International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 8 Issue 11 ǁ 2020 ǁ PP. 43-47

Breast Cancer Classification Using Data Mining

SRAVYA MANDADI
Software Engineer, BSH Hausgeräte, Bangalore, India

TEJASHWINI B.
Product Engineer, Talisma, Bangalore, India

SANJAN VIJAYAKUMAR
Data Scientist, eMotionRx, Boston, USA

ABSTRACT:
Medical professionals identified over 100 types of cancer, declaring it the world's most dire diseases worldwide killing over 8 million people every year. There are many treatment options that the patient can consider, including Chemotherapy, Radiation Therapy, surgery or Bone Marrow Transplant costing the infected more than 1.5 lakhs. Over 2 million people in India are infected with cancer and this number has seen a constant increase every year. The types of cancers are mostly gender specific, with rate of survival depending on the type of cancer and the stage it is in. One such cancer, Breast cancer is specifically common among women and has the lowest survival rate. In this paper, various data mining approaches which can be used for breast cancer diagnosis and prognosis have been discussed. Many times, breast cancer goes unnoticed and in order to avoid that, a system using data mining is presented. Diagnosing breast cancer involves distinguishing benign lumps from malignant ones whereas prognosis predicts whether the cancer will revert back to the patients even after being extirpated.

***KEY WORDS:* Breast cancer; Diagnosis; Prognosis; Data Mining; Classification.**

Date of Submission: 28-10-2020 Date of acceptance: 09-11-2020

I. INTRODUCTION

Statistics have shown that 1 in 8 women are prone to be infected with breast cancer in their lifetime[1]. Breast cancer is gender specific to women, and around 13% of women have lost their life due to this as seen in figure 1. In order to reduce the number of deaths due to breast cancer, the cells must be detected in an early stage. Physicians must be able to differentiate between non-harmful or early stage breast tumors from more serious ones without surgical biopsy. This can sometimes be difficult for the naked eye to see, and must require a precise and reliable diagnosis procedure. The goal of this diagnosis is to demarcate the patients who have noncancerous cells, in which case their tumor is non-threatening from the patients who have harmful or cancerous breast tumors.

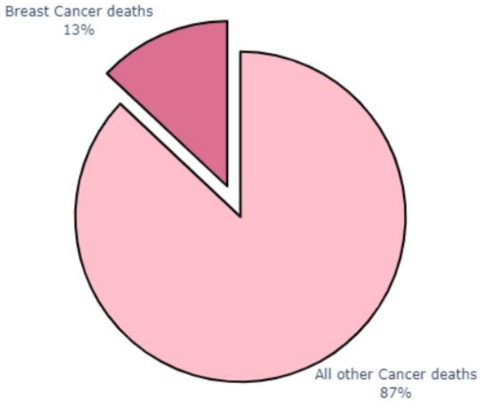

Figure 1: *Statistics of Breast Cancer deaths*

1.1 CLASSIFICATION

Prognosis is a prediction whether the cancer would relapse into the patient even after the tumor cells have been surgically removed[5]. This method can help one to know if there is a chance of recurring cancer cells on a long term basis. The objective of this is to classify the patients as a recurring patient, where the disease is observed even after the tumor excision and as non-recurring, where the tumor is not seen and is not predicted to relapse into the patient. The two forms of diagnosis are equally important and are thus, essential for breast cancer classification. Although this may be hard to detect using the naked eye, it can be developed using algorithms to surpass the efficiency of that of humans. Many computational methods have been used in order to classify these tumors, using Machine Learning, Computational Intelligence and even Image Processing.

Data mining is one such approach that can be used to differentiate these tumors, which involves the process of knowledge discovery in databases. This technology (KDD) is used in many computational methods and is used to train the algorithm and pick up patterns from the database intelligently. Data mining has one of the widest applications, and is frequently used in Health analytics. This technique can range from simple to challenging tasks, complete with the use of automated computing, myriad of data being collected and the data being widespread and available to the public for any medical research. This is why Knowledge Discovery in Databases (KDD) has become one of the most go-to techniques for any medical practitioner in order to identify and classify patterns and any common relationships between the data. This technology is largely used in medical practices since it can be used to predict the result of a disease when the past cases have been recorded in the dataset. Similarly, this can also be used to classify between the cases of breast cancer.

II. KNOWLEDGE DISCOVERY AND DATABASE (KDD) AND DATA MINING

Knowledge Discovery and Database is an essential part of data mining and analysis. This is used to convert unstructured data into meaningful information, and is one of the primary steps in any analysis. KDD refers to the process of identifying any recurring patterns that are found in any database system[5].

2.1 KNOWLEDGE DISCOVERY

On applying different analytics and algorithms on data, KDD extracts a number of patterns used throughout the data. This is extensively using in data mining and analysis where data is made sense of and converted into information, hence the name Knowledge Discovery.

There are a number of steps that are involved in the Knowledge Discovery process.

- **Data Collection**: Firstly, unstructured data is collected from different sources and concatenated to a single file. This step involves creation of a database from which the patterns are depicted.
- **Data Selection**: Next, the data is selectively chosen with respect to the requirements. Most data would not be of use and thus is exempted from the database, and only required data is left.

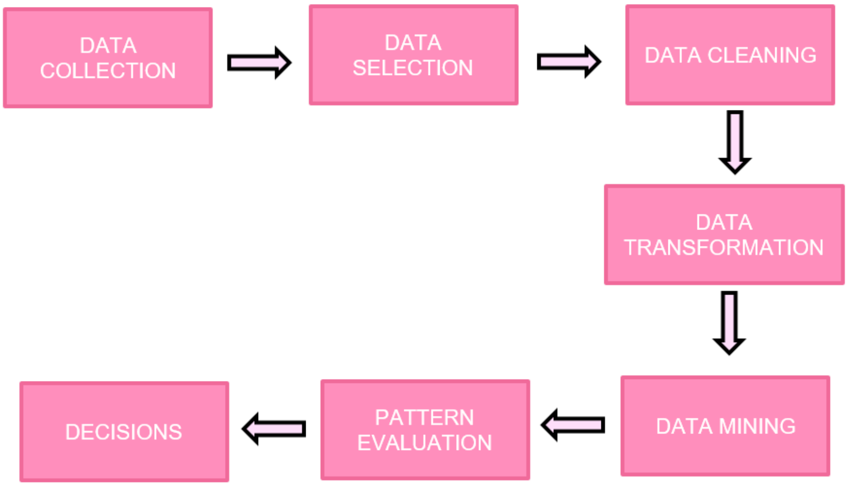

Figure 2: *Steps involved in KDD*

- **Data Cleaning**: Data Cleaning involves the application of different techniques to fix any bugs that the collected data might contain. This includes removing any corrupt or inaccurate data from the database by either replacing or deleting it entirely.

- **Data Transformation:** In this process, techniques such as smoothing, normalization and aggregation are applied in order to make the data compatible with the mining process. Even after cleaning, the data might not be in the required format of the mining technique, which is why transformation is required.
- **Data Mining:** Following transformation, the data is ready to be processed using data mining techniques. This is the most important step, here patterns are detected using schemes such as clustering and association analysis.
- **Pattern Evaluation:** This step is basically to convert the unstructured data into useful information using visualization and data transformation. This process consists of analyzing the data patterns found while mining and removing any recurring patterns that were detected.
- **Decisions:** This process mainly involves learning how to make use of the information acquired to make decisions.

2.2 DATA MINING PROCESS

Data mining is an essential step to retrieve information from the large unprocessed amount of data. The main steps of this data mining include converting this myriad of data into meaningful new information and forming some patterns. There are many tasks that can be applied for data mining, but mainly consist of 2 different types, predictive and descriptive methods. Predicative data mining, as the name suggests is used to perform predictive analysis on the data given using techniques such as classification, regression, etc. Descriptive data mining on the other hand, generally is used to describe the given data in any table or form. This is done using techniques such as clustering and summarization.

III. IMPLEMENTATION

The objective of this experiment is to use diagnosis to classify breast cancer, as well as prognosis to predict any future occurrences. The proposed system consists of using data mining techniques in order to recognize patterns in the data available, through which the system itself can identify the tumor by itself. Data mining software and techniques are thoroughly used in this system. The classification can be represented by using a data-tree which will depict all the possible classification factors as well. The system has been trained by using the data set from University of California, Irvine consisting of a large amount of unstructured data[9]. This is taken from the ML database and contains over 600 instances in order to train the algorithm. There are 10 attributes that are used to define the cancer cells including factors such as clump thickness, mitoses, shape uniformity etc.

3.1 DATA MINING SOFTWARE

There are multiple data mining softwares that are compatible with our end result such as WEKA, Kaggle, Rapid Miner etc. The proposed system has been implemented by using the former due to its easy accessibility and quality of user friendliness. Weka is inbuilt with most data mining functionalities such as feature selection, regression clustering etc. Containing a collection of visualization tool algorithms for data analysis and predictive analysis, Weka is one of the most widespread and essential software tools used for data mining. Weka is written in Java and can be easily associated with SQL and Java databases giving it an edge over other systems. Weka is advantageous because of its free availability, portability, descriptive collection of multiple data mining and processing tools all incorporated in one software and used GUI for user experience. Weka requires the processed data to be all stored in one file into an ARFF format.

3.2 CLASSIFICATION ATTRIBUTES

The dataset has 10 input factors and 2 outputs[9].

- Radius: The radius of the cancerous cell is measured by the average length of the exterior border of the cell to the center.
- Texture: Texture can be measured by firstly converting all values into gray scale intensities and further finding the variance and standard deviation.
- Perimeter: The total circumference of the cells is called the perimeter.
- Area: Nuclear area can be calculated by the sum of the pixels inside the cell and half the cells of the perimeter.
- Smoothness: This factor can be calculated by referencing a single radial line and finding the difference from the average of the radial lines around it. Similar to curvature energy, it is found using local variation in radial lengths.
- Concavity: Since all the shapes may not be a perfect symmetrical shape, this factor helps us approximate the actual shape of the body. Any dips between the radial lengths is calculated with respect to the nucleus. This measures the magnitude of each of the discrepancies in the cell.

- Concave Points:Instead of measuring the magnitude of the discrepancies, this factor calculates the number of deformities in the cell.
- Compactness: Compactness is indirectly proportional to the size of the cell. It can be calculated using the given formula:
-

$$\text{Compactness} = \frac{\text{perimeter}^2}{\text{area}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

- Symmetry: In order to find the symmetry of a cell, first the longest axial chord of the body is recognized. After this, the difference between each of the perpendicular axes is calculated.
- Fractal Dimension: Using smaller scales will give a more accurate measure of dimensions of the body. This is calculated by coastline approximation.

3.3 DATA TREE FOR CLASSIFICATION

The figure explains how the classification of benign and malignant cancer cells is implemented. There are 14 leaves and the size of the tree is 27. This dataset from University of California, Irvine has had an accuracy of 98% for this given data mining model. Out of 700 models in the data set, 686 were classified correctly and 14 were classified in correctly.

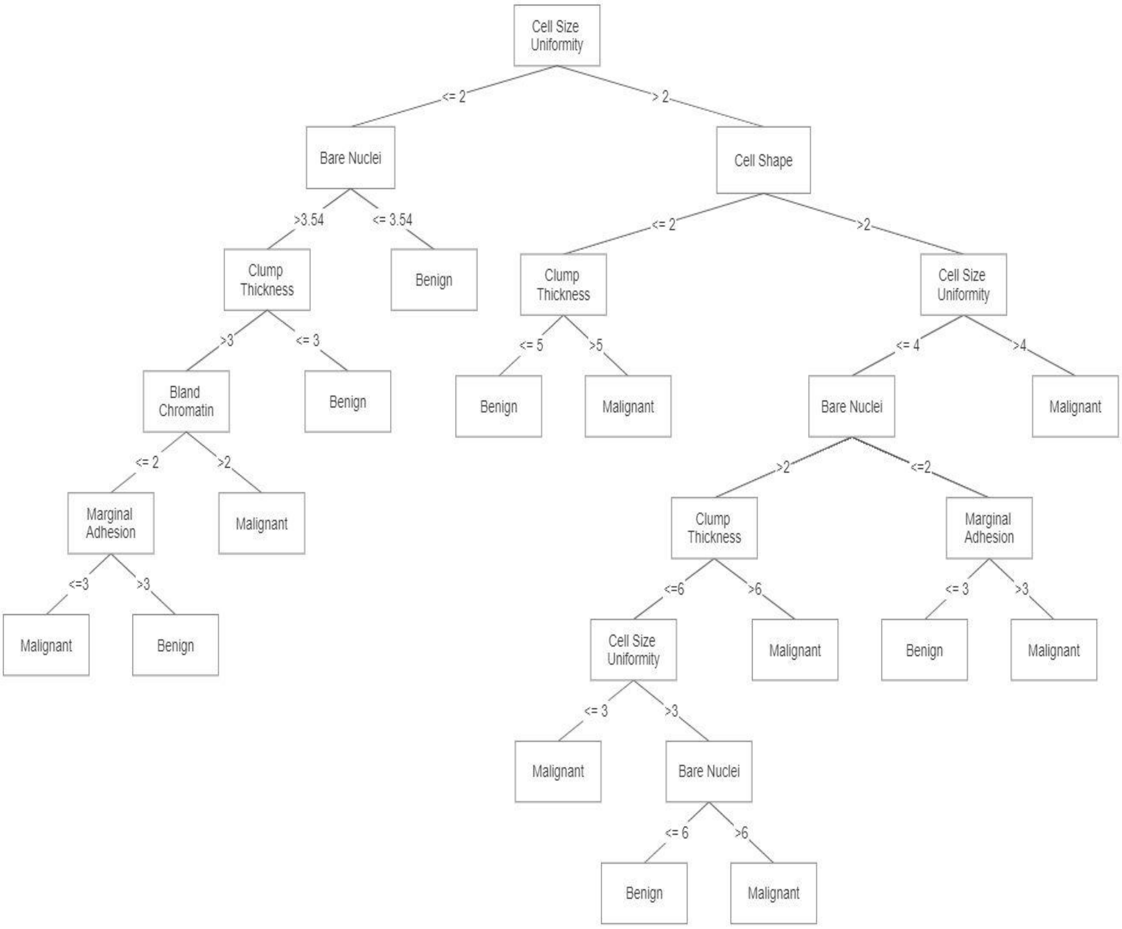

Figure 3: Classification Datatree

IV. CONCLUSION

In this paper, we have implemented data mining technology in order to classify and diagnose the cancerous cells to benign and malignant. These cases were dependent on 10 input parameters, which were defined in the dataset by UCI. During the data mining process, patterns were observed and from this a predictive model was employed. In the future, this data tree structure can be further expanded to classify reiterated cancer cells as well.

REFERENCES

[1]. Dr. K. Usha Rani, Parallel Approach for Diagnosis of Breast Cancer using Neural NetworkTechnique.

[2]. Chang Pin Wei and Liou Ming Der, (2010)―Comparision of three Data Mining techniqueswith Ginetic Algorithm in analysis of BreastCancer data‖.[Online].Available:http://www.ym.edu.tw/~dmliou/Paper/compar_threedata.pdf.

[3]. Quinlan J.(1996) Induction of decision trees.Mach Learn 1986; 1:81—106.

[4]. El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T., Azzedin F. and Al-Suhaim F., "Evaluation of breast cancertumor classification with unconstrained functional networks classifier," Computer Systems and Applications, IEEEInternational Conference, 2006, pp. 281 – 287.

[5]. BellaachiaAbdelghani and ErhanGuven(2006),"Predicting Breast Cancer Survivability usingData Mining Techniques," Ninth Workshop onMining Scientific and Engineering Datasets inconjunction with the Sixth SIAM InternationalConference on Data Mining,‖ 2006.

[6]. Anunciacao Orlando, Gomes C. Bruno, Vinga Susana, Gaspar Jorge, Oliveira L. Arlindo and RueffJose, "A Data Mining approach for detection of high-risk Breast Cancer groups," Advances in SoftComputing, vol. 74, pp. 43-51,2010.

[7]. Burke H. B. Et al , "Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction", Cancer, 1997,vol.79, pp.857-862.

[8]. Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. [Online]. Available:webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf

[9]. archive.ics.uci.edu/ml/datasets.html.

[10]. Abdelaal Ahmed Mohamed Medhat and FarouqWaelMuhamed (2010), ―Using data mining forassessing diagnosis of breast cancer,‖ in Proc.Internationalmulticonfrence on computer scienceand information Technology, 2010, pp. 11-17.

[11]. Maria-LuizaAntonie, Osmar R . Za¨iane, Alexandru Coma, .Application of Data Mining Techniquesfor Medical Image Classification.Proceeding of second International worshop on Mutimedia datamining(MDM/KDD'2001),in conjuction with ACM SIGKDD conference.SANFRANCISCO,USA,AUG 26,2001.

[12]. Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, MorganKauffmann Publishers,2001.

[13]. JyotirmayGadewadikar ,Ognjen Kuljaca1, KwabenaAgyepong, Erol Sarigul3, Yufeng Zheng andPing Zhang, Exploring Bayesian networks for medical decision support in breast cancer detection,African Journal of Mathematics and Computer Science Research Vol. 3(10), pp. 225-231, October2010.

[14]. Satyanandam N., Satyanarayana Ch.,Md.Riyazuddin,(2012) Data Mining MachineLearning Approaches and Medical DiagnoseSystems : A Survey ,International Journal ofComputer & Organization Trends –Volume2Issue3- 2012 ,PP 53-60 ISSN: 2249-2593 http://www.internationaljournalssrg.org.