iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.41045

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Big Data Analytics to Predict Breast Cancer

Hardi Patel[1], Dr. Mehul P. Barot[2]
[1]*Research Scholar,* [2]*Associate Professor, Department of Computer Engineering, 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.
Keywords: Breast Cancer, Malignant, Benign, Machine Learning, Big Data Analytics.

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 about the dataset of the breast cancer. Section 5 discusses the literature survey. Section 6 explains proposed architecture to compare the accuracy of different algorithms. Finally, Section 7 includes conclusion of the survey.

II. BREAST CANCER

Normally, cells in the body divide (reproduce) only when new cells are needed. Sometimes, cells grow and they divide out of control, which creates a mass of tissue called a tumour. If the tumor is benign then the cells that are growing out of control that are normal cells. If, however, the cells are growing out of control are abnormal and don't function like the body's normal cells, the tumor is called malignant.
Cancers are named after the body part from which they originate. The cancer which is originates in the breast tissue is called Breast Cancer. Like other cancers, breast cancer can grow into the tissue surrounding the breast. It can also travel from breast to other parts of the body and create new tumors, a process called metastasis[2].

A. Types of Tumors

Tumors can be benign or malignant.

1) *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].
2) *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 [4]. 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.

B. 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:

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

C. Stages of Breast Cancer

Breast Cancer has four stages.

1) T0: There is no evidence of cancer in the breast.[3]
2) 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]:

a) 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.

b) **T2:** The tumor is larger than 20 mm but not larger than 50 mm.
c) **T3:** The tumor is larger than 50 mm.
d) **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.

III. BIG DATA ANALYTICS AND MACHINE LEARNING

1) 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 period 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].

2) Machine Learning

Machine Learning is a learning program from experience to improve its performance without human instruction
There are two types of learning:

a) Supervised Learning
b) Unsupervised Learning

A. 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:

1) *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.
2) *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.
3) *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.
4) *Decision Tree Algorithm(J48):* 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 many 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.
5) *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.

B. 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:

1) *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.
2) *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.
3) *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].
4) *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. BREAST CANCER DATASET

For the prediction of breast cancer, we used breast cancer Wisconsin(original) dataset. The dataset includes 699 instances and 11 attributes along with the class label. The distribution of class will be 458 instances belong to the benign class and other 241 instances belong to the malignant class.

Data Description

Attribute	Domain
1. Sample code number	Id number
2. Clump thickness	1–10
3. Uniformity of cell size	1–10
4. Uniformity of cell shape	1–10
5. Marginal adhesion	1–10
6. Single epithelial cell size	1–10
7. Bare nuclei	1–10
8. Bland chromatin	1–10
9. Normal nucleoli	1–10
10. Mitoses	1–10
11. Class	2 for benign, 4 for malignant

Breast Cancer Dataset [9]

1) Sample code number indicates id number.
2) Clump thickness determines whether it contains single or multi layered cells.
3) Uniformity of cell size means in the given samples it determines the size of cells which are consistency.
4) Uniformity of cell shape: It recognizes marginal differences and determines the cell shapes.
5) Marginal adhesion: It evaluates how many cells present on the external of the epithelial and they are stick together.
6) Single epithelial cell size: It identifies the epithelial cells that are necessarily expanded, and it also describes the uniformity of cells.
7) Bare nuclei: it computes the hypothesis of the bunch of cells that are not encircled by the cytoplasm.
8) Bland chromatin: it ranks the pattern of a nucleus from admirable to rude.
9) Normal nucleoli: it identifies either the nucleoli are tiny, hardly apparent or huge, most clearly visible.
10) Mitoses: mitoses illustrate the level of the mitotic state.

V. LITERATURE SURVEY

A. 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) 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.

B. 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.

C. 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%.

D. 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%.

E. 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.

F. 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.

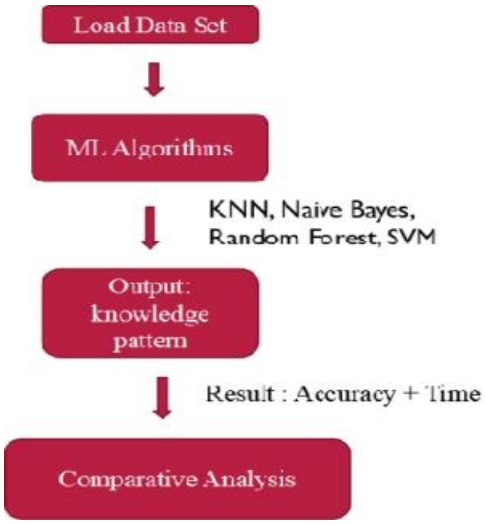
G. Performance Analysis of Different Classifiers in Prediction of Breast Cancer

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 act 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 the means by which 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.

VI. PROPOSED ARCHITECTURE



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.

A. Confusion Matrix

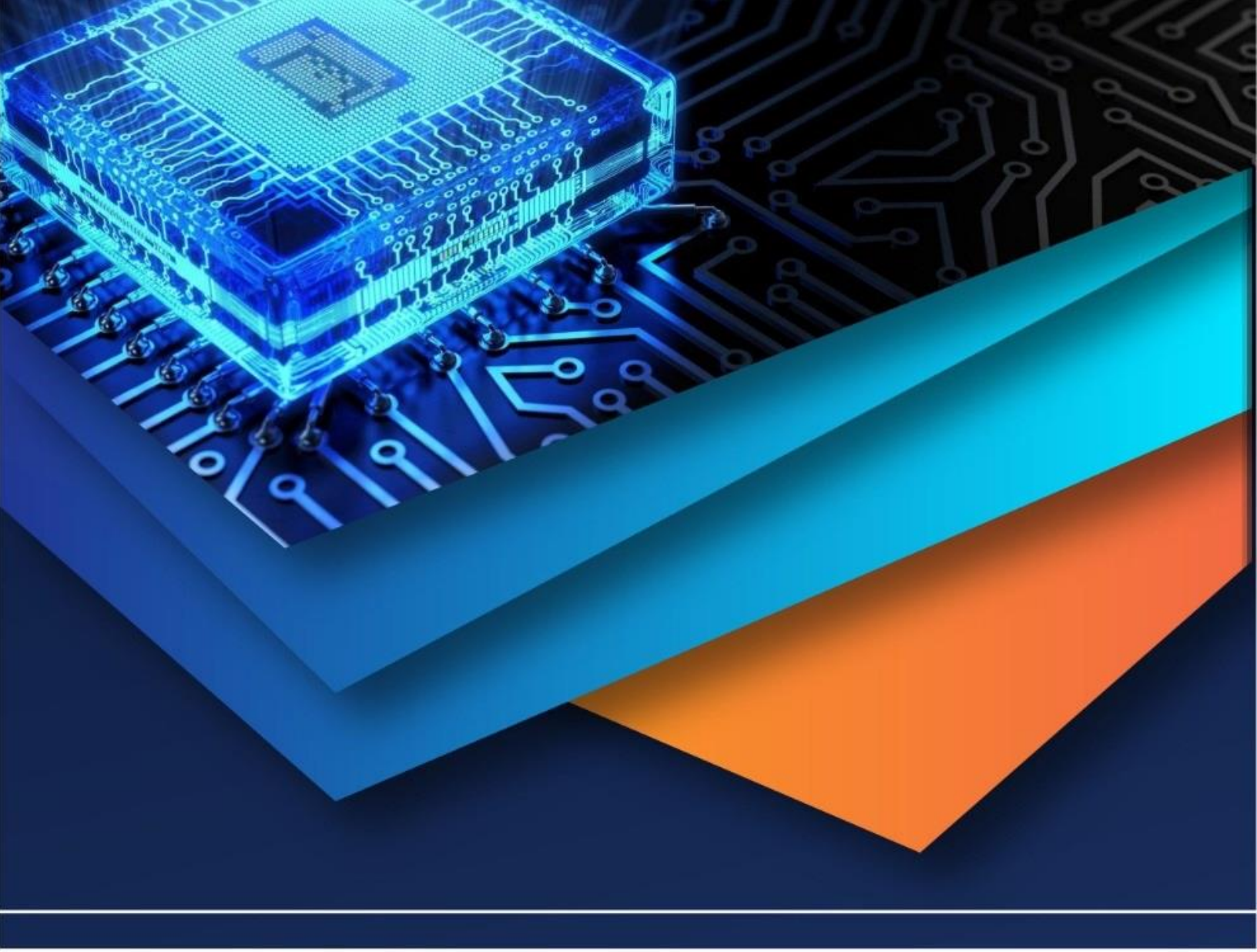
Algorithm	Benign	Malignant	Class	Accuracy
Naïve Bayes	436 6	22 235	Benign Malignant	95.99%
SVM	445 10	13 231	Benign Malignant	96.71%
KNN	445 20	13 221	Benign Malignant	97.6%
Random Forest	443 7	15 234	Benign Malignant	96.85%

VII. CONCLUSION

In this paper, we compared different type of machine learning algorithms to find the most accurate algorithm to classify the breast cancer dataset into two different classes benign and malignant. we performed these algorithms on WEKA tool. This experiment shows different accuracy of all the algorithms. KNN got the highest accuracy of 97.6%.

REFERENCES

[1] 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.
[2] https://my.clevelandclinic.org/health/diseases/3986-breast-cancer
[3] https://www.cancer.net/cancer-types/breast-cancer/stages
[4] https://jamanetwork.com/journals/jamaoncology/fullarticle/2768634
[5] https://www.ibm.com/in-en/analytics/hadoop/big-data-analytics
[6] **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6340124/**
[7] Saria Eltalhi. "Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review." IOSR Journal of Dental and Medical Sciences (IOSR JDMS), vol. 18, no. 04, 2019, pp 85-94.
[8] **https://www.cdc.gov/cancer/breast/basic_info/symptoms.htm**
[9] https://www.researchgate.net/figure/Breast-cancer-dataset_tbl1_323952426
[10] 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.
[11] 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
[12] K. Shailaja, " Prediction of Breast Cancer Using Big Data Analytic", International Journal of Engineering & Technology, volume 7, 2018.
[13] 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.
[14] 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.
[15] 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
[16] 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY