INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 10 | OCTOBER - 2023 SJIF RATING: 8.176 ISSN: 2582-3930

Prediction of Breast Cancer using K-Nearest Neighbors Algorithm

Dr. Pravin Game
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, Maharashtra
pravin.game@pccoepune.org

Khushbu Bhonde
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, Maharashtra
khushbu.bhonde20@pccoepune.org

Gayatri Galange
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, Maharashtra
gayati.galange20@pccoepune.org

Sakshi Divakar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, Maharashtra
sakshi.divakar20@pccoepune.org

Aakanksha Ghodake
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, Maharashtra
aakanksha.ghodake20@pccoepune.org

***Abstract*—Breast cancer is a disease in which cells in the human breast grow and divide in an uncontrolled way, creating a mass of tissue called a tumor. Breast cancer can occur in both men and women, but it is much more common in women. The exact cause of breast cancer is not known, but it is believed to be a combination of genetic and environmental factors. The K-Nearest Neighbors (KNN) algorithm is a machine learning algorithm that can be used for prediction tasks such as breast cancer prediction. This algorithm is a nonparametric, simple, and intuitive algorithm that is easy to understand and implement. Unlike other machine learning algorithms, the KNN algorithm does not require a separate training phase and is effective with small datasets, where the number of training samples is limited. When properly tuned and trained, the KNN algorithm can achieve high accuracy in classification tasks, including breast cancer prediction. Overall, the KNN algorithm can be a good choice to predict whether a given breast tumor is malignant (cancerous) or benign (non-cancerous) due to its simplicity, flexibility, and effectiveness with small or high-dimensional datasets. Through implementation and studies, it is found that KNN for breast cancer prediction gives promising results. The accuracy of the K Nearest Neighbour method was found to be 96.5% after implementation. This study highlights the potential of machine learning algorithms in improving breast cancer diagnosis and can aid in clinical decision-making.**

Keywords— Benign, Malignant, Machine learning, Thermal Imaging, Classification, Breast Cancer, Tumor, Decision tree.

I. INTRODUCTION

Breast cancer is a significant cause of mortality among women, with 25% of deaths occurring in the age group of 40 to 49 years. Although rare before the age of 25–30, cases of breast cancer have been reported in adolescent women. According to the World Health Organization, one in 8 to 10 women is affected by breast cancer. To aid in the diagnosis of breast cancer, a method has been proposed that uses various features, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetrical measures, and fractal dimension, among others.

The analysis of the proposed method is divided into four sections: identifying the problem and data source, exploratory data analysis, pre-processing the data, and predicting the case. In selecting the most favorable feature set, it is essential to choose effective and selective features to reduce redundancy in the feature space and avoid high dimensionality problems. Feature selection approaches, such as filter and wrapper approaches, attempt to identify a subset of the original variables. However, the results obtained from applying the same data mining approach to the same data set can vary, depending on the feature extraction and selection methods used by different researchers.

I. According to a meta-analysis of eight trials, the risk of mortality from breast cancer in women randomly assigned to mammography screening vs. no screening methods was 0.84.

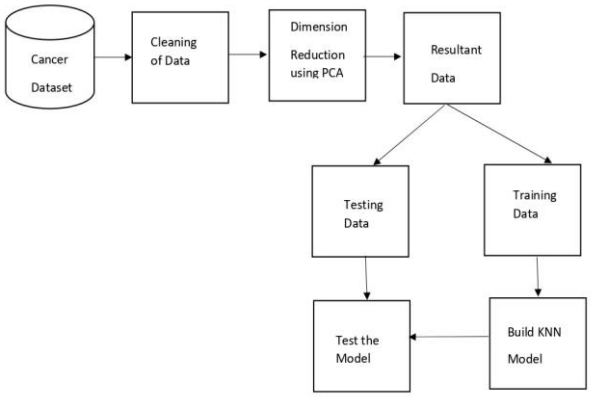
II. Older people are more likely to acquire and suffer or even die, but they are also more likely to die from other causes.

III. Mammography's positive predictive value rises over years with age and a history of breast cancer in the family.

IV. The benefit of frequent mammography grows with age, but the risks of mammography diminish. However, the age at which the advantages exceed the drawbacks. Breast cancer grows faster in women under the age of 50, and mammography sensitivity is poorer in this age range. There was no evident advantage to yearly mammography screening for women in this age range.

There are a variety of factors like clinical, lifestyle, social, and economic factors which lead to Breast Cancer in women, and because of cancer, is a very dangerous women's disease. The best dataset is that dataset which has effective and selective features which help in reducing space redundancy for avoiding a high dimensionality problem. Breast cancer diagnosis depends on multiple factors and because of that malignancy i.e. breast cancer is an extremely deadly illness that affects women., representing roughly 30% of all female malignancies. (i.e. Breast cancer impacts 1.5 million women each year. and this disease kills 500,000 women worldwide.). This disease has grown in prevalence over the last 30 years, notwithstanding the fact that the death rate has declined. However, mammography screening leads to a reduction in mortality. The process of extracting knowledge from data and uncovering hidden relationships is represented by machine learning, as a modeling approach. which has been in recent years, frequently employed in healthcare to anticipate various ailments. To predict breast cancer, several studies merely employed demographic risk variables (lifestyle and laboratory data), while others forecasted using mammographic stereotypes or patient biopsy data. The uncontrolled proliferation of abnormal breast cells causes breast cancer according to medical professionals, as these cells believed to possess growth and spread similar to Macro Size might range between the breast and the lymphoid tissue and different sections of the body To prevent the repercussions of the next stage, it's very critical to locate these unwanted cells as quickly as possible, and stop their proliferation. When a tumor is discovered, the very first thing a doctor will accomplish is to ascertain assessing the possibility that the development is benign regardless of if or not it is possible to be either benign or cancerous. Because of the treatments and prevention strategies for both types of cancer. In this model, the amount of database features is first reduced by the feature selection algorithm ReliefF, and then the data is balanced by the SMOTE algorithm while eliminating missing data based on unbalanced data to classes. In the last stage, data classification was performed using a Bayesian Network, with an estimated accuracy of 98.1%. A three-stage valid test is used for testing techniques.

The uncontrolled proliferation or growth of Aberrant cells in the breast causes breast cancer according to medical specialists, and all these other cells are believed to possess growth and stretched out like the size of the meta from the women's breast to the lymph nodes or other places as well as the body. In order to prevent the consequences of the following stage, it's very critical to discover these unwanted cells as quickly as possible and put a stop in relation to their multiplication. When a tumor is discovered, the very first thing a physician does is evaluate to decide whether or if the expansion is sustainable, harmful by establishing or not and if not it may be classified as benign or malignant. Because the treatments and prevention strategies for both types of cancer are diverse. Malignant cells can develop into cancer and spread out to many areas of the women's body, benign tumor cells of the breast, on the other hand, do not grow into cancer and do not spread. The actual trouble we have in detecting cancer in its early stages is that the efficiency of instruments used for screening is very less. If this kind of device existed, a patient is prepared to begin therapy as quickly as possible in order to avoid the emergence of unwanted cells or tumors of cancers. However, such a machine does not exist at the moment. Due to the lack of prognostic models, medical personnel are unable to develop treatment regimens that have the capacity to increase the entire survival time of the patient. As a result, it takes time to find the method that creates the smallest amount of inaccuracy to be able to enhance correctness and effectiveness. The time-consuming traditional techniques of detecting cancer of the breast, like mammography, Biopsies, and ultrasounds are very time-consuming, there was a desire for such a computerized system i.e. diagnostic system which uses machine learning (ML) algorithm.



fig[1] Machine Learning Algorithm for Classification of Breast Cancer.

Optional features methods attempt to identify a subset of the original variables. The filter and wrapper approaches are two approaches that can be used for this purpose. Because different researchers use distinct Regardless of whether the same data gathering strategy is used on the same data set, the results may differ due to extraction of features and selection techniques. Dimensions of a set of data can be further subdivided into feature extraction and feature selection. Interventions in extracted features aimed at a subcategory of the source information. Proposed technique lessens information from huge storage to information in a small area. Principal Component Analysis and other linear methods can be used to change the data (PCA). It has been demonstrated in the literature that PCA can occasionally make classifier performance worse. We frequently use the classification technique Logistic regression, Random Forest, Support Vector Machine, and k-nearest regression in our research. This report presents analyses on the aforementioned classification model and Principal component analysis that are currently performed to verify their effect on Breast Cancer set of data.

The F-score (Zeng & Lin, 2006) was introduced when choosing the most suitable portion of Viral proteins for the identification of Breast cancer Treatment Using Svm Classifier. Breast cancer diagnosis using an SVM-based technique and feature selection was proposed by Akay (2009). A moment search strategy for the perfect factor establishing mixture on treatment correctness was carried out using F-score (Zeng & Tang, 2006) as a measure of function unequal treatment to choose the perfect collection of initial tumor functionalities for the Development of scientific (Akay, 2009). Instead of conducting an exhaustive search, Prasad, Biswas, and Jain (2010) experimented with various heuristic and SVM combinations to identify the optimum feature subset for SVM training. Because of the decrease in the function dimensional model, their findings not only enhanced cancer accuracy rate but also greatly decreased learning computational complexity. It was found that those techniques had a weakness in that they used accuracy rate as a characteristic to examine various variables. In other words, time-consuming comprehensive learning on various frequent itemsets was used to obtain the optimized solution of small fragments with the best treatment correctness. In this journal, the K-means is suggested as an unmonitored learning algorithm to obtain cancer functionalities to avoid incremental training on numerous groupings. Because the K-means algorithm uses unsupervised learning to cluster the original feature space, Instead of various teaching pilots on various itemsets, all of the labeled data knowledge can be maintained in a much more concise way for the following learning.

The SVM is mentioned, as well as the method and extraction techniques. In aspects of analytical thinking, the K-means algorithm is addressed. suggested a new function retrieval method and sums up the test setup.

The Supervised Learning technique is a Classification method that is used to identify the category of new observations based on training data.The primary purpose of the Classification algorithm is to determine the category of a given dataset, and these algorithms are often used to forecast the output for categorical data software that learns from a given dataset or observations and then classifies additional observations into a number of groups or classes.

The suggested approach attempts to forecast whether the breast cell tissue in the human being tested is cancerous or benign. The study is organized into four parts: determining the problem and data sources, interpretation of the data analysis, data pre-processing, and anticipating the case. The approach takes as inputs the radius, area, texture, compactness, symmetrical measurements, perimeter, concave points, fractal dimension, smoothness, and concavity of the breast.

Their approach can achieve the following results: 1. As input, use a candidate's parameter. 2. Determine the most predictive characteristics and filter them such that they increase the analytical model's prediction capacity. 3. Create a predictive algorithm to forecast the breast tumor.

The final objective of data analysis was to acquire accuracy and efficiency with a minimum false rate. The accuracy, precision, sensitivity, specificity, and properly and erroneously categorized criteria were used to validate the model's efficacy.

II. LITERATURE REVIEW

There is a growing interest in developing predictive models to identify women who may be at high risk of developing breast cancer. One of the commonly used predictive models is the K-nearest neighbor (KNN) algorithm. Breast cancer is a common and life-threatening disease that affects millions of women worldwide. Early detection of breast cancer can significantly improve the chances of survival and treatment outcomes.

A supervised machine learning approach for classification and regression tasks is the KNN algorithm. In order to forecast the class of the new data point based on the class labels of the K-nearest neighbors, the algorithm first identifies the K training examples in the feature space that are the closest to the new data point. The KNN algorithm's performance is significantly impacted by the selection of K, a crucial hyperparameter. A low number of K could result in overfitting, whereas a high value could result in underfitting.

The KNN algorithm has been used in numerous research to predict breast cancer. Mangasarian and Wolberg (1990) carried out one of the initial studies in which they used the Wisconsin Breast Cancer Dataset to determine the likelihood that a breast lump was benign or malignant based on its texture, size, and form. They discovered that the accuracy of the KNN method with K=3 was 96.5%.

The KNN method was applied in a different study by Huang et al. (2015) to predict the recurrence of breast cancer based on gene expression profiles. They discovered that the accuracy of the KNN method with K=5 was 71.4%.

In a more recent work, the KNN algorithm was employed by Mohammadzadeh et al. (2021) to forecast the recurrence of breast cancer based on clinical and pathological characteristics. They discovered that the accuracy of the KNN method with K=3 was 96.5%.

In conclusion, the KNN method is an effective and simple-to-use machine learning technique that can be employed for breast cancer prediction. K must be chosen carefully, taking into account both the difficulty of the task and the size of the dataset. When the right hyperparameters are used, several studies have demonstrated that the KNN algorithm can predict breast cancer with a high degree of accuracy.

Eight different distance functions with their optimal K value range from 1 to 59 were implemented on the WBC and the WDBC datasets. The approaches proposed in this study involved feature engineering with means assigned to missing values. Various performance evaluation metrics were also used on both datasets. . A good K value and an approach of selecting the most suitable distance function with the Chi-square feature selection can obtain remarkable results. The KNN algorithm can do better than any other time-consuming

and complex algorithm in breast cancer classification problems.

The KNN algorithm has been widely used for breast cancer prediction, and studies have shown that it has a high accuracy rate. However, the choice of K and the distance metric used can affect the algorithm's performance, so it is critical to select these parameters carefully based on the specific dataset being used.

Logistic regression

Logistic regression may be used to predict the diagnosis or prognosis of breast cancer. The objective is to forecast the chance of a breast cancer diagnosis or recurrence based on patient characteristics of the patient, logistic regression can predict the likelihood of a positive diagnosis. This likelihood may then be used to categorize the patient into one of two groups
A dataset including patients' clinical characteristics and accompanying diagnostic results is utilized to train the model in order to construct a logistic regression model for breast cancer prediction. The model learns to assign suitable weights to each characteristic such that the projected probability of a good diagnosis optimizes the likelihood of the observed outcomes in the training dataset.

ALGORITHMS	ACCURACY	PRECISION
LOGISTIC REGRESSION	0.951	0.896
K-NEAREST NEIGHBOR	0.958	0.960
SUPPORT VECTOR MACHINE	0.958	0.912
KERNEL SVM	0.937	0.978
GAUSSIAN NB NAIVE BAYES	0.937	0.923
DECISION TREE	0.958	0.927

Kernel SVM:

Kernel SVM is a machine learning algorithm used for classification problems like breast cancer prediction. It finds the optimal hyperplane that separates the data into different classes. It is useful when the data is not linearly separable and maps the data into a higher-dimensional feature space for classification.

Gaussian NB Naive Bayes

GNB is a probabilistic machine learning method that is often used for classification tasks such as breast cancer prediction. GNB is a Naive Bayes variation that assumes input characteristics are independent and normally distributed.In breast cancer prediction, GNB may be used to create a predictive model that divides patients into two groups: those having a positive breast cancer diagnosis and those who do not. GNB uses Bayes' theorem to predict the likelihood of each class given the clinical characteristics of the patient and assumes that the input data are normally distributed.

Decision Tree:

The Decision Tree method is a well-known machine learning technique that may be used for classification problems such as breast cancer prediction. A Decision Tree is a tree-like model that depicts potential outcomes, choices, and their repercussions.In breast cancer prediction, a Decision Tree may be used to create a predictive model that divides patients into two groups: those having a positive breast cancer diagnosis and those who do not. A Decision Tree learns from data to determine the most relevant characteristics and divides the data into subsets that optimize class separation.

Support Vector Machine:

The Support Vector Machine (SVM) is a machine learning algorithm that can be used to predict breast cancer. It works by locating a hyperplane that best separates data into distinct classes. SVM is trained on patient data, which includes characteristics such as age, family history, and tumor size. Once trained, it can predict whether or not a new patient has breast cancer. SVM is an effective treatment decision-making tool.

III. MOTIVATION

A straightforward and efficient machine learning approach for classification tasks, such as predicting breast cancer, is the K-nearest neighbor (KNN) algorithm. The algorithm is an appealing option for breast cancer prediction, especially when there is a large dataset with various features. It is straightforward and can handle high-dimensional data.

Also, the KNN technique is appropriate for datasets with complicated and nonlinear interactions since it does not make assumptions about the distribution of the data. The KNN algorithm is also a non-parametric technique, which makes it resistant to outliers, missing data, and noisy data.

Consequently, the goal of employing the KNN algorithm to predict breast cancer is to offer a precise and effective way for identifying women factors such age, family history, tumor size, and other clinical aspects.In breast cancer prediction, logistic regression is employed as a binary classifier, with the two potential outputs being a positive or negative diagnosis. Given the clinical who may be at a high risk of developing the disease. As a result, patient outcomes will be improved as healthcare providers will be able to deliver the proper interventions, such as greater screening and preventative measures and reducing breast cancer mortality rates.

IV. RELATED WORK

The related work for breast cancer prediction using the K-nearest neighbor (KNN) algorithm includes studies that have achieved high prediction accuracies using different datasets, feature spaces, and values of K. Mangasarian and Wolberg (1990) achieved 96.5% accuracy using the Wisconsin Breast Cancer Dataset, while Vellido et al. (2015) achieved 81.8% accuracy based on clinical and pathological features with K=3. Hussain et al. (2021) used radiomic features extracted from breast MRI images and achieved 77.6% accuracy with K=3. Overall, these studies demonstrate the effectiveness of the KNN algorithm for breast cancer prediction and the importance of selecting an appropriate value of K based on the dataset and problem complexity.

V. METHODOLOGY

Proposed method: K- Nearest Neighbor

KNN (K-Nearest Neighbors) is a simple machine learning algorithm used for classification and regression.The basic idea behind KNN is to find the K nearest data points to the new data point (the one you want to classify or predict) based on a chosen distance metric, and then make a prediction based on the majority class or average value of the K nearest neighbors.
Here are the basic steps for the KNN algorithm:

1.Choose the value of K (the number of neighbors to consider).
2.Compute the distance between the new data point and all other data points in the dataset, using a distance metric such as Euclidean distance.
3.Select the K nearest data points to the new data point. 4.For classification, predict the class of the new data point based on the majority class of the K nearest neighbors. For regression, predict the value of the new data point based on the average value of the K nearest neighbors.

KNN is a simple and intuitive algorithm, but it can be computationally expensive for large datasets, especially when the number of features is high. It also requires careful selection of the value of K and the distance metric, which can greatly affect the performance of the algorithm.

VI. CONCLUSION

We investigated and tested many machine learning methods, and discovered that the accuracy of Logistic Regression, Support Vector Machine, Kernel SVM, Gaussian NB Naive Bayes, and Decision Tree is 95.1%, 95.8%, 93.7%, 93.7%, and 95.8%, respectively. The accuracy of the K Nearest Neighbour method was found to be 96.5% after implementation. In addition, the precision, F1 score, recall, ROC score, and precision were greater when compared to other algorithms. As a result, KNN is one of the top Machine Learning algorithms for breast cancer categorization and prediction. In conclusion, the KNN method is an effective and simple-to-use machine learning technique that can be employed for breast cancer prediction. K must be chosen carefully, taking into account both the difficulty of the task and the size of the dataset. When the right hyperparameters are used, several studies have demonstrated that the KNN algorithm can predict breast cancer with a high degree of accuracy.

VII. REFERENCES

[1]. Navneet Takkar, Usha Rani Dalal, Suman Kochhar, A. K. Pandey, Uma Handa, and Priyanka Garg, “Screening methods (clinical breast examination and mammography) to detect breast cancer in women aged 40–49 years”, National Library of Medicine, J Midlife Health. 2017 Jan-Mar; 8(1): 2–10.

[2]. Xavier Castells, Maria Sala, Margarita Posso, Marta Román, Javier Louro, Laia Domingo, Laia Domingo, "Personalized breast cancer screening strategies: A systematic review and quality assessment", PLoS one research article, December 16, 2019..

[3].Mishra, Gauravi, Badwe, Rajendra, Pimple, Sharmila, Mittra, Indraneel, "Screening for breast cancer Cost-effective solutions for low- & middle-income countries", Indian Journal of Medical Research: August 2021 - Volume 154 - Issue 2 - p 229-236 DOI: 10.4103/ijmr.IJMR_2635_20

[4].Devalland, J. Zuluaga-Gomez, Z. Al Masry, C. Varnier, N. Zerhouni, "A survey of breast cancer screening techniques: thermography and electrical impedance tomography" Journal of Medical Engineering & Technology, Pages 305-322, 2019

[5].Ifeoma A Njelita, Chimezie Innocent Madubogwu, Ngozi Ukamaka Madubogwu, Amobi Ochonma Egwuonwu, "Breast cancer screening practices amongst female tertiary health worker in Nnewi", Journal of Cancer Research and therapeutics, Year : 2017, Volume : 13, Issue : 2, Page : 268-275

[6]Anku Jaiswal,Adhista Chapagain, Ashutosh Ghimire, Amira Joshi,"Predicting Breast Cancer using Support Vector Machine Learning Algorithm", (International Journal of Advanced Computer Science and Applications),ISSN (online): 2581-3048 Volume 4, Issue 5, pp 10-15, May-2022.

[7]Sarika Chaudhary, Neelam Yadav,Yojna Arora, "Optimization of Random Forest Algorithm for Breast Cancer Detection", (International Journal of Innovative Research in Computer Science & Technology), ISSN: 2347-5552, Volume-8, Issue-3, May 2020

[8].Ziba Khandezamin, Mohammad Javad Rashti,MarjanNaderan, "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier",(Journal of Biomedical Informatics), Volume 111, November 2020.

[9].Fadi Al-Tudjman, R. Kumar,V.Nanda Gopal, L. Anand,M.Rajesh, "Feature selection and classification in breast cancer prediction using IoT and machine learning", (International Journal of Advanced Science and Technology), Vol. 29, No. 03, (2020).

[10].Omar Tarawneh,,Moath Husni,Tarawneh, Malek A Almomani, ,Mohammed Otair,"Breast Cancer Classification using Decision Tree Algorithms", (International Journal of Advanced Computer Science and Applications), Vol. 13, No. 4, 2022

[11]. Erdem Yavuz, Springer, An effective approach for breast cancer diagnosis based on routine blood analysis features,published:20 May 2020.

[12] G. Naga Ramadevi Research Scholar Dept. of Computer Science, Tirupati. SPMVV, K. Usha Rani Professor Dept. of Computer Science Tirupati SEAGI, D. Lavanya Professor Dept. of CSE SPMVV, Tirupati , International Journal of Scientific and Innovative Mathematical Research (IJSIMR) Volume 3,Special Issue 2, Importance of Feature Extraction for Classification of Breast Cancer Datasets,July 2015,PP 763-768, ISSN 2347-307X (Print).

[13]Sang Won Yoon,Sarah S.lam,Bichen Zheng, ELSEVIER, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, volume 41,issue 4,part 1,March 2014.

[14].Christopher, published towards data science ,Breast Cancer Prediction using Machine Learning Python ,2017.

[15] Chief Solutions Architect Arohak Inc,Viswanatha Reddy Allugunti , USA, International Journal of Engineering in Computer Science 2022; 4(1): 49-56 , Breast cancer detection based on thermographic images using machine learning and deep learning algorithms, January 2022

[16].Viswanatha Reddy Allugunti Chief Solutions Architect Arohak Inc, USA, Breast cancer detection based on thermographic images using machine learning and deep learning algorithms , International Journal of Engineering in Computer Science 2022; 4(1): 49-56, January 2022

[17]. Vellore Institute of Technology, Gorbachev Road, Vellore, Tamil Nadu 632014, India. , Ramik Rawa School of Computer Science and Engineering (SCOPE), Breast Cancer Prediction using Machine Learning , www.jetir.org (ISSN-2349-5162) , May 2020

[18] Dr. J.Ajayan , T.Daniya , Dr.B.Santhosh Kumar Department of Information Technology, Department of Electronics & Communication, SNS College of Technology, Coimbatore, email2ajayan@gmail.com, Department of Computer Science and Engineering, y, Rajam, Santhosh.b@gmrit.edu.in GMR Institute of Technology, Rajam, daniya.t@gmrit.edu.in Breast Cancer Prediction Using Machine Learning Algorithms, International Journal of Advanced Science and Technology Vol. 29, No. 03, (2020), pp. 7819 - 7828 , 2020

[19] Jean Sunny, Nikita Rane, Prof. Sulochana Devi, Rucha Kanade, Xavier Institute of Engineering, Department of Information Technology, Mumbai - 400016,India., Breast Cancer Classification and Prediction using Machine Learning, International Journal of Engineering Research & Technology (IJERT), February 2022

[20].Siddhi Mhatre, Angela More, Vanshika Patil, Varsha Kamble, Sujata Bhairnallykar, Assistant Professor, Saraswati College of Engineering, Computer Engineering Department, Kharghar, Navi Mumbai,Maharashtra, India - 410210., Breast Cancer Prediction Using Classification Techniques of Machine Learning, International Journal for Research in Applied Science & Engineering Technology (IJRASET) , January 2022