

Detection of Breast Cancer Using Data Mining Algorithms

¹R.V.L. Manasa, ²M. Sai Krishna, ³M. Naveen, ⁴G. Abhishek, ⁵S.S.N.L. Priyanka

^{1,2,3,4}UG Student, ⁵Asst.Professor(Guide), Computer Science Engineering department, Anil

Neerukonda Institute of Technology and Sciences College, Visakhapatnam, Andhra Pradesh, India.

¹rvmnasa1335@gmail.com, ²machasaikrishna155@gmail.com, ³mnaveen.16.cse@anits.edu.in,

⁴gabhishek.15.cse@anits.edu.in, ⁵priyankaIT55@gmail.com

Abstract:- Over two decades, the most observed deadliest and dreadful disease is considered to be Cancer. The positive rate of cancer and the death rate over the years is rapidly increasing at an alarming rate. Among women, Breast cancer is the most diagnosed cancer. More than the treatment, the initial clinical examination of breast cancer itself is a very painful process for many patients. Even a small mistake can lead to false-negative and false-positive results that will be a burden on the life of a patient. To make this task easier and accurate we will be dealing with a novel approach by using technology. However, the present technology can make this painful clinical examination comparatively easier. This paper presents data mining algorithms like Decision tree, SVM which will be executed in the Spyder(anaconda) platform where the input is taken as values that differentiate the patient's record whether the cancer tissues are benign or malignant with higher probability based on the training dataset. This paper also summarizes the cons of a traditional breast cancer diagnosis.

Keywords - Benign, Cancer, Data Mining, Decision tree, Malignant, SVM, Spyder

I. INTRODUCTION

Thousands of females fall victim to breast cancer every year. The human body comprises millions of cells each with its unique function. When there is the unregulated growth of the cells it is termed as cancer. In this, cells divide and grow uncontrollably, forming an abnormal swelling tissue part called a tumor. Tumor cells grow and invade digestive, nervous and circulatory systems disrupting the bodies' normal functioning. Though every single tumor is not cancerous. Cancer is classified by the type of cell that is affected and more than 200 types of cancers are known. This paper is focused on Breast cancer. Breast cancer is the most common type of cancer among females across the world.

As per the National Breast Cancer Foundation [4], "Breast cancer is the most commonly diagnosed cancer in women". Breast cancer is the second largest cause of cancer death among women. Women generally approach clinics with a mild to serious pain in their breasts. After the examination of the breasts, the doctors usually suggest an ultrasound scan. The proceedings after the scan are more painful. Some women feel that the pain is intensely increased only after clinical proceedings. It is because of the painful process that is followed to detect whether the lymph node is malignant or benign. Malignant tumors are harmful or cancerous and

benign tumors are harmless and can be removed through surgery. The treatment is then decided after thoroughly examining the state of the tumor. With the help of this paper, the detection of the state of the tumor can be decided with the help of data mining algorithms.

II. TRADITIONAL CLINICAL PROCEDURE

The traditional clinical proceedings for breast cancer are painful. Lymph nodes found in women's breasts might not be always a malignant tumor or cancer.

A patient must undergo clinical examination by a breast surgeon then the result like x-ray mammography(Figure 1) an ultrasound scan(Figure 2) or MR mammography will be evaluated by a radiologist. If the report was suspicious, a needle biopsy is recommended followed by surgical removal of a lesion. The above procedure is very painful for women as they take a blood sample or a part of the lesion from the affected part. As it is most painful the pain stays for days to months. As per reports, diagnostic errors play a role in around 10% of patient deaths and breast cancer is no exception. Research says, "Overall, screening mammograms miss about 20% of breast cancers that are present at the time of screening because mammograms cannot find the affected area for all skin types. False-negative results can delay treatment and a false sense of security for affected women". On the other hand, false-positive results would let the

patient go through unwanted painful and expensive procedures.

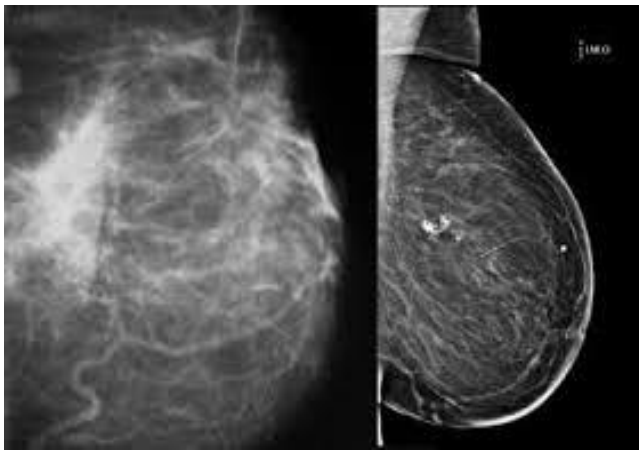


Figure 1: X-ray mammography



Figure 2: Ultrasound Scan

From analyzing the references [1], [10] we have got some facts for causing and checking breast cancer.

Some Causes of Breast Cancer:

1. Inheritance: - Inherited genes from the family having breast cancer can be a cause for the patient.
2. Menopause: - Beginning a period at younger (below 12) or older (after 55).
3. Radiation: - If a patient had radiation treatment on the chest at a younger age.
4. Alcohol: - Drinking alcohol can increase the rate of breast cancer.
5. LCIS: - If Lobular Carcinoma in Situ (LCIS) in the breast, it can lead to breast cancer.
6. Gender: - the probability is 100 times more in women than man

7. Hormone replacement: - Hormone replacement is done to relieve from menopause but it causes a risk of breast cancer

And some factors like aging, obesity, etc. can be factors of cancer.

Symptoms of Breast Cancer:

1. Pain: - Severe pain in the breast.
2. Breast Lumps:- Formation of painless breast lumps (hard mass with irregular edges but sometimes it can be soft).
3. Change in nipple i.e. retraction of nipple inward or discharge of nipple other than milk.
4. Change in the shape and size of the breast.
5. Change in the skin of the breast i.e puckering or dimpling of the skin.
6. Bloodstain discharge from the breast can also lead to breast cancer.
7. Crusting, peeling of the pigmented area of skin surrounding the areole or breast skin.
8. Causing small red holes over the skin of the breast.

These are the symptoms of breast cancer for self-diagnosis for a patient. If these symptoms are notified, it is better to contact the consulting doctor for further confirmation as early as possible. Neglecting these symptoms can cause risk for lives.

III. THE ARCHITECTURE OF BREAST CANCER CLASSIFIER

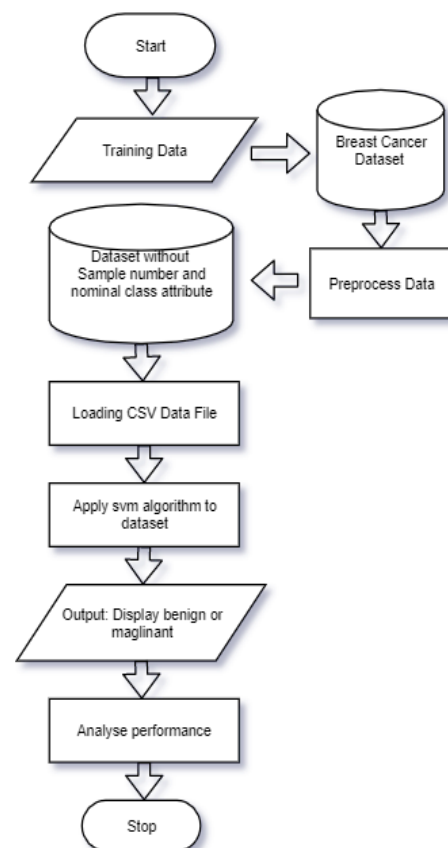


Figure 3: Architecture of Classifier

Process of executing the classifier:-

1. Collect the Wisconsin breast cancer dataset from the UCI repository and assign it as the training dataset.
2. The training dataset [7], [8] is preprocessed for applying the data mining algorithm(SVM) as:
 - i. The nominal class attribute i.e target class (benign or malignant) and sample no are removed from the training dataset.
 - ii. The target class is changed to binary number i.e '0' for benign and '1' for malignant.
3. Load the dataset in a CSV format for analysis using the anaconda(spyder) python libraries.
4. The accuracy of the given dataset is tested by supervised algorithms(Decision tree/SVM).
5. Find the algorithm having maximum accuracy for the following dataset.
6. The SVM algorithm is implemented on the dataset by forming an ideal two-dimensional hyperplane by divided the 2 target classes.
7. Analyzing for the new record apart from the dataset:
 - i. The values for required fields are taken as either in an individual value or as a query having single spaced between the values.
 - ii. The output is decided by calculating the record values distance from the support vectors to decide the class it belongs either '0' or '1'.
 - iii. The result is given on the web page along with the probability of the result and also displays the performance of the machine i.e. by giving the time taken for the analysis.

IV. TRAINING THE BREAST CANCER CLASSIFIER

The first part of the project is to train the classifier using the dataset [7], [8]. The dataset consists of variable values extracted from the scans like x-rays, ultrasound scans, mammographies. The scan reports and needle biopsy results are considered as training datasets into the machines. The training sets are thoroughly learned by the machines. The more the samples of datasets available, the more is the accuracy. Each sample is considered as an attribute with the unique identifier and the diagnosis information (malignant or benign). For each attribute, the cell nuclei are taken as main input which computes ten real-valued features.

The below table shows the readings:

S.no	Feature	Description
1	Radius	mean of distances from the center to points on the perimeter
2	Texture	The standard deviation of gray-scale values
3	Perimeter	In units (centimeters)
4	Area	In units (centimeters)
5	Clump Thickness	1 – 10
6	Class	2 for benign, 4 for malignant

7	Smoothness	local variation in radius lengths
8	Uniformity of Cell Size	1 – 10
9	Uniformity of Cell Shape	1 – 10
10	Single Epithelial Cell Size	1 – 10

Table 1. Input Features for Decision Making

The classification methods like decision trees, support vector machines are used in detecting the node whether it is cancerous or not. With improved accuracy, the result can assist radiologists in reviewing each mammogram, flagging potentially cancerous results that may have been missed for human review. As discovering breast cancer in its earliest stages saves lives.

V. ACCURACY TESTING

The second part of the project is accuracy testing for algorithms. Generally, for the selection of algorithms, the neural networks are best but as we are taking the extracted data values from scans than images [5] i.e x-ray, mammography. For implementing the classifier [9], we have analyzed dataset with 2 algorithms i.e. Decision tree and support vector machine as our output is binary prediction i.e. either benign or malignant.

Decision trees:

Decision Trees [2], [3] are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decisions and leaves. The leaves are the decisions or outcomes. And the decision nodes are where the data is split.

The benefits of having a decision tree are as follows: -

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

In the decision tree, the nodes are divided on the base of two parameters entropy and information gain formulas. The formula differs by the no of variables from the target class.

The drawback of Decision tree algorithm:-

- The accuracy of a Wisconsin breast cancer dataset is given to 94.5%.
- The time complexity depends on the depth of the tree hence more attributes lead to more run time complexity.
- The space complexity is more as we store more values.

Support Vector Machine (SVM):

SVM is a type of supervised machine learning algorithm. Where it is used for regression and classification analysis. In this algorithm, the dataset is represented by plotting in n-dimensional space. In our project, we have two output

classes i.e. benign and malignant. The value of each feature is given a specific coordinate then, we find an ideal hyperplane that differentiates between two classes.

The working principle of this algorithm is firstly we will find the closest points to both classes and these points are known as support vectors. Then we find proximity between the hyperplane and support vectors called margin. When the margin reaches to its maximum level, the hyperplane becomes the ideal one. In this project, we will be dealing with this algorithm as it gives us an accuracy of 98.01% for the given dataset [9].

Advantages of SVM:-

- The accuracy of a Wisconsin breast cancer dataset is 98.01%
- It is a decision-maker with a smaller dataset as it is memory efficient (stores support vector values).
- It requires less time for processing.

Drawbacks of SVM:-

- Overfitting occurs if no of variables > no of samples.
- Instead, the direct probability it uses expensive cross-validation for probability checking.

The accuracy of the result is dependent on large sets of data. However, for clinical purpose data analytics would need a huge amount of authentic patient data for deep analysis and creation of many different patterns for successful detection of tumors. Without the effort of patients dealing with multiple visits to the clinic to conclude whether the lymph node is cancer or benign tumor, this project can examine the breasts concluding whether the state is malignant or not. And further medication or treatment is suggested by the doctors. Using technology, women need not shy away to be directly touched by doctors and radiologists. The machine which is developed with Data Analysis will compare the values to the available dataset and finalize the category of the tumor.

VI. OUTPUT GENERATION

The final part of our project is to identify the target class value (benign or malignant tumor) for a new record apart from the training dataset records. For taking input of new records will be using python libraries i.e. flask app which will be acting as web interface between the web pages and database. The input given by the users is taken as individual values or as a query.

The input values are given to SVM classifier where it first pre-processes the machine by training dataset then the ideal hyperplane is created using support vectors by using svc classes imported from SVM. the input values are compared with both support vectors then the probability to be in each class is calculated if the probability of one class1 is greater than class2 the output class will be class1 otherwise class2. The result of the analysis will be shown on the Html page

including the target class value along with its probability and time taken for processing.

VII. RESULTS

INPUT 1:-

The samples from the scans are considered for learning purposes to the machines. The essential input features as discussed above are read and feed for each sample with the unique ID as shown in the below figure:-

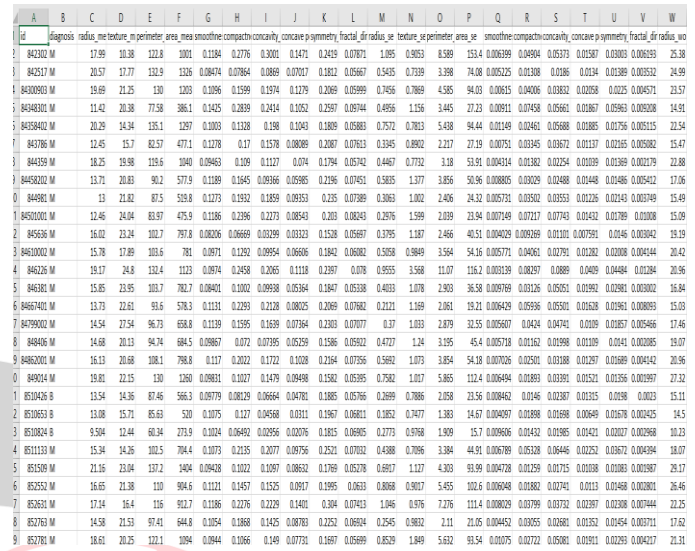


Figure 4: Training Dataset

For each sample all the features which are necessary to calculate whether the tumor is “BENIGN” or “MALIGNANT”. Where these samples are used for training the classifier.

INPUT 2:-

The second input is the record without the target attribute (cancerous or not) i.e. new record apart from training dataset where these record attribute values are taken as an input to the classifier which will decide whether the record belongs to the benign or malignant tumor.

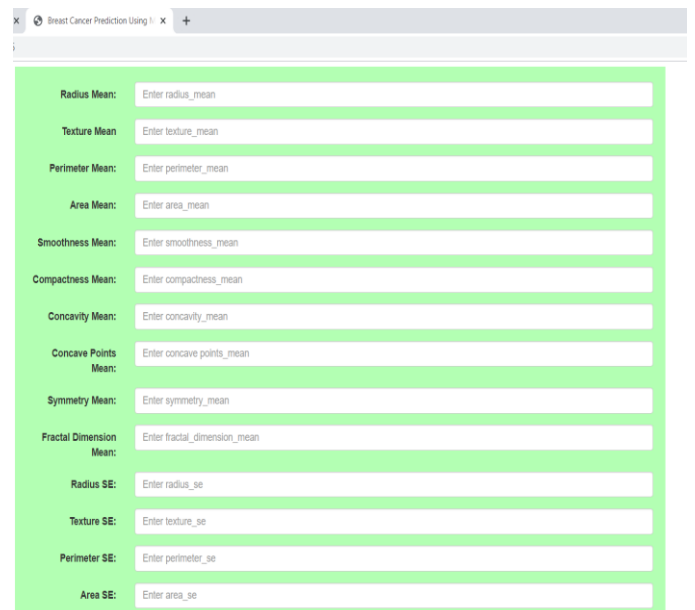


Figure 5: Input Page For entering the individual values

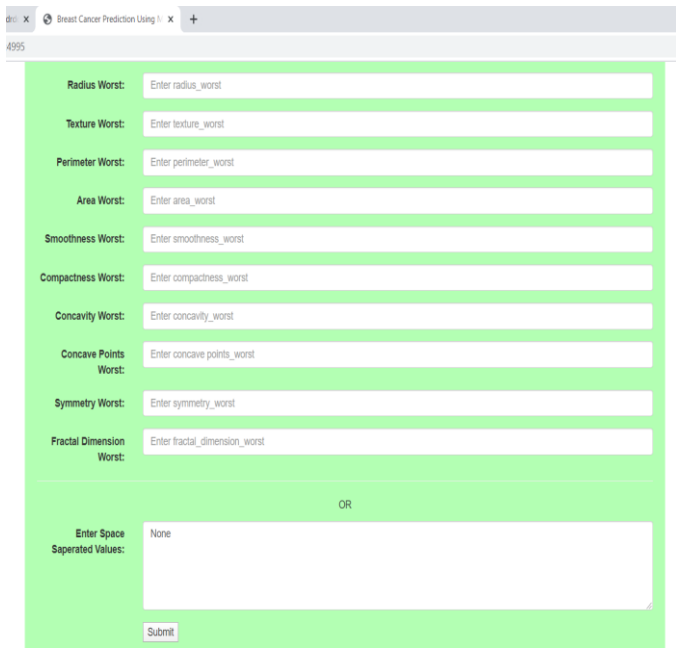


Figure 6: Input page for entering values as query

OUTPUT: -

The output is displayed whether the cancer is benign or malignant with the probability of the status by comparing it with the training dataset and the time taken for the prediction (analyzing the performance of the machine). If the result is benign the patient can undergo surgery for removal of the tumor otherwise if it is malignant the patient must undergo cancer treatment like chemotherapy, radiation therapy or removal of some or all part of the cancer tissue.



Figure 7: Result Page

VIII. CONCLUSION

Data Mining and Big Data are two emerging disruptive technologies that are changing the game in many sectors like e-commerce, agricultural sector, health sector, industrial sector, etc. By using these technologies can greatly reduce the effort of the patient as well as the health sector. In our project, we have dealt with the most common cancer i.e breast cancer where One in 100 women is diagnosed with either breast tumor or breast cancer. For survival and treatment of a patient is more expensive, as a helping hand for reduction of cost and increase the survival rate of the patient, early detection of cancer must be done to attain an immediate necessary treatment [6]. In our project we have focused on this issue, we have developed a machine with improved accuracy which could assist radiologists in reviewing each mammogram, flagging potentially cancerous results that may have otherwise been missed for human review. With flawless and wonderful technology, we have designed this project to have accurate results. And by the motto, “**Discovering cancer in its earliest stages saves lives and money**”.

IX. EXTENSION OF PROJECT

The future scope of this project is instead of dealing with data values extracted from scans as inputs, it can be extended by using the images i.e ultrasound scans, mammography and x-rays as inputs by using technology like image processing [5], neural networks (Convolutional NN) or Unsupervised learning algorithms. Even the additional feature can be added i.e using blockchain (SHA) for storing the new records of patients to increase the training dataset which increases the performance of the classifier.

REFERENCES

- [1] American Cancer Society <http://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html>
- [2] Sumbaly Ronak & N.Vishnusri & S.Jeyalatha “Diagnosis of Breast Cancer using Decision Tree Data Mining Technique,” Dept of computer science, BITS Pilani, Dubai, 2014.
- [3] Manish Mathuria “Decision Tree Analysis on J48 Algorithm for Data Mining,” Dept. of C.E. & I. T., Govt. Engineering College, Ajmer, India
- [4] National Cancer Institute for breast cancer <https://www.cancer.gov/types/breast>
- [5] conversion patch to image https://raw.githubusercontent.com/lisheh/end2end-all-conv/master/ddsm_train/Fig%20%20convert%20patch%20to%20whole%20image%20classifier.png

- [6] Forbes Innovation
<https://www.forbes.com/sites/suparnadutt/2018/08/21/how-this-indian-startup-is-using-ai-to-improve-detection-of-breast-cancer-at-a-low-cost/#197d64cdd3bc>
- [7] Wisconsin Breast Cancer Diagnostic
<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>
- [8] Wisconsin Breast Cancer Datasets
[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Original\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))
- [9] Building a machine learning model on breast cancer
<https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>
- [10] Mayo clinic on Breast Cancer symptoms
<https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470>

