International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 04 | April -2019 ISSN: 1847-9790 || p-ISSN: 2395-0126

Decision Tree and Random Forest for Breast Cancer Detection

Ms. Moram Lakshmi Rekha[1], Mr.G L N V Kumar[2], Mrs N Lakshmi[3], Mr.S Suresh[4]

Ms Moram Lakshmi Rekha, Asst.Prof,
Department Of MCA,
BVC Institute of Technology and Science,
Amalapuram, E.G.Dt.,AP
Email:rekhamoram@gmail.com

Mr GLNVS Kumar, Asso.Prof,
Department Of MCA,
BVC Institute of Technology and Science,
Amalapuram, E.G.Dt.,AP
Email:kumar4248@gmail.com

Mrs N Lakshmi, Asso.Prof,
Department Of MCA,
BVC Institute of Technology and Science,
Amalapuram,E.G.Dt.,AP
Email:n.laksmi27@gmail.com

Mr S Suresh, Asst.Prof,
Department Of CSE,
BVC Institute of Technology and Science,
Amalapuram,E.G.Dt.,AP
Email:sureshsappa@gmail.com

Abstract

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. This capability is particularly well-suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. More recently machine learning has been applied to cancer prognosis and prediction.

Breast cancer is the second leading cancer for women in developed countries including India. Many new cancer detection and treatment approaches were developed. The most effective way to reduce breast cancer deaths is detect it earlier. The frequent occurrence of breast cancer and its serious consequences have attracted worldwide attention in recent years. Problems such as low rate of accuracy and poor self-adaptability still exist in traditional diagnosis. In order to solve these problems, a Decision Tree classification algorithm and Random Forest algorithm is proposed in this project for the early diagnosis of breast cancer. The effectiveness of the proposed methods are examined by calculating its accuracy, confusion matrix which give important clues to the physicians for early diagnosis of breast cancer.

Key Words: *statistical, probabilistic, optimization prediction, Linear Regression, Vector Regression,, dataset, Detection, consequences*

1. INTRODUCTION

Cancer is not a single disease, but rather many related diseases that all involve uncontrolled cellular growth and reproduction. It is leading cause of death in the developed world and second in the developing world, killing almost 8 million people ayear. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. For better clinical decisions, it is important to accurately distinguish between benign and malignant tumours. Conventionally, statistical methods have been used for classification of high risk and low risk cancer, despite the complex interactions of high dimensional medical data.

To overcome the drawbacks of conventional statistical methods, more recently machine learning has been applied to cancer prognosis and prediction. Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard to discern patterns from large, noisy or complex data sets. This capability is particularly well suited to medical applications, especially those that depend on complex proteomic and genomic measurements. As a result, machine learning is frequently used in cancer diagnosis and detection. This latter approach is particularly interesting as it is part of a growing trend towards personalized, predictive medicine. In this review different types of machine learning methods being used, the types of data being integrated and the

performance of these methods in cancer prediction and prognosis. Now to predict whether it is malignant or benign, when we say benign that means that the tumor is kind of not spreading across the bodies of the patient is safe somehow. If it's malignant that means it's a cancerous

Currently, Mammograms are the most used test available, however, still, they have false positive (high-risk) results which shows abnormal cells that can lead to unnecessary biopsies and surgeries. Sometimes surgery is done to remove lesions reveals that it is benign which is not harmful. This means that the patient will go through unnecessary painful and expensive surgery.

Our task is to classify tumors into malignant or benign tumors using features of pain from several cell images. Let's take a look at the cancer diagnosis and classification process. So the first step in the cancer diagnosis process is to do what we call it final needle aspirate or if any process which is simply extracting some of the cells out of the tumor. And at that stage, we don't know if that human is malignant or benign. When you say malignant or benign as you guys can see these are kind of the images of the would be benign tumor and this is the malignant tumor. And when we say benign that means that the tumor is kind of not spreading across the bodies of the patient is safe somehow. It's if it's malignant that means it's a cancerous. That means we need to intervene and actually stop the cancer growth And what we do here in the machine learning aspect so now as we extracted all these images and we wanted to specify if that cancer out of these images is malignant or benign that's the whole idea. So what we do with that we extract out of these images some features when we see features that mean some characteristics out of the image such as radius, for example the cells such as texture perimeter area smoothness and so on. And then we feed all these features into kind of our machine learning model in a way which is kind of a brain in a way.

The idea is to teach the machine how to basically classify images or classify data and tell us OK if it's malignant or benign for example in this case without any human intervention which is going to change the model once the model is trained we're good to go we can use it in practice to classify new images as we move forward. And that's kind of the overall procedure or the cancer diagnosis procedure

1.1 RELATED WORK:

Decision Tree

The decision tree is an important algorithm for predictive modeling and can be used to visually and explicitly represent decisions. It is a graphical representation that makes use of branching methodology to exemplify all possible outcomes based on certain conditions. In decision tree internal node represents a test on the attribute, branch depicts the outcome and leaf represents decision made after computing attribute.

Types of decision tree are based on the type of target variable we have. It can be of two types:

1. **Binary Variable Decision Tree:** Decision Tree which has binary target variable then it called as Binary Variable Decision Tree. Example:- In above scenario of student problem, where the target variable was "Student will play cricket or not" i.e. YES or NO.
2. **Continuous Variable Decision Tree:** Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree.

Random Forest Regression:

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

1]Pick N random records from the dataset.
2]Build a decision tree based on these N records.
3]Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
4]In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

1.2 PREVIOUS WORK:

The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. For better clinical decisions, it is important to accurately distinguish between benign and malignant tumors. Conventionally, statistical methods have been used for classification of high risk and low risk cancer, despite the complex interactions of high-dimensional medical data. Most of the studies concentrated on mammogram images.

Mammography.:

The most important screening test for breast cancer is the mammogram. A mammogram is an X-ray of the breast. It can detect breast cancer up to two years before the tumor can be felt by you or your doctor.

1]Mammogram images sometimes have a risk of false detection that may endanger the patient's health.
2]It is vital to find alternative methods which are easier to implement and work with different data set.

2. Proposed functioning:

Machine learning has emerged as a promising technique for handling high dimensional data, with increasing application in clinical decision support. This system highlights new research directions and discusses main challenges related to machine learning approaches in cancer detection. This system proposes a hybrid model combined of several Machine Learning (ML) algorithms including Support Vector Machine (SVM), Decision Tree (DT) and Random Forest Classifier for effective breast cancer detection.

1]In Proposed system machine learning has emerged as a promising technique for handling high-dimensional data, with increasing application in clinical decision support.
2]It improves the accuracy of diagnosis. It is much more effective than the previously developed works
3]The proposed work can easily predicts by giving user inputs to the system whether the tumor is benign or malignant

2.1 System Architecture:

First Extract the dataset and then convert categorical data into numerical values then refining the data of features containing null values and then feature scaling. Now split the dataset into Training Data and Test Data which can be used to Train the System. Next step is select the model based on refined dataset for predicating the best results. Train the machine using selected model then predict the output for new input values

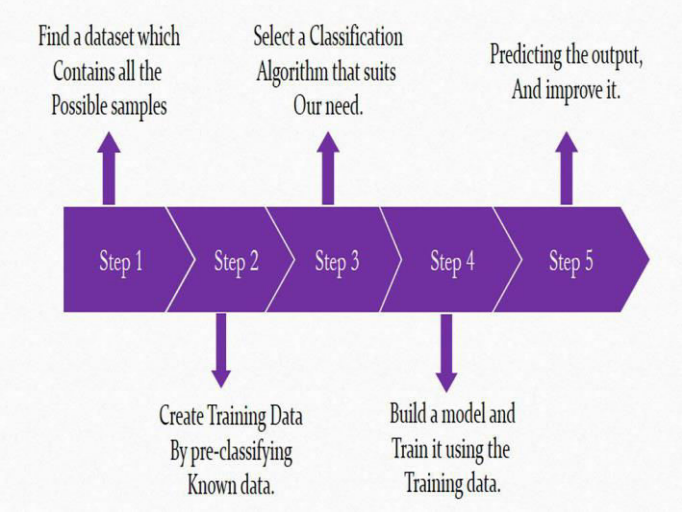


Fig: System architecture

Viewing of the dataset:

This is the first module functionality of Admin module in this system. Admin already upload the dataset which is used for prediction process.

When Admin want to view the already uploaded dataset the Admin may login into page. Then the Admin can view the uploaded dataset. Admin is only one person who can modify the dataset for different operations performed in that dataset for prediction.

Count the users:

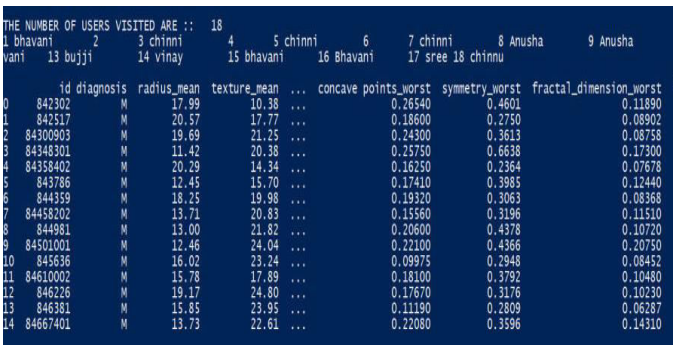
This is the second functionality of Admin module in this system. It will display how many users are used this system. For every user the count will be incremented by one. Admin can view the dataset as well as the count of user. How many users can used this system will maintained by this count the user

Prediction of Admission:

The second module in this system is Prediction of Diagnosis. User had an ability to perform operations such as preprocessing techniques, feature extraction, prediction on the dataset which is uploaded by admin. In this the user can view the predicted output after performing the algorithms on the dataset. It will display whether the patient is having breast cancer or not.

NAME	Bhavani
R-mean	13.54
P-mean	87.46
A-mean	566.3
Con-mean	0.06664
cpointsmean	0.04781
compact-se	0.014600
R-worst	15.11
P-worst	99.70
A-worst	711.2
Con-worst	0.2390

Table -1: Sample user input Data



```
THE NUMBER OF USERS VISITED ARE ::  18
1 bhavani    2      3 chinni      4      5 chinni      6      7 chinni      8 Anusha      9 Anusha
vani    13 bujji      14 vinay      15 bhavani      16 Bhavani      17 sree 18 chinnu

         id diagnosis  radius_mean  texture_mean  ...  concave points_worst  symmetry_worst  fractal_dimension_worst
0    842302         M        17.99         10.38  ...               0.26540          0.4601                  0.11890
1    842517         M        20.57         17.77  ...               0.18600          0.2750                  0.08902
2  84300903         M        19.69         21.25  ...               0.24300          0.3613                  0.08758
3  84348301         M        11.42         20.38  ...               0.25750          0.6638                  0.17300
4  84358402         M        20.29         14.34  ...               0.16250          0.2364                  0.07678
5    843786         M        12.45         15.70  ...               0.17410          0.3985                  0.12440
6    844359         M        18.25         19.98  ...               0.19320          0.3063                  0.08368
7  84458202         M        13.71         20.83  ...               0.15560          0.3196                  0.11510
8    844981         M        13.00         21.82  ...               0.20600          0.4378                  0.10720
9  84501001         M        12.46         24.04  ...               0.22100          0.4366                  0.20750
10   845636         M        16.02         23.24  ...               0.09975          0.2948                  0.08452
11 84610002         M        15.78         17.89  ...               0.18100          0.3792                  0.10480
12   846226         M        19.17         24.80  ...               0.17670          0.3176                  0.10230
13   846381         M        15.85         23.95  ...               0.11190          0.2809                  0.06287
14 84667401         M        13.73         22.61  ...               0.22080          0.3596                  0.14310
```

Table-2: predicted Output

3. CONCLUSIONS

In this algorithm was experimented on the Breast cancer dataset. The simulation results proved that the approach achieved a very high accuracy rate than the existing methods like mammograms and statistical methods. We also demonstrated a certain level of accuracy in the classifier, and for finding accurate results there must be sufficient preprocessing of data done. Missing data, data imbalance and other peculiar cases are to be considered in order to derive an accurate result. Finally we also demonstrated that we can attain accuracy in diagnosing breast cancer disease using the Random Forest Classifier, Decision Tree and Support Vector Machines.

REFERENCES

1]International Journal of Computer Applications Technology and Research Volume 7–Issue 01, 23-27, 2018,ISSN:-2319–8656.2.

2]http://www.imaginis.com/general-information-onbreast-cancer/what-is-breast-cancer3. Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4), 1054-1062.4.

3]https://www.safaribooksonline.com/library/view/da ta-science-for/9781449374273/ch04.html5.

4]https://medium.com/machine-learning-101/chapter2-svm-support-vector-machine-theory6.

f0812effc72Shrivastava, Shiv, Anjali Sant, and Ramesh Aharwa. "An Overview on Data Mining Approach on Breast Cancer Data." International Journal of Advanced Computer Research (2013): n. pag. Web.