After that he can search the patients by the Id. The patient data is classified into two groups which are structured and unstructured. Structured data such as name, age, gender, etc. of the patients are classified using Decision tree algorithm, and unstructured data such as patient's BP level, Insulin level, number of cells affected, etc. are classified using CNN algorithm. Finally the overall risk level of the patient's cancer is predicted. Also the user can determine the type of cancer by providing their symptoms and can view the result which contains the risk level of their symptoms.

V. SYSTEM ARCHITECTURE



ALGORITHMS USED:

1. Convolutional Neural Network (CNN) Algorithm
2. Decision Tree Algorithm.

Convolutional Neural Network (CNN) Algorithm :

The term "convolutionary neural network" indicates the network is using a mathematical method called convolution. Convolution is a linear operation of a specific nature.

A CNN's hidden layers usually consist of a collection of convolutionary layers that coexist with a result of multiplication or other dots.

Since the layers are called convolutions in colloquial terms, that is by convention only. Technically, it is a moving dot element or a cross-correlation. It has significance for the matrix indices, in that it influences how weight is measured at a given index level. Usually, we start with low number of filters for low-level feature detection. The deeper we go into the CNN, the more filters we use to detect high-level features. Feature detection is based on 'scanning' the input with the filter of a given size and applying matrix computations in order to derive a feature map.

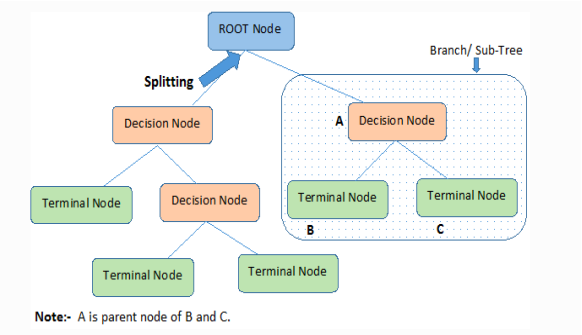
Decision Tree Algorithm:

Decision Tree algorithm is part of the supervised learning algorithms family. The decision tree algorithm can also be used to solve regression and classification problems, unlike other supervised learning algorithms.

The aim of using a Decision Tree is to build a training model which can be used to predict the class or value of the target variable by learning basic rules for decision rules inferred from prior data (training data).

In Decision Trees we start from the root of the tree to predict a class mark for a record. We compare the root attribute values with the attribute of that record. We follow the branch corresponding to that value on the basis of comparison, and move to the next node. Decision Trees obey expression on the Sum of Product (SOP). Material Sum (SOP) is also known as Standard Disjunctive Form. For a class, each branch from the tree's root to a leaf node having the same class is a value conjunction (product), different branches ending up in that class form a disjunction (sum).

The key challenge in implementing the decision tree is defining which attributes are to be considered as the root node and each stage. Handling this is to be known as selection of attributes. We have different selection measures of attributes to classify the attribute that may be considered as the root note at each point.



VI . MODULES

1. ADMIN.
2. USER.

Admin:

- Login into his account.
- Upload dataset of patient record and cancer symptoms.
- View dataset.
- Predict risk group wise.
- Search by patient's Id.
- Predict overall risk level.
- Prepare Prescription.
- Logout.

User:

- Register themselves.
- Login into their account.
- Search the type of cancer by symptoms.
- View Results.
- Logout.

ATTRIBUTES :

In this work it is proposed to find and recommend a list of relevant attributes for different classifiers which yield high accuracy. Relevant features are determined using the given steps.

Step - 1 Rank the attributes according to correlation measure.

Step - 2 Perform classification of known data using three commonly used classifiers, namely, MLP, SMO and NB and compare the accuracy of different classifier models.

Step - 3 Recommend relevant features for the chosen classifiers based on accuracy.

To perform the above steps, three experiments have been conducted. It is proposed to use Cleveland dataset and Weka

3.6.9 tool in Windows 7 operating system. Data are collected from Cleveland database of UCI repository. UCI includes four different databases such as Cleveland (303), Hungarian (294), Switzerland (123), and Long Beach VA (200) for cancer disease prediction. This database contains 76 attributes. There class labels are integer, valued from 0 (no presence) to 4(presence). Among these four databases, Cleveland dataset has less number of missing values (only six records contains missing values) than the other datasets. So Cleveland database has been taken up for experiment work. Further the details of attributes of the dataset are given in Table.

S.no	Attribute	Value	Description
1.	Age	29 - 62	Age in years
2.	Sex	0 -male, 1 - female	Gender
3.	Cp	1 - typical angina; 2 - atypical angina; 3 - non-anginal pain; 4- asymptomatic	Chest pain type
4.	Trestbps	Numeric value(140 mm/hg)	Resting bp in mm/hg
5.	Chol	Numeric value(289 mg/dl)	Serum cholesterol in mg/dl.
6.	Fbs	1 - true; 0-false	Fasting bp>120 mg/dl
7.	Restecg	0-normal; 1-having ST-T; 2-hypertrophy	Resting electro cardiographic results.
8.	Thalach	140, 173	Maximum heart rate achieved.
9.	Exang	1-yes; 0-no	Exercise induced angina.
10.	Oldpeak	Numeric value	ST depression induced by exercise relative to rest.
11.	Slope	1 - unsloping; 2 - flat; 3 - downsloping	The slope of the peak exercise ST segment.
12.	Ca	0-3 vessels	No. Of major vessels colored by flouroscopy.
13.	Thal	3-normal; 6-fixed defect; 7-reversable defect	Thalassemia
14.	Num	0: <50% diameter narrowing 1: >50% diameter narrowing	Diagnosis of heart disease(angiographic disease status).

EXPERIMENTATION AND RESULTS:

There may be many attributes related to a given prediction problem. But not all the attributes have strong association with the prediction. Hence finding the relevant attributes for a given prediction problem is important. In this work, relevant attributes for heart disease prediction are determined using correlation measure. In order to find the weight or rank of these attributes an experiment has been conducted. In this experiment the correlation between each attribute and class label is found out. Attributes along with their correlation values are given in below Table. In order to determine which feature set produces optimal accuracy, second experiment is conducted with three popularly used classifiers, namely, NB, MLP and SMO. While doing the above experiment, attributes are added one by one up to 13 attributes by choosing the attribute with highest weight as the first attribute. Accuracy of these classifiers is computed for different feature sets as given in below Table.

S.no	Attribute	Rank
1	Thal	0.4862
2	Ca	0.4608
3	Exang	0.4368
4	Oldpeak	0.4307
5	Thalach	0.4217
6	Cp	0.3817
7	Slope	0.3564
8	Sex	0.2809
9	Age	0.2254
10	Restecg	0.1664
11	Trestbps	0.1449
12	Chol	0.0852
13	Fbs	0.0280

VI. CONCLUSION

In conjunction with more accurate diagnostics, AI has the potential to bring down the cost of unwanted interventions for cervical cancer screening. Early detection will promise a greater rate of patients' prognosis especially in case of non – invasive cancer. Our paper discussed above made use of independent data sources, consequently a base for comparing

algorithms on a single scale was hard to define. CNN (Convolutional Neural Network) has proved to yield highest accuracy for classifying cancer cells. CNN's can predict with greater accuracy because they considerably reduce data – dimensionality, thus the computational overheads. Upon analysis of accuracy of the machine learning algorithms, it can be inferred that CNN can give maximal accuracy for cancer cell classification.

A graph convolutional network-based cancer survival prediction method GCGCN integrating multiple genomic data and clinical data was proposed in this paper, where multiple genomic data included gene expression, copy number alteration, DNA methylation and exon expression. First of all, multiple genomic data and clinical data were integrated using the similarity network fusion algorithm, sample similarity matrix was obtained, cancer survival related features were extracted using min-redundancy max-relevance RMR feature selection algorithm, the influence of useless features was mitigated, and classification training and prediction were conducted through the graph convolutional network.

VII. REFERENCES

[1] Rajesh, K., Anand, S (2012). Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012. ISSN 2278-1021. http://www.ijarcee.com pg. 72 , 77.

[2] Shajahaan, S.S; Shanthi, S., Chitra, V.M. (2013).Application of Data Mining Techniques to model Breast Cancer Data. International Journal of Emerging Technology and Advanced Engineering Vol. 3, Issue 11, November 2013. ISSN 2250-2459. http://www.ijetac.com pg 362 – 369.

[3] Mangasarian,D.S.;Street, W.N.,Wolberg, W.H (1995). Breast cancer diagnosis and prognosis via linear programming, Operations Research, 43(4), pages 570-577, July-August 1995.

[4] Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H.,(1999) "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.

[5] Delen, D., Walker, G., Kadam, A. (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, vol. 34, pp. 113-127, June 2005.

[6] V. Manikantan & S.Latha,"Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering, Volume-2, Issue-2, pp.5-10, 2013.

[7] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., ... & Bray, F. (2015). GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer; 2013.

[8] Frank, A., & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. School of information and computer science, 213, 2-2.

[9] J. Das, K.M.G., F. Bunea, M.H. Wegkamp,, H. Yu, ENCAPP:elastic-net-based prognosis prediction and biomarker discovery for human cancers,. BMC genomics, 2015. 16 p. 263.

[10] M. Khademi, N.S.N., Probabilistic graphical models and deep belief networks for prognosis of breast cancer, , in in: Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conferenceon, 2015. p. 727–732.

[11] Güler, I., & Übeyli, E. D. (2003). Detection of ophthalmic artery stenosis by least-mean squares backpropagation neural network. Computers in Biology and Medicine, 33(4), 333-343.

[12] Brenton, J.D., et al., Molecular classification and molecular forecasting of breast cancer: ready for clinical application? J Clin Oncol, 2005. 23(29): p. 7350-60.

[13] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, Nikahat Kazi – Breast cancer Diagnosis and Recurrence prediction using Machine learning Techniques,IJRET-International Journal of Research in Engineering and Technology, April 2015.