Predict of Breast Cancer Using Machine Learning Algorithms

HARDI PATEL[1], DR MEHUL P. BAROT[2]

[1]*Research Scholar, Department of Computer Engineering, LDRP ITR, KSV*

[2]*Associate Professor, Department of Information Technology, LDRP-ITR, KSV*

Abstract— Breast Cancer is the second cause of death among women. Early prediction of breast cancer will help with the survival of breast cancer patient. Machine Learning and Data Mining have been widely used in the prediction of breast cancer and on the early detection of breast cancer. This paper compares the machine learning techniques which are used for the prediction of breast cancer and deriving results which is the best algorithm to predict breast cancer. In this paper the machine learning algorithms are compared to achieve most accurate prediction result, we are using Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM) and K-nearest Neighbor (KNN).

Index Terms— Benign, Breast Cancer, Machine Learning, Malignant, Random Forest, Tumor.

I. INTRODUCTION

In the whole world, breast cancer is the most common and dangerous cancer in women. According to the WHO report in 2020, *"It is estimate that worldwide over 685000 women died due to breast cancer."*

Data mining and machine learning have been widely used in the diagnosis of breast cancer. Also, machine learning and data mining assist the medical researchers to identify relationships between different variables and make them able to predict the outcome of disease using datasets.

Machine learning can be applied to improve breast cancer detection. Also, it could be an assistance to accurate decision making. Therefore, the aim of this research is to analyse the data mining and machine learning techniques in breast cancer detection. This research is organized as follows; Section 2 introduces of breast cancer. Section 3 explains the algorithms and tools of data mining and machine learning which are used to predict breast cancer. Section 4 discusses the literature survey. Section 5 explains proposed architecture to compare the accuracy of different algorithms. Finally, Section 6 includes the implementation of prediction of survey paper, and Section 7 includes conclusion of the survey.

II. BREAST CANCER

Breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more rapidly than healthy cells and continue to gather, forming a lump or mass. The cells may spread (metastasize) to the lymph nodes or other parts of the body. Breast cancer mostly begins with cells in the milk-producing ducts (invasive ductal carcinoma) or in the glandular tissue called lobules (invasive lobular carcinoma) or in other cells or tissue within the breast [4].

Researchers have identified hormonal, lifestyle and environmental factors that may increase the risk of breast cancer. But it is unclear why some women who have no risk factors develop cancer, yet others with risk factors never do. It is likely that breast cancer is caused by a complex interaction of genetic makeup and environment factors.

Studies say that over 1,70,000 new breast cancer cases are likely to develop in India by 2020. According to research, 1 in every 28 women is likely to get affected by the disease. While breast cancer occurs almost entirely in women, around 1-2% men are likely to get affected, too [22].

2.1 Types of Tumors:

Tumors can be benign or malignant.

(i) *Benign*: Benign tumors are those that stay in their primary position without overrunning other parts of the body. They do not spread to distant parts of the body. Benign growths will often develop gradually. Benign cancers have unmistakable lines [4].

Benign tumors are not problematic. However, they can end up massive and compress constructions nearby, inflicting ache or different scientific complications. For example, a giant benign lung tumor ought to purpose issue in breathing. This would want to press surgical operation to get rid of the most cancers from the physique. Benign tumors are unlikely to recur once removed. The two common benign tumors are fibroids in the uterus and lipomas in the skin.

Some benign tumors can flip into malignant tumors. These kinds of tumors are monitored intently and may additionally require surgical operation to dispose of it. For example, colon polyps can end up malignant consequently it wishes surgical operation to eliminate [4].

(ii) *Malignant*: Malignant tumors have cells that develop uncontrollably and unfold to the different components of the body. These sorts of tumors are cancerous. They unfold to different phase of the physique by way of the bloodstream or the lymphatic system. This spread is called metastasis. Metastasis can occur anywhere in the body and mostly it is found in the liver, lungs, breast, brain, and bone [15].

Malignant tumors can spread frequently and require surgery or treatment to avoid spread. If we can find it early, then it can be prevented by treatment. Treatments for malignant tumor is like chemotherapy or radiotherapy. If the cancer has spread, the treatment is likely to be systemic, such as chemotherapy or immunotherapy.

2.2 Symptoms of Breast Cancer:

Different people have different types of symptoms of breast cancer. Some people do not have any symptoms at all [8].

Some different types of symptoms are as follows:

- New lump will be created in the breast or underarm.
- Thickening of section of the breast vicinity or swelling of section of the breast area.
- Irritation of breast pores and skin or dimpling of breast skin.
- Redness or flaky pores and skin in the nipple location or the breast.
- Pulling in of the nipple or ache in the nipple area.
- Nipple discharge different than breast milk, consisting of blood.
- Change in the dimension or the form of the breast.
- Pain in place of the breast.

2.3 Stages of Breast Cancer:

Breast Cancer has four stages.

T0: There is no evidence of cancer in the breast.[3]

T1: The tumor in the breast is 20 millimetres (mm) or smaller in size at its widest area. This is a little less than an inch. This stage is then broken into 4 substages depending on the size of the tumor [3]:

- T1mi is a tumor that is 1 mm or smaller.
- T1a is a tumor that is larger than 1 mm but 5 mm or smaller.
- T1b is a tumor that is larger than 5 mm but 10 mm or smaller.
- T1c is a tumor that is larger than 10 mm but 20 mm or smaller.
- T2: The tumor is larger than 20 mm but not larger than 50 mm.
- T3: The tumor is larger than 50 mm. T4: The tumor falls into 1 of the following groups:
- T4a means the tumor has grown into the chest wall.
- T4b is when the tumor has grown into the skin.
- T4c is cancer that has grown into the chest wall and the skin.
- T4d is inflammatory breast cancer.

2.4 Risk Factors:

- Inherited Breast Cancer:

Gene mutations passes through generations of a family causes about 5 to 10% of breast cancers. The most well-known inherited mutated genes that can increase the likelihood of breast cancer are breast cancer gene 1 (BRCA1) and breast cancer gene 2 (BRCA2), both of which significantly increase the risk of breast cancer.

- Increasing Age:

The risk of breast cancer increases as age is increases.

- A Personal History of Breast Conditions:

If you have had a breast biopsy that found lobular carcinoma in situ (LCIS) or atypical hyperplasia of the breast, there is an increased risk of breast cancer

- A Personal History of Breast Cancer:

If you had breast cancer in one breast, you have an increased risk of developing cancer in the other breast

- A Family History of Breast Cancer:

If your mother, sister, or daughter was diagnosed with breast cancer, especially at a young age, your risk is increased [25].

- Inherited Genes that Increase Cancer Risk:

Some gene mutations that increase the risk of breast cancer can be passed from parents to children. The most well-known gene mutations are BRCA1 and BRCA2. These genes can greatly increase the risk of breast cancer and other cancers, but they do not make cancer inevitable

- Radiation Exposure:

If you received radiation treatments to your chest as a child or young adult, your risk of breast cancer is increased [26].

- Obesity:

Being obese or overweight increases the risk of breast cancer. With more body fat, your body stores more estragon and estragon stimulates tumor growth. Maintain a healthy weight by being physically active and eating a healthy, balanced diet.

- Early Menstruation or Late Menopause.:

Beginning your period early (before 12) and menopause at an older age (after 55), increases the risk of breast cancer. The longer a woman menstruates, the higher her lifetime exposure to the hormone’s estragon and progesterone. A higher lifetime exposure to estragon is linked to an increase in breast cancer risk [29].

- Having the First Child at an Older Age:

Women who give birth to their first child after age 35 may have a slightly increased risk of breast cancer.

- Never Had a Pregnancy:

Women who have never been pregnant have a greater risk of breast cancer than women who have had one or more pregnancies.

- Postmenopausal Hormone Therapy:

Women who take hormone therapy that combine estragon and progesterone to treat menopause have an increased risk of breast cancer.
The risk decreases when women stop taking these medications.

- Drinking alcohol:

Drinking alcohol increases the risk of breast cancer. The risk increases with the amount of alcohol consumed. The way alcohol is metabolized in a woman's body may increase estragon levels in the bloodstream, which increases the risk.

- Being Physically Inactive:

Regular physical activity reduces breast cancer risk, especially in menopausal women.

- Long Term Use of Oral Contraceptives:

Women using oral contraceptives (birth control pills) have a slightly higher risk of breast cancer. Once the pills are stopped, this risk goes back to normal within about 10 years.

- Dense Breasts:

Breast tissue is called dense if there is a lot of fibrous or glandular tissue and not much fat in the breasts. Women whose breasts are dense, rather than fat, have a higher risk of breast cancer, and the risk increases with higher breast density.

- Smoking:

Smoking increases the risk of at least 15 cancers including breast cancer. Women who smoke are more likely than non-smokers to develop breast cancer. Long-term exposure to second hand smoke may also increase the risk [26].

III. BIGDATA ANALYTICS AND MACHINE LEARNING

Big Data:

Big data analytics is the use of advanced analytic techniques against large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from [5]. Big data is a time utilized to datasets whose measurement or kind is beyond the capability of relational databases to capture, control and system the statistics with low latency. Big data has following characteristics: high volume, high velocity, high variety, veracity, and value.

Applications of big data analytics can improve the services which are patient based, to detect diseases earlier, generate new patterns into disease mechanisms, monitor the quality of the medical and healthcare institutions as well as provide better methods of treatments [6].

Machine Learning:
Machine Learning is a learning program from experience to improve its performance without human instruction

There are two types of learning:
(i) Supervised Learning
(ii) Unsupervised Learning

- Supervised learning: In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.
- Unsupervised learning: This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined [27].

3.1 Data Mining Algorithms:
There are many algorithms such as Naïve Bayes, K-Nearest Neighbor, k-mean, Random Forest; They are used for analysing a huge amount of data.

Some popular Data Mining Algorithms are discussed as follows:

(i) *Naïve Bayes*:
It is a probabilistic classifier [10]; it is one of the efficient classification algorithms based on applying Bayes' theorem with strong (naïve) independent assumptions. It assumes the value of the feature is independent of the value of any other features, given the class variable. Based on the maximum probability. It detects the class membership for the given tuple to a particular class.

(ii) *K-Nearest Neighbor*:
KNN algorithm is also called as Instance-Based Learning. KNN is the simplest approach for classification of samples. Here different distance measures are used for classifying samples. K-nearest Neighbor finds the number of samples from training data which is near to the test samples and assigns to the frequent class label [14].

In this algorithm, training samples generate the classification rules without considering extra information. It has excessive likelihood when associated cases belonging to the same type [14]. Based on K training samples KNN algorithm identifies the test samples. For every situation, K value will be a positive integer.

(iii) *Support Vector Machine*:
Support Vector Machine (SVM) which is designed in 1990's. To achieve machine learning tasks support vector machine is used, and it is a simple and prominent process. During this technique, a collection of training samples is given each sample is divided into different categories. Support vector machine mainly used for classification and regression problems.

(iv) *Decision Tree:*
Decision tree algorithms are successful machine learning classification techniques. They are the supervised learning methods which use information gained and pruned to improve results. Moreover, decision tree algorithms are commonly used for classification in much research, for example, in the medicine emergency and health issues. There are many types of decision tree algorithms like ID3 and C4.5. However, J48 is the most popular and useful decision tree algorithm. J48 is the implementation of an improved version of C4.5 and is an extension of ID3.

(v) *Random Forest*:
A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning. Ensemble learning is a technique which combines many classifiers to provide solutions to complex problems.

A random forest algorithm contains many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating.

The algorithm establishes the outcome based on the predictions of the decision trees. It takes the mean or average of the output from the various trees and then predict the outcome. To increase the precision of the outcome we must increase the number of trees.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations in packages.

3.2 Data Mining Tools:

Data mining tools provide ready to use an implementation of the mining algorithms. Most of them are free opensource software. Some of the popular data mining tools are discussed in the following:

(i) *WEKA*:

The Weka is a collection of machine learning algorithms and data pre-processing tools for Knowledge Learning. WEKA stands for Waikato Environment for Knowledge Analysis. It is a computer program that was developed at the University of Waikato (New Zealand). The program is written in Java, and it runs on almost any operating system. It is a free data mining software. WEKA supports evaluating, visualizing, and preparing the input data. It supports different machine learning algorithms like classification, clustering, and regression.

(ii) *Tanagra*:

Tanagra is a free machine learning software for research and academic purposes. It was developed by Ricco Rakotomalala at the Lumière University, France. Tanagra supports different types of data mining tasks like visualization, descriptive statistics, regression, clustering, classification, and association rule learning.

(iii) *Orange*:

Orange is a Python-based tool for machine learning and data mining. Its visual programming interface is clean and easily understood. The orange may be more suited for novice researchers and small projects [7].

(iv) *MATLAB*:

MATLAB as a data mining tool has an interpreted language and graphical user interfaces. It also has hundreds of mathematical functions to support multi-paradigm numerical calculations which make it suitable to the computing environment.

IV. LITERATURE REVIEW

(i) Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model

In this paper, K. Sivakami uses Decision tree and Support Vector Machines (DT-SVM) [6] both are hybrid methods. To introduce a disorder status prognosis, they employ DT-SVM methods. The experiment was performed through Weka tool. The authors have considered the Wisconsin breast cancer dataset that includes 699 instances; in those 458 instances belong to not cancer (benign) class and other 241 instances belong to cancer (malignant) class. Finally, the author compared the output of the DT-SVM model with Naive Bayes, instance-based learning (IBK), and sequential minimal optimization (SMO) and conclude that DT-SVM gives better accuracy i.e., 91% compared to NB, IBK, and SMO.

(ii) Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach

In this paper, D.R. Umesh and B. Ramachandra [1] have utilized Expectation Maximization (EM) algorithm for identifying the breast cancer recurrence. To find out the classification accuracy they have used SEER dataset which contains 2,20,811 instances with 17 attributes. The authors have performed their experiment through Amazon cloud computing environment (EC2) and declare expectation maximization algorithm gives 88.54% of accuracy.

(iii) Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review

In this paper, Hiba Asri et al. [7] performed this experiment to determine the efficiency and effectiveness of various algorithms like Support

Vector Machine (SVM), K Nearest Neighbor (K-NN), Decision Tree (C4.5), and Naive Bayes (NB). They utilized Wisconsin breast cancer (original) dataset taken from UCI machine learning repository contains 699 instances with 11 attributes. The experiment is performed on WEKA tool and outcomes show that the SVM gives higher accuracy 97.13% compared to K-NN, C4.5 i.e., 95.27%, 95.13%.

(iv) Prediction of Breast Cancer using Big Data Analytics:

In this paper, K. Shailaja et al [12] uses KNN algorithm to classify cancer tumor as either benign or malignant. This approach is evaluated and compared using Wisconsin Breast Cancer dataset. The authors have applied feature selection on the dataset to remove duplicate and irrelevant features. The experiment result shows the accuracy, precision, recall and F-measure are increased by the proposed method when compared with different models. Accuracy before feature selection is 96.6% and after feature selection is 98.14%.

(v) Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis

In this paper, Hiba Asri et al [11] employed four main algorithms: SVM, Naïve Bayes, KNN, C4.5 on the Wisconsin Breast Cancer (original) Dataset. The authors try to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity, and specificity to find the best classification accuracy. SVM reaches at higher accuracy of 97.13%. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.

(vi) Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier

In this paper, P. R. Anisha et al [16] used six main machine learning algorithms to predict and diagnose the breast cancer. Comparison of the six algorithms: Logistic Regression, Decision Tree, K- nearest Neighbor, Naïve Bayes, Support Vector Classifier and Random Forest Classifier. The author got higher accuracy 98% of the Random Forest classifier.

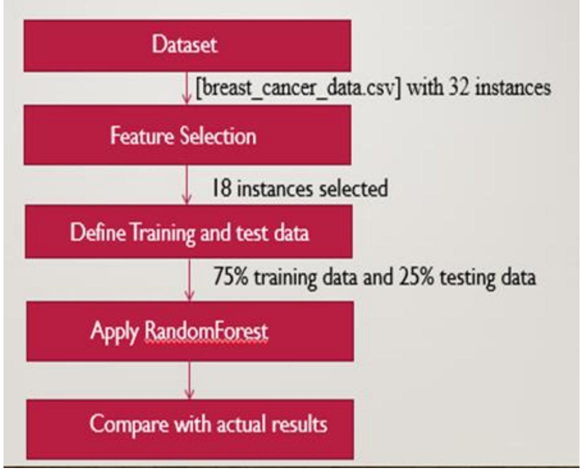
(vii) *Performance Analysis of Different Classifiers in Prediction of Breast Cancer*

In this paper, S. Roobini et al [14] performed different methodology and perform analysis of different classifiers in prediction of breast cancer.

In this research, 10-fold cross validation is used to validate the results. The dataset is divided into ten equal subsets randomly. One of the partition acts as a testing set, whereas the rest of the partitions act as training set to train the model. A relative report on the execution of existing and proposed grouping model is talked about dependent on Accuracy, Error rate, F - measure, exactness, and review. Precision quantum's how profound the settled tuples are being ordered effectively, TP embodies to positive tuples and TN epitomizes to negative tuples characterized by the essential classifiers. So also, FP ascribes to positive tuples and FN attributes to negative tuples which is inaccurately grouped by the classifiers.

The performance of Fuzzy C-Means Clustering [FCM] with Naive Bayesian classifier provides a better prediction when compared to other classifiers.

V. PROPOSED ARCHITECTURE


Dataset
[breast_cancer_data.csv] with 32 instances
Feature Selection
18 instances selected
Define Training and test data
75% training data and 25% testing data
Apply RandomForest
Compare with actual results

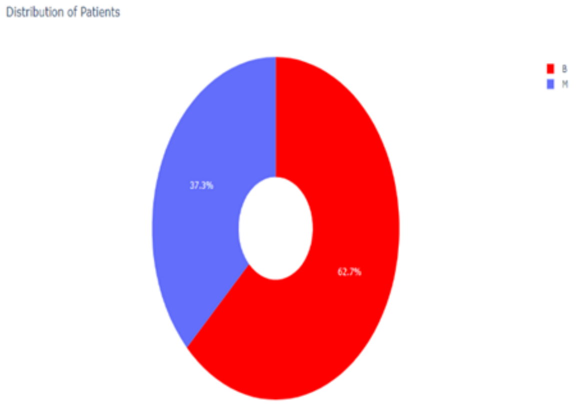
[Proposed Architecture]

VI. IMPLEMENTATION

- Database:

Wisconsin (Original) database has been used for prediction of breast cancer.
Wisconsin database contains 32 different attributes.
Distribution of patients into two parts based on their diagnosis which is benign(B) and malignant(M).

- Data Visualization:



[Data Visualization for distribute of patients]

- Training and testing dataset
- Training Data:
- The observations in the training set form the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed output variable and one or more observed input variables.
- Testing Data:
- The test set is a set of observations used to evaluate the performance of the model using some performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it.
- Splitting training and test dataset:

Training Dataset:75%

Testing Dataset:25%

- Comparative Analysis:

To understand the efficiency of different algorithms, we construct the confusion matrix to compare different algorithms like Naïve Bayes, SVM (Support Vector Machine), KNN and Random Forest and Decision Tree.

Confusion Matrix:

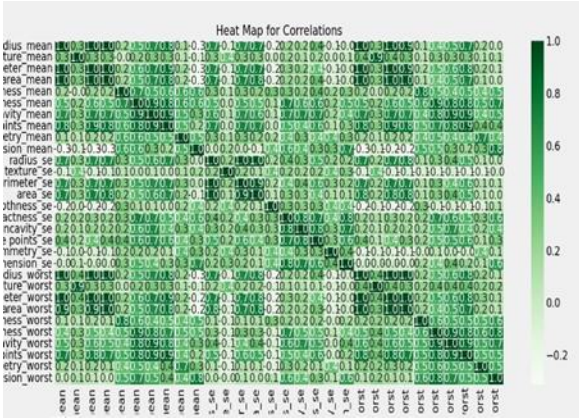
Algorithms	Training Accuracy	Testing Accuracy
Logistic Regression	98.82%	97.20%
Decision Tree	100%	96.50%
Random Forest	100%	97.20%
Naïve Bayes	93.89%	95.80%
Support Vector Machine (SVM)	98.82%	97.20%
K-nearest Neighbor (KNN)	97.65%	97.90%

[Confusion Matrix for comparison of Algorithms]

- Heat map for correlation:

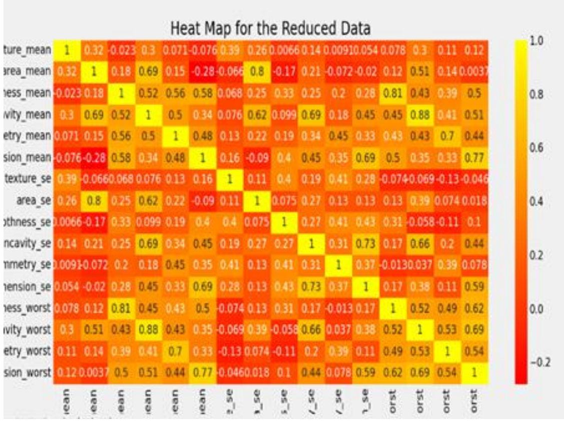
By referring the above diagram, some attributes can be eliminated which has lowest correlation.

Deleted attributes after observing the heat map are 18 like perimeter_mean, redius_mean, compactness_mean, concave point_mean, radius_se, perimeter_se, radius_worst, perimeter_worst, compactness_se, concave point_se, texture_worst, area_worst.



[Heat map for correlations]

- Heat Map for the Reduced Data:

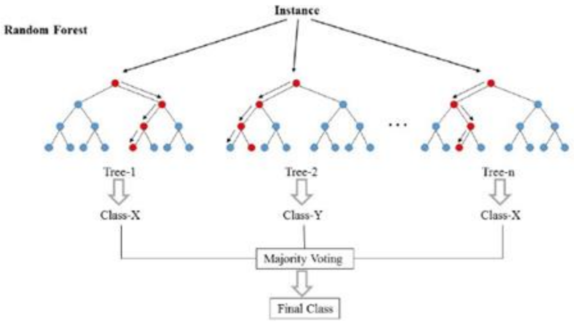


[Heat map of reduced data]

- Applying Random Forest:

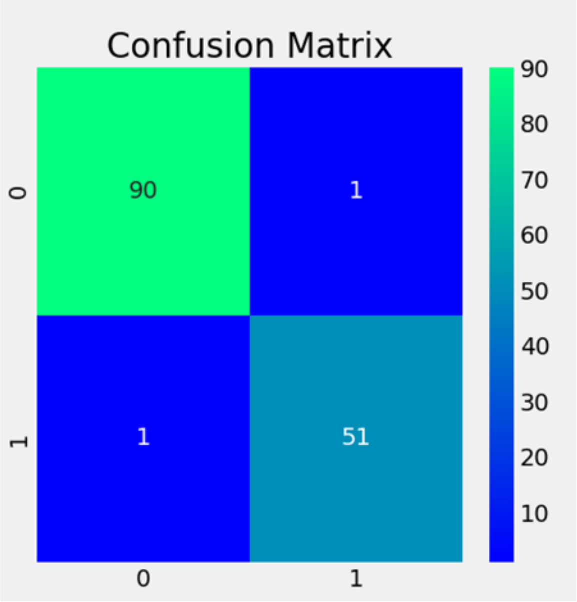
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Steps involved in random forest algorithm:
Step 1: In Random Forest n number of random records are taken from the data set having k number of records.
Step 2: Individual decision trees are constructed for each sample.
Step 3: Each decision tree will generate an output.
Step 4: Final output is considered based on Majority Voting or Averaging for classification and regression respectively.



[Working of Random Forest]

- Results:



[Confusion Matrix of Result]

Results of the classification using random forest is shown in above confusion matrix. In confusion Matrix Benign is denoted by 0 and Malignant is denoted by 1.

Training Accuracy: 100%

Testing Accuracy: 98.60%

CONCLUSION

- In this research, the Wisconsin database is used to predict the breast cancer.
- By comparing the different type of algorithms on the dataset, I got different accuracy.
- After eliminating irrelevant features(attributes) if the dataset, the accuracy is increasing.
- Random forest gives the accuracy of 97.20% before feature selection and after feature selection, accuracy is 98.60%

REFERENCES

[1] G. Sumalatha et al., “A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques”, International Journal of Innovative Research in Computer and Communication Engineering, volume 5,2017.

[2] D.R Umesh et al., “Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach”, International Journal of Computer Applications, volume 7, 2016.

[3] Hiba Asri, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis”, The 6th International Symposium on Frontiers in Ambient and Mobile Systems, pp.1064-1069

[4] K. Shailaja,” Prediction of Breast Cancer Using Big Data Analytic”, International Journal of Engineering & Technology, volume 7, 2018.

[5] Eltalhi, Saria & Kutrani, Huda. (2019). Breast Cancer Diagnosis and Prediction using Machine Learning and Data Mining Techniques: A Review. IOSR Journal of Dental and Medical Sciences. 18. 85-94.

[6] S. Roobini and J. Fenila Naomi, “Performance Analysis of Different Classifier in Prediction of Breast Cancer”, International Journal of Science and Technology, volume 12(8) , 2019.

[7] Emanelwerfally, & Kutrani, Huda & Eltalhi, Saria & Ashleik, Naeima. (2021). Predicting Breast Cancer Treatment Using Decision Tree Algorithms and Statistical Metrics. IOSR Journal of Dental and Medical Sciences. 20. 48-54

[8] V. Sivakumar et al, “Feasibility Study on Data Mining Techniques in Diagnosis of Breast Cancer”, International Journal of Machine Learning and Computing”, Volume 9 ,2019.

[9] Eltalhi S, Kutrani H. Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. IOSR J. Dental Med. Sci. 2019 Apr;18(4):85-94.

[10] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1; 13:8-17.

[11] Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. The Morgan Kaufmann series in data management systems. 3rd ed. Burlington: Morgan Kaufmann Publishing; 2011.

[12] Sornlertlamvanich V, Potipiti T, Charoenporn T. Automatic corpus-based Thai word extraction with the C4. 5 learning algorithms. InCOLING 2000 Volume 2: The 18th International Conference on Computational Linguistics 2000.

[13] D.R Umesh et al., “Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach”, International Journal of Computer Applications, volume 7, 2016.

[14] Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. Jama. 2005 Mar 9;293(10):1245-56.

[15] Organization World Health, “WHO | Breast cancer: prevention and control,” WHO. 2016, Accessed: Jan. 11, 2021. [Online]. Available: https://www.who.int/cancer/detection/breast cancer/en/.

[16] Abulkasim MA. The prevalence of breast cancer in Africa and establishment of The Libyan Breast Cancer Registry (Master's thesis, Faculty of Health Sciences).

[17] Yadav P, Varshney R, Gupta VK. Diagnosis of breast cancer using decision tree models and SVM. International Research Journal of Engineering and Technology (IRJET) e-ISSN. 2018 Mar:2395-0056.

[18] Kutrani H, Eltalhi S. Cardiac catheterization procedure prediction using machine learning and data mining techniques. IOSR Journal of Computer Science. 2019;21(1):86-92

[19] https://www.python.org/about/gettingstarted/

[20] https://searchdatamanagement.techtarget.com/definition/5-Vs-of-big-data?amp=1

[21] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[22] https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html

[23] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[24] https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

[25] https://www.fromthegenesis.com/pros-and-cons-of-k-nearest-neighbors/amp/

[26] htts://my.clevelandclinic.org/health/diseases/3986-breast-cancer

[27] https://www.cancer.net/cancer-types/breast-cancer/stages

[28] https://jamanetwork.com/journals/jamaoncology/fullarticle/2768634

[29] https://www.ibm.com/in-en/analytics/hadoop/big-data-analytics