International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-8, Issue-3, May 2020
https://doi.org/10.21276/ijircst.2020.8.3.4
www.ijircst.org

# Optimization of Random Forest Algorithm for Breast Cancer Detection

**Sarika Chaudhary, Yojna Arora, Neelam Yadav**

**ABSTRACT-** Today, cancer is a big issue and the most common disease all over the world. Cancer starts due to the abnormal growth of cells in your body. So, cancer takes place anywhere in the body. There are more than 100 types of cancers. The most common cancers are blood cancer, skin cancer, lung cancer, breast cancer etc. Nowadays, women can die because of breast cancer. There are several techniques like machine learning algorithms, big data & hadoop algorithms, and data mining algorithms to addressing breast cancer. Many techniques claim that their results were faster and more accurate. This paper presents an optimized random forest algorithm for cancer detection. Experimental results show that random forest gives the accuracy of 98.60%. All experiments are executed within anaconda package the scientific python/R development environment and spider software.

**KEYWORDS**- Anaconda, breast cancer, benign, malignant, machine learning, random forest algorithm, Spider**,** tumor.

## I. INTRODUCTION

Breast cancer is the second leading cause of women's death. Cancer is a malignant tumour in an organ that starts growing cells abnormally and crowded the malignant. Cells can do particular jobs in the body, normal cells can divide orderly and then die when damaged and new cells take place of the dead cells. If cancer cells spread, then it's hard for the body to work normally. Some symptoms of breast cancer are swelling on a particular area, skin dimpling, pain, swollen lymph nodes, skin turned red, and dry. The Indian Council of Medical Research (ICMR) recognizes, in 2016 India had 14 lakh cancer patients and the number of patients will be increased in the future**.** The ICMR report stated that till 2019 rate of diagnosis is 25.8 per one lack and the rate will be increased by 35 per one lakh women till 2029. The U.S., China and India are the most diagnosis countries all over the world. 1 in 28 women are likely to survive from Breast Cancer in their lifetime. Kerala, Tamil Nadu, Delhi have the highest rate for breast cancer 2000 new cases diagnosed every day, 1200 are detected on a later stage. On a later stage survival rate reduced by 4 to 17 times. Cost for late detection cases 1.5 to 2 times higher than the early-stage cases. In 2019, an estimated 286,600 new cases of breast cancer malignant will be diagnosed in the U.S. and also 62,930 new cases are benign. Approximately 62% of cases will diagnosis on early-stage, in that stage women can survive 99% and also the cost in that stage is less as compared to the later stage. In 2019, approximately 41,760 women die because of breast cancer in U.S. Men can also diagnosis breast cancer but it's a rare case. Only 1 in 1,000 can diagnose from breast cancer [1]. This year only 2,670 men were diagnosed with breast cancer in the U.S. and approximately 450 died. Overall 3.5 million breast cancer survivors present in U.S. There are many classifiers for prediction of breast cancer outcomes. This paper enhance the performance of random forest classifier which is the most influential machine learning algorithm. The rest of the paper is arranged as follows. Section 2 describes the literature review. Section 3 discusses the methodology. Section 4 shows related experimental setup. Section 5 describes the result and discussion followed by conclusion in Section 6.

**Manuscript received May 7, 2020**

**Sarika Chaudhary**, Assistant Professor, Department of Computer Science and Engineering, Amity University, Gurugram, INDIA, 9911193360, email: schaudhary@ggna.amity.edu

**Yojna Arora**, Assistant Professor, Department of Computer Science and Engineering, Amity University, Gurugram, INDIA,

**Neelam Yadav**, Department of Computer Science and Engineering, Amity University, Gurugram, INDIA,8010092441,

## I. LITERATURE REVIEW

Choosing the best classifier is one of the most important tasks in machine learning. Data scientists showed very good results by applying various algorithms on different medical dataset. Delen et al. Lu [2] works with 202,932 records of breast cancer patients, which can be categorised into two groups of malignant (93,273) and benign (109,659). The result of the prediction model is 93% accurate using random forest. Chandrasekar [3] implements the breast cancer prediction model using data mining techniques. The study aimed to develop accurate prediction models for breast cancer using neural network classification technique. The classification technique includes tree random forest. Data were collected from the WBCD dataset. The dataset contained 286 cancer patient information in which 201 of them were benign and 85 were malignant. The dataset described by 10 attributes such as

age, tumour size and class. The results shows that tree random classifier achieved accuracy of 98%. The experiment was analysed by WEKA software. Pietro Lio & Sarinder Kaur Dhillon[4] proposed random forest algorithm for predicted cancer diagnosis. They work with the dataset retrieved from the University Malaya Medical Centre with 8066 records. The dataset was then split into a training set (70%; 5646 records) and a testing set (30%, 2420 records) for the model evaluation using Random Forest and achieved the accuracy of 82.2%. Mamta Jadhav & Zeel Thakkar [5] performs Random forest by using WBCD dataset. They were splitting the dataset into training set(75%)and testing set(25%) and find the accuracy of 95%.Manisha Bahl &Regina Barzilay[6] describes Random forest classifier with dataset of 1006 cancer patients and divided the dataset as 671 training dataset and 335 testing dataset and achieved the accuracy of 97.4%. Cameron Wolfe [7] proposed random forest using sklearn breast cancer dataset of 569 patients. He removed the variables which were highly correlated (>0.9) and also using 50 decision trees. He achieved the 96.71% accurate result. Ashish [8] suggested random forest algorithm for breast cancer prediction model and achieved accuracy of result was 83%. Table 1 below shows the summary of literature review carried out in recent years.

Table 1: Summarized Literature Review

| Sr. No. | Author | Year | Technique | Findings |
|---|---|---|---|---|
| 1. | Delen et al. Lu[2] | 2005 | Random Forest | 93% |
| 2. | Chandrasekar [3] | 2013 | Random Forest | 98% |
| 3. | Pietro Lio & Sarinder Kaur Dhillon[4] | 2019 | Random Forest | 82.7% |
| 4. | Mamta Jadhav & Zeel Thakkar[5] | 2019 | Random Forest | 95% |
| 5. | Manisha Bahl &Regina Barzilay[6] | 2017 | Random Forest | 97.4% |
| 6. | Cameron Wolfe[7] | 2018 | Random Forest | 96.71% |
| 7. | Ashish[8] | 2016 | Random Forest | 83% |

## II. PROPOSED METHODOLOGY

In order to improve the performance of random forest classifier, the methodology as described below is proposed in Fig-1. This methodology will work efficiently with random forest algorithm. Fig.1 shows the flow diagram of proposed methodology.

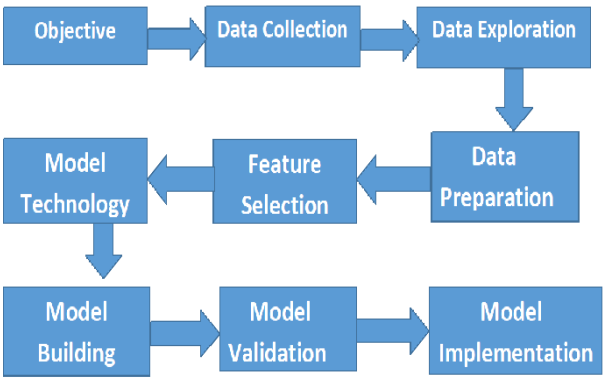



Fig. 1: Flow diagram of proposed methodology

1) Objective: The objective is to build a prediction model to identify breast cancer is benign or malignant.
2) Data Collection: Collection of medical dataset is carried out in this step.
3) Data Exploration: Exploring the data and check the quality of the dataset. Dataset may have some missing values.
4) Data Preparation: If missing data exists then we need to fill the missing values by using mean/median.
5) Feature Selection: In feature selection we are going to select the best variables. The variables which contain the higher value is the part of the model and the variable which has the value nearly 0, we will remove the variable.
6) Model Building: In this step, we have to start building our model by splitting the data into two parts as shown in Fig.2. Training data will be 70%-80% & testing data will be 20%-30%.

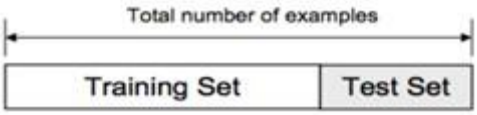



Fig.2: Data splitting

7) Model Technology: This is the most important phase where algorithm selection is done for the developing system. Data Scientists use different types of ML algorithms.ML algorithms are of two types: supervised learning and unsupervised learning. For this prediction model, we only need supervised learning.
8) Supervised Learning: In supervised learning the model building is based on the historic dataset. The input and output both required for supervised learning model. On the basis of the input and output, the model will learn how to process the future data and how the model will support future judgments.

### *A. Random Forest Algorithm*

Random forest classification is a technique in which the number of trees is higher than it will give the high accuracy results. Random forest algorithm can handle the missing data by itself. For this dataset, we have already handled missing values of attributes. If it includes many trees, then it doesn't over fit the model. This algorithm can use for both classification and the regression task.

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-8, Issue-3, May 2020
https://doi.org/10.21276/ijircst.2020.8.3.4
www.ijircst.org

**Steps of Random Forest Algorithm: -** In these steps, we will see how random forest will work with the dataset as mentioned in Fig-3.

- **Step 1**: Select the random samples from the dataset.
- **Step 2**: Each sample have a decision tress which will be constructed by the algorithm. Then the result will be predicted from every decision tree.
- **Step 3**: Voting will be performed by every training sample for the prediction result.
- **Step 4**: At the end, select the final prediction by choosing the prediction on average basis.

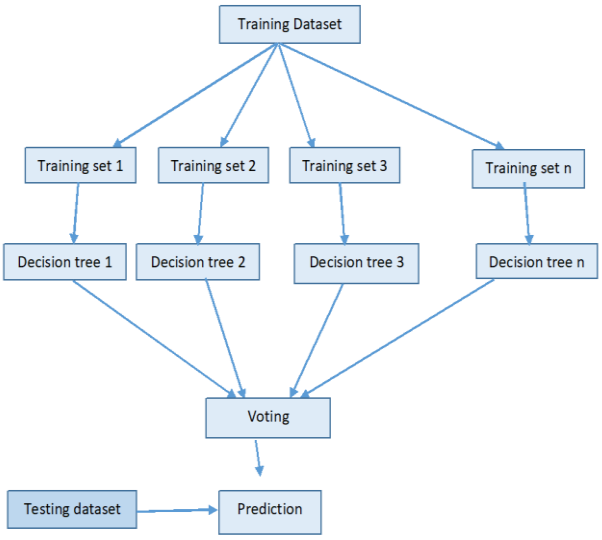



Fig.3: Random forest steps

### *B. Model Validation*

In this step checking the model will work properly with the testing data or have to train the model again. Apply confusion matrix find the accuracy as shown in fig.4.

1) True Positive (TP): Observation is positive, and is predicted to be positive.
2) False Negative (FN): Observation is positive, but is predicted negative.
3) True Negative (TN): Observation is negative, and is predicted to be negative.
4) False Positive (FP): Observation is negative, but is predicted positive.

| | | Prediction | |
|---|---|---|---|
| | | 0 | 1 |
| Actual | 0 | TN | FP |
| | 1 | FN | TP |

Fig.4: Confusion matrix

### *C. Model Implementation*

Final step is to implement the model.

## III. EXPERIMENTAL SETUP

In order to compare the performance of Random forest, we conducted an experiment that focuses on assessing both the effectiveness, and the efficiency of the algorithm.

### *A. Experiment Environment*

All experiments were executed using anaconda. Anaconda is a package which contains the software environment for implementing python and R. we are using spyder, with contains the random forest algorithm and also some important libraries such as pandas, numpy.

### *B. Breast cancer dataset*

The Wisconsin breast cancer (original) datasets from the UCI machine learning repository is used in this study. Breast-Cancer-Wisconsin has 569 instances (Benign: 357 Malignant: 212), 31 parameters such as radius_mean, Parameter_mean, texture_mean.

## IV. RESULTS AND DISCUSSION

This paper, implements a random forest classifier. Table 2 mention below shows the summary of the performance of the classifier. We take data of 569 cancer patients and also have 31 parameters like radius_mean, texture_mean, parameter_mean etc. After preparing our data, we will split data into testing data (75%) and training data (25%). Find the value of M & B in the data as shown in Fig.5.

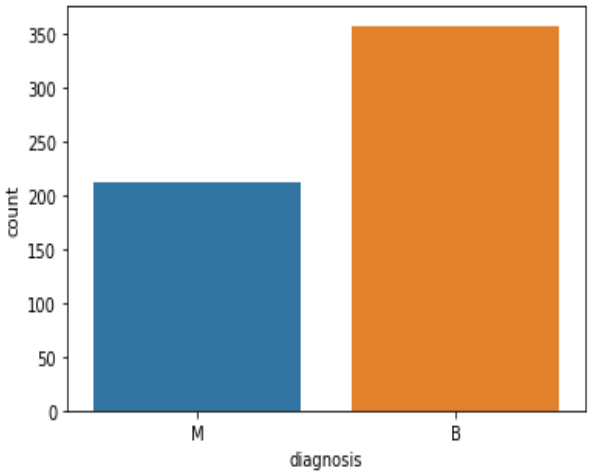
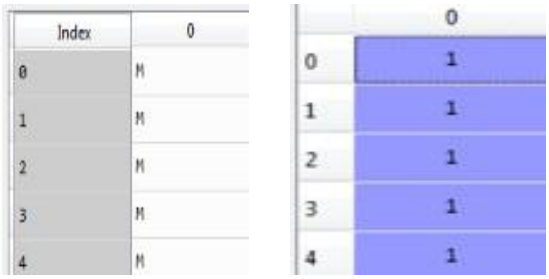



Fig.5: Graph displaying Malignant (cancerous) & Benign (non-cancerous) diagnosis

There are some variables which have the value near zero, so we will remove those variables like unnamed 32: have only null values in complete column and also removing the id column because id is a unique identity number so, it does not help in making the model well. May be this parameter will reduce our result. In this dataset, we had two categorical data which are B & M (benign & malignant). B further converted into numeric data as 0 and M as 1 as shown in Fig.6 because machine learning model doesn't recognize the character/string. It only understands the numbers.



| Index | 0 |
|---|---|
| 0 | M |
| 1 | M |
| 2 | M |
| 3 | M |
| 4 | M |

| | 0 |
|---|---|
| 0 | 1 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |

Fig.6: Covert M=1 & B=0

Implementing variance into the model and remove the parameters which have nearly zero value. Through that we will have the accuracy of 97.9% and the number of training set(n) is 12 and when n=16, accuracy will be 97.20. Applying the standard deviation, we will get the result accuracy 95.80%. When we apply Random forest only by removing id and null value it will give accuracy of 93.70% with n=2.In this way, we apply the random forest in different number of training set and find the accuracy of 98.6% with n=6 by using confusion matrix as shown in Fig.7 and table2 represents the classifier accuracy performance.

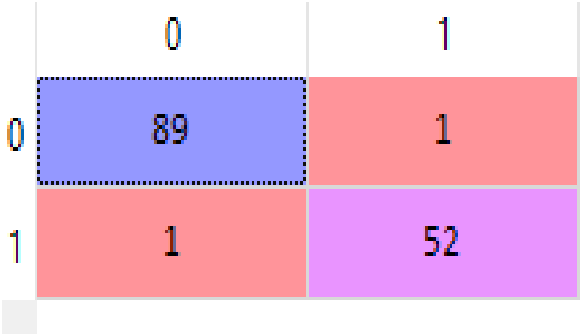



Fig.7: Resultant matrix

Table 2: Classifier Performance Accuracy

| Random Forest Classifier | | | |
|---|---|---|---|
| **Experiments** | **Incorrect variables** | **Number of training set** | **Result** |
| I. | id, unnamed 32: | 2 | 93.70% |
| II. | id, unnamed 32: | 4 | 95.10% |
| III. | id, unnamed 32: | 6 | 98.60% |
| IV. | id, unnamed 32: | 8 | 98.60% |
| V. | id, unnamed 32: | 10 | 98.60% |
| VI. | smoothness_se, fractal_dimension_se, id, unnamed 32: | 10 | 97.90% |
| VII. | smoothness_se, fractal_dimension_se, id, unnamed 32: | 16 | 97.20% |
| VIII. | smoothness_se, fractal_dimension_se, id, unnamed 32:, perimeter_mean, texture_mean, concave points_worst, radius_se, radius_worst, perimeter_worst, area_worst | 16 | 95.80% |

## V. CONCLUSION

In this paper, we find Random forest is widely used technique to find the accurate result for the medial dataset. In related study, various type of datasets is affecting the results. Some of the results will affect by the splitting of dataset into different portions such as 70%-30%, 60%-40%. Results will also affect by taking higher count estimators. In this paper, we found that prediction models are the fastest way to diagnosis malignant. Prediction model can diagnosis malignant at the early stage. The proposed technique spilt the data in the portion of 75%-25% and removed two variables. After 8 successful tests finally, the achieved accuracy of result is 98.60% by taking 6 estimators.

## REFERENCES

[1] H. Asri, H. Mousannif, H. A. Moatassime and T. Noel, "Using Machine Learning Algorithm for Breast Cancer Risk Prediction and Diagnosis" pp.1064-1069,2016

[2] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, pp. 113–127, 2005.

[3] Chandrasekar RM, Palaniammal V, Phil M. "Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis". IOSR Journal of Computer Engineering (IOSR-JCE). 2013; 15(5):39-44.

[4] Pietro Lio & Sarinder Kaur Dhillon,"Predicting factors for survival of breast cancer patients using machine learning techniques",Article no.48(2019).

[5] Mamta Jadhav and Zeel Thakkar," Breast Cancer Prediction using Supervised Machine Learning Algorithms", vol.06,2019.

[6] Manisha Bahl &Regina Barzilay," High-Risk Breast Lesions: A Machine Learning Model to Predict Pathologic Upgrade and Reduce Unnecessary Surgical Excision",vol 286,10-17,2017.

[7] Cameron Wolfe," Training a Random Forest to Identify Malignant Breast Cancer Tumors",2018.

[8] Ashish, "A Random Forest approach to predicting breast cancer in working class women"2016