Analysis and Prediction of Cancer using Structured and Unstructured Data

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 analysis of this sickness physically takes extended periods of time and their less accessibility in frameworks, there will be process to build up the programmed finding framework for initial location in malignancy. Data mining methods gives a great deal in advancement of that framework. There are different calculations and approaches utilized for robotized screening of malignancy by portioning and ordering disease cells into various classes. 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 are 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 the users.

Keywords— Cancer Prediction, CNN Algorithm, Decision Key Algorithm.

I. INTRODUCTION

As per WHO (2002) Cancer was liable for the passing's of a huge number of individuals worldwide with a remarkable 50 percent ascend for developing nations and 70 percent of in general disease passing's. Creating countries have only 5 percent of worldwide assets for malignant growth counteraction, as indicated by past examinations, and next to no human and material assets are accessible in these nations too. The American Cancer Society (2008) portrays malignant growth as a general term for a wide number of sicknesses that may influence any piece of the body; dangerous tumors and neoplasms are different names. For instance, Breast malignancy is one kind of disease in which it influences the cancer tissue which is the most regularly in the form of internal coating of milk channels and the modules that will supply the pipes with milk. Bosom malignancy is brought about by various elements called chance variables; they are named modifiable or non – modifiable elements. Different scientists likewise proposed that smoking tobacco seems to expand the danger of disease which is higher relying upon to what extent the individual have been smoking. Haul smokers has an expanded danger of 38% with half. Their danger of malignancy increments with the expanded eating routine particularly for those with fat eating regimen, liquor admission and weight. Radiation introduction likewise expands the odds of malignant growth hazard. Additionally, presentation to pesticides, synthetics and to natural solvents are accepted to build disease dangers. As indicated by some specialist's hereditary qualities is likewise accepted to be the reason for 5% to 10% of disease cases with those with none, a couple of influenced family members with malignant growth separately. Those with first degree relative with the sickness face twofold the hazard than a typical individual. Order of the data mining procedure in which populace and information focuses are isolated into number of gatherings with the end goal that information point in one gathering are progressively like information focuses in a similar gathering however not at all like information focuses in different gatherings. This targets utilizing information mining methods to arrange malignant growth dangers utilizing informational indexes of patients' data which contains the hazard factors and the disease classes (far-fetched, likely and generous). The decision trees and CNN arrangement of malignant growth was likewise performed.

II. LITERATURE SURVEY

Title: Suggestion of the Attributes for Heart Disease Prediction utilizing Correlation Measure.

Description: By and large, channel and wrapper strategies are being utilized for highlight determination for anticipating heart maladies. In channel techniques where highlight determination is free of the forecast calculation, distinctive factual factors, Correlation, LDA (Linear Discriminant Analysis) and ANOVA (Analysis of Variance) are utilized for discovering pertinence. As wrapper techniques are computationally over the top expensive, channel strategies are every now and again utilized by and by. Subsequently, we made an examination on inquire about works which utilize channel techniques. From writing, a few research works have utilized the thirteen characteristics, in particular, age, gender , chest torment, rest circulatory strain (Rbps) , cholesterol , fasting pulse (ftbs), resting electro cardiographic outcomes (rest ecg), most extreme pulse accomplished (thalach), practice prompted angina (exang), ST sorrow incited by practice comparative with rest (old pinnacle), the slant of the pinnacle practice ST segment (slope), number of significant vessel shaded by fluoroscopy and thalasemia for forecast of heart diseases.

Title: Prediction of Breast Cancer Using Ensemble Learning

Description: Data mining systems to demonstrate the breast cancer malignancy information utilizing decision trees to anticipate the nearness of disease. Data gathered contained 699 occasions (quiet records) with 10 characteristics and the yield class as either kind or threatening. Information utilized containing example code numbered , bunch thickness, cell size. Shape consistency, cell development and different outcomes physical assessment. The aftereffects of the regulated learning calculation applied appeared the arbitrary tree calculation had the most elevated exactness of 100% and blunder pace of 0 while CART had the least precision with an estimation of 92.99% yet guileless Bayes' had the an exactness of 97.42% with a mistake pace of 0.0258. The examination included the utilization of three irregular 500 records structure the pre-prepared information of 1183 and was utilized as preparing information and the most reduced mistake rate accomplished was 0.599. During the testing stage, the C4.5 order rules were applied to a test and the calculation indicated had a precision of 92.2%, affectability of 46.66% and an explicitness of 97.4%. Future upgrade of the work will require the act of spontaneity of the C4.5 calculation to improve order rate to accomplish more prominent precision.

Title: Cancer precaution Method Based on GCN network.

Description: Both indicative and prognostic of breast cancer data. The arrangement technique received by them for analytic information is called Multi Surface Method – Tree (MSM – T) that utilizes a direct programming model to iteratively put a progression of isolating planes in the component space of the models. On the off chance that the two arrangements of focuses are directly divisible, the principal plane will be put between them. In the event that the sets are not straightly detachable, MSM – T will build a plane which limits the normal separation of misclassified focuses to the plane, subsequently about limiting the quantity of misclassified focuses. The system is recursively rehashed. Besides, they have moved toward the prognostic information utilizing Recurrence Surface Approximation (RSA) that utilizes direct programming to decide a straight mix of the info highlights which precisely predicts the Time – To – Recur (TTR) for a repetitive bosom disease case. The preparation division and the forecast exactness with the MSM – T approach was 97.3% and 97 % separately while the RSA approach had the option to give precise expectation just for every individual patient. Their disadvantage was the inherent linearity of the predictive models.

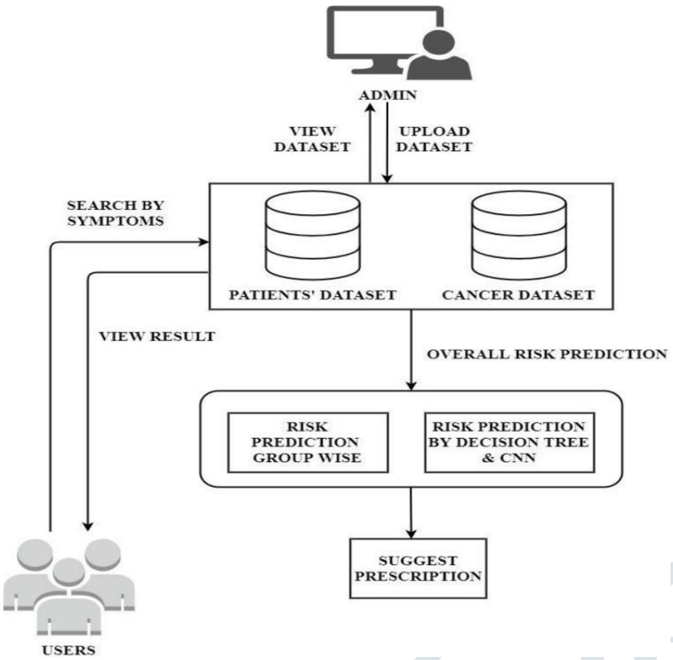
III. EXISTING SYSTEM

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&amp;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 tumors, a histopathology test is essential. It is an old method used to predict invasive cancer cells from H&amp;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. Generally, we start with low number of channels for low-level element location. The more profound we go into the CNN, the more channels we use to recognize significant level highlights. Highlight identification depends on 'checking' the contribution with the channel of a given size and applying network calculations so as to infer a component map.

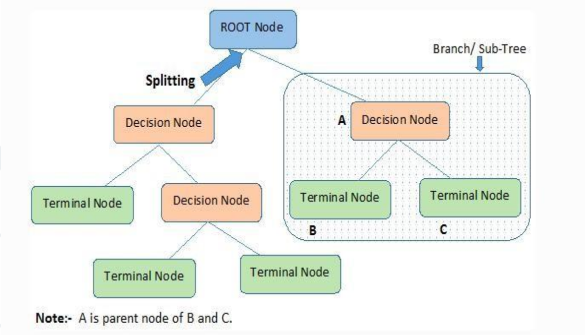
Decision Tree Algorithm:

Decision Tree algorithm is part of the regulated learning algorithms family. The decision tree algorithm can also be used to take care of relapse and characterization issues, unlike other supervised learning algorithms.

The point of utilizing a Decision Tree is to fabricate a preparation model which can be utilized to foresee the class or estimation of the objective variable by learning essential guidelines for choice principles deduced from earlier information (preparing information).

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 test in executing the decision tree is characterizing which credits are to be considered as the root hub and each stage. Dealing with this is to be known as choice of characteristics. We have diverse choice proportions of ascribes to arrange the trait that might be considered as the root note at each point.



Note:- A is parent node of B and C.

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 our paper work it is done to discover and to suggest a proper similar attributes for different classifications which has high level accuracy. Comparable highlights are resolved utilizing the below steps.

Step - 1 Line up the attributes as indicated by relationship measure.

Step - 2 Perform order of realized information utilizing three ordinarily utilized classifiers , and look at their exactness of different models using classification.

Step - 3 Suggest important highlights for their picked up classifications dependent on value.

To conduct their above advances, several assessments has been directed. It was proposed to use Cleveeland data set and Wekha.

3.6.8 tool in Windows 8 working framework. Information have gathered from Cleveeland data base of UCG vault. UCI incorporates five distinct data bases , for malignant growth illness expectation. This database contains 76 qualities. There class names are number, esteemed from 1 (with out nearness) to 5 (active). Of all five data bases, Cleveeland data set has small number of absent qualities (just six records contains missing qualities) than the different datasets. So Cleveeland data base have been taken for test work. Later the subtleties of traits of the data set are collected in our project.

No	Attributes	Values	Illustration
1	Age	28 - 65	Age group in years
2	Gender	0 -men , 1 - women	Gender (0,1)
3	Chp	1 - Normal angina; 2 - atypical angina; 3 - non anginal pain; 4- asymptomatic	Chest torment type
4	Rbps	Nume value (140 mm per hg)	Resting bp in mm per hg
5.	Cholesterol	Num value (288 mg per dl)	Serum cholesterol in mg per dl
6	Ftbs	1 - valid 0-false	Fast bp > 120 mg per dl
7	Rest ecg	0 - normal 1 - having ST-T 2- hypertrophy	Rest electro cardiographic outcomes
8	Thalache	141, 174	Max pulse accomplished
9	Exag	1 - yes 0 - no	Exercise induced angina.
10	Old peak	Num value	ST dep actuated to practise comparitive with resting

11	Level of Slope	1 - unsloping;
		2 - level;
		3 - down sloping
			The incline of the pinnacle practice ST section

12	Ca	0 - 3 vessels	No. of significant vessels coloured by flouroscopy
13	Thalasemia	3 - normal 6 - fixed deformity 7 - reversible defect	Thalasemia level
14	Numeriv Value	0 < 50% 1 > 50% Width narrowing	Width narrowing diagnosis of heart disease (angiographic sickness status).

EXPERIMENTATION AND RESULTS:

There might be numerous credits identified with a given expectation issue. However, not all the properties have solid relationship with the expectation. Subsequently finding the pertinent properties for a given expectation issue is significant. In this work, pertinent qualities for coronary illness expectation are resolved utilizing connection measure. So as to discover their weight or rank of their properties and analysis had been led. In the investigation the relationship between their every property and their class discovered. Properties alongside their connection esteems are given in beneath Table. So as to figure out which list of capabilities produces ideal exactness, second trial is led with three famously utilized classifiers, in particular NB , MLP and SMO.

No	Attributes	Ranks
1	Thalasemia	0.4872
2	Ca	0.4618
3	Exag	0.4378
4	Old peak	0.4317
5	Thalache	0.4227
6	Chp	0.3827
7	Slope Level	0.3574
8	Gender	0.2819
9	Age	0.2264
10	Rest ecg	0.1674
11	Rbps	0.1450
12	Cholesterol	0.0853
13	Ftbs	0.0281

While checking the before examination, traits added individually up to 14 qualities by picking their property with in most noteworthy load as primary characteristic. Precision of the classifiers are registered for the various capabilities as taken in beneath Table.

VII. CONCLUSION

Related to increasingly exact diagnostics, AI can possibly cut down the expense of undesirable mediations for cervical malignant growth screening. Early location will guarantee a more noteworthy pace of patients' visualization particularly if there should be an occurrence of non – intrusive malignant growth. Our paper examined above utilized free information sources, thus a base for looking at calculations on a solitary scale was difficult to characterize. CNN (Convolutional Neural Network) has demonstrated to yield most elevated precision for grouping malignant growth cells. CNN's can anticipate with more noteworthy precision since they significantly diminish information – dimensionality, along these lines the computational overheads. Endless supply of exactness of the AI calculations, it tends to be deduced that CNN can give maximal precision.

A diagram convolutional organize based malignancy endurance expectation technique GCGCN coordinating different genemic information and clinical information was proposed in our paper , here various geneic information included quality articulation, duplicate number modification, DNA melioration and axon articulation. Above all else, various genemic information and clinic information were to be coordinated utilizing likeness organize combination calculation, test closeness grid was acquired, malignant growth endurance related highlights were separated utilizing min-excess max-significance mRMR include choice calculation, the impact of futile highlights was alleviated, and characterization preparing and expectation were led through the chart convolutional arrange.

VIII. REFERENCES

[1] Rajesh, KN., Anaand, S (2013). Analysis of the SER data set 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] Shajahan, S.S; Shanti, S., Chithra, K.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 - December 2014.ISSN 2251 -2460. http://www.ijeatac.com page no: 363 – 368.

[3] Mangasaerian, D.K.;Stret, W.N.,Walberg, W.K (1996). Breast cancer diagnosis and prognosis via linear programming, Operations Research, 43(4), pages 570- 577, July-August 1995.

[4] Laundin M., Laundin J., BurkeeB.H.,Taoikkannen SI., 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] Delean, DJ., Waalker, G., Kadham, A. (2006) has Predicted breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, vol. 34, pp. 113-127, June 2005.

[6] V. Manikandhan & Latha S ,”Prediction and 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] Ferlaey, J., Sorjomataram, I., Eervik, M.Dikshitha, R., Eseer, 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] Fraank , A., & Asunnion, A. (2011). UIC in 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.K. Das, M.G.K., F. Buneae , M.G.H. Weagkamp ,, H. Yua , ENCAPP:elastic-net-based prognosis prediction and biomarker discovery for human cancers,. BMC genomics, 2015. 16 p. 263.

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

[11] Gueler, L., and Ubaeeyli, E. V. (2005). Likelihood of ophthalmic course stenosis by least - mean squares back propagational neural systems. PCs in Biology and Medicine, 33(4), 333-343.

[12] Breneton, J.D., Molecular order and guaging of the breast malignant growth , prepared for clinical application. J ClineOncol, 2005. 24(28): p. 7340-50.