International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com

# Breast Cancer Detection Using Machine Learning Algorithms

Dr. T. Amalraj Victoire,[1] M. Vasuki[2], N. Mohana krishnan[3]
[1]*Professor, Department of MCA, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry – 605107.*
[2]*Associate Professor, Department of MCA, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry – 605107.*
[3]*PG Student, Department of MCA, Sri Manakula Vinayagar Engineering College, Madagadipet, Puducherry – 605107.*

Date of Submission: 06-05-2024 Date of Acceptance: 18-05-2024

**ABSTRACT**
Breast cancer is the most common reason for death due to cancer. It is very necessary to detect cancer at its early stages. We can be able to detect the affected cells by using machine learning algorithm. On an average1 in 39 women have breast cancer as stated by the research results. The existing methods of this cancer were visualized by machine learning techniques including k-nearest neighbors (KNN), support vector machine (SVM), Decision tree, Naïve Bayes, and Random-forest were applied to the dataset. Before, beginning the data processing, we have to sure that the data is already pre-processed. The previous research is conducted through many machine learning techniques with some enhancement and augmentation in the dataset for better performance. But it is observed that machine learning methods gives finer results on linear data. It is also concluded from the previous research when the data is in the form of images where the machine failed. Machine learning not only just provides the diagnosis of cancer with better accuracy as compared to manual diagnosis, but also helps in minimizing the cost and time of diagnosis.

**Key words:** Machine Learning, Cancer, Supervised Learning, Naive Bayes, Decision Tree

## I. INTRODUCTION:

Breast cancer ranks as one of the most commonly occurring cancers in women globally. Just like any other cancer, it initiates when healthy cells undergo changes and start growing abnormally, forming a cluster known as a tumor. Doctors explain that breast cancer arises from the abnormal growth of cells within the breast, with the potential for these cells to spread beyond the breast to lymph nodes or other body areas. Detecting and halting the growth of these unwanted cells is crucial. In developing nations, breast cancer constitutes 23% of all cancer cases, with an estimated 1.6 million new cases affecting women worldwide. A benign tumor refers to a growth in a specific body part. Breast cancer manifests in four types: Ductal Carcinoma in Situ, located in the lining of breast milk ducts and considered a precursor to breast cancer; the most prevalent type, accounting for 70-80% of diagnoses; Inflammatory Breast Cancer, a rapidly advancing form where cancer cells invade the breast's skin and lymph vessels; and Metastatic Breast Cancer, which spreads to other body parts.

## II. LITERATURE REVIEW

[1] Breast cancer remains the leading cause of cancer-related deaths and early detection is crucial. Various Machine Learning methods are available for breast cancer diagnosis. This study introduces a Machine Learning model for automated breast cancer diagnosis, using a CNN as a classifier and Recursive Feature Elimination (RFE) for feature selection. Additionally, the study compares five algorithms: SVM, Random Forest, KNN, Logistic Regression, and Naïve Bayes. The model was tested on the BreaKH is 400X Dataset and evaluated based on accuracy and precision. The findings indicate that CNN outperforms existing methods in terms of accuracy, precision, and dataset size.

[2] Breast cancer is a prevalent disease affecting millions of women worldwide. To reduce unnecessary breast biopsies, computer-supported frameworks have been proposed recently. Machine Learning offers significant potential for computer-aided disease diagnosis. This dataset is widely used due to its large, nearly noise-free instances. Most machine learning algorithms achieve prediction accuracies exceeding 95%, demonstrating the

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com

effective use of machine learning in breast cancer prediction.

[3] Breast cancer is highly prevalent among women in the UAE and globally. Early and accurate diagnosis is crucial for effective treatment. However, mammogram-based detection faces uncertainties. Machine Learning techniques can develop tools for physicians to enhance early breast cancer detection, thereby improving patient survival rates. This study compares three popular ML techniques for breast cancer detection—Support Vector Machine (SVM), Random Forest (RF), and Bayesian Networks (BN). Results show that SVMs offer the highest accuracy, specificity, and precision, while RFs excel in correctly classifying tumors.

[4] As per recent data, breast carcinoma ranks as the most widespread cancer globally, resulting in nearly 900 thousand deaths annually. Early detection and accurate diagnosis can significantly enhance survival rates and reduce fatalities. Identifying tumors relies on machine learning algorithms that excel at pattern recognition, particularly in distinguishing between malignant and benign conditions. The CNN achieved an accuracy of 99.67%, while SVM and RF scored 89.84% and 90.55% respectively. This method could contribute to the advancement of more sophisticated CAD systems in the future.

[5] Breast cancer is highly prevalent among Indian women, with a 50% fatality rate among diagnosed cases. This study compares popular machine learning algorithms like Random Forest, KNN, and Naïve Bayes for breast cancer prediction using the Wisconsin Diagnosis Breast Cancer dataset. Each algorithm achieved over 94% accuracy in distinguishing between benign and malignant tumors, with kNN demonstrating the best accuracy, precision, and F1 score among them.

[6] Predicting breast cancer is crucial for medical decision-making. This study evaluates various machine learning algorithms based on decision tree learners using routine blood analysis. The algorithms are categorized into basic decision tree, random forest, and gradient boosting, with the gradient boosting algorithm combined with feature selection achieving the highest sensitivity and specificity scores (85% and 80% respectively).

[7] Cancer ranks as the second leading cause of death worldwide, largely due to delayed detection. Approximately 60% of breast cancer cases are diagnosed at advanced stages. This paper focuses on enhancing an image processing algorithm for early breast cancer detection using X-ray mammogram images. Pre-processing techniques such as Gaussian filtering and edge detection improve image quality, while features are extracted using Wavelet Transform and GLCM. The proposed algorithm achieved a 96% accuracy in early breast cancer diagnosis using the DNN classification algorithm.

[8] According to the Breast Cancer Organization, breast cancer is one of the most dangerous diseases affecting women worldwide. In this study, an ensemble voting machine learning technique was used for breast cancer analysis. The proposed approach achieved an impressive precision of 98.50%. The analysis focused on 16 specific features related to breast cancer.

[9] This paper provides an overview of the evolution of big data in healthcare and applies four learning algorithms to a breast cancer dataset. The goal is to predict breast cancer, a major cause of death among women globally, emphasizing the importance of early detection and prevention. The machine learning algorithms used include Random Forest, Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (K-NN), with SVM demonstrating the highest accuracy at 97.9%. These findings aid in selecting the most effective machine learning algorithm for breast cancer predictions.

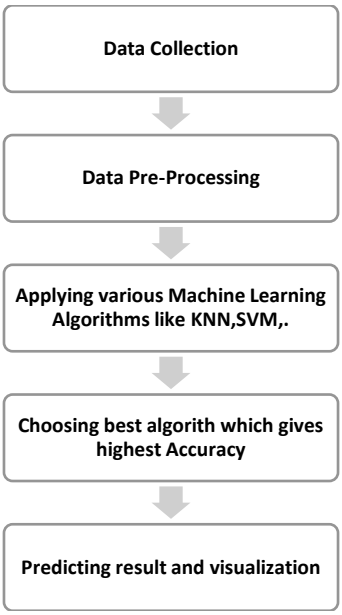
[10] Breast cancer is a type of cancer originating from breast tissue, and the first signs of it are breast lumps or an abnormal mammogram. There are two main types of biopsies: needle biopsy and surgical biopsy. Needle biopsy types include fine needle biopsy (FNA), core needle biopsy, vacuum assisted breast biopsy. Through FNA, data on cytological characteristics can be obtained and assessed for breast cancer diagnoses. Table and graphs of the accuracy results of the WHAVE methods, each individual ML method, and the un-weighted majority voting method are produced to show the minimum, maximum and average accuracies of each method. Graphs that show the difference in accuracy between 90% and 10% training data and the variation of each method are generated.

**PROPOSED WORK**

One of the solutions that we have done in this paper is we have compared various algorithms such as logistic regression, KNN, Naïve bayes, decision tree etc to predict the accuracy.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com

**ARCHITECTURE DIAGRAM**




**MODULE DESCRIPTION**

**A.Datasets:**

We acquired the Breast Cancer Wisconsin (Diagnostic) Dataset from Kaggle, containing data from 569 patients. Each instance includes 32 attributes along with a diagnosis and features. The dataset features numeric values representing cancerous and non-cancerous cells, and our goal is to predict cancer based solely on these features. The 'Target' attribute indicates whether the patient has 'Benign' (non-cancerous) or 'Malignant' (cancerous) cancer. 'Benign' signifies the absence of cancer, while 'Malignant' indicates the presence of cancer.

**Data Set Characteristics:**

Number of Instances: 569

Number of Attributes: 30 numeric, predictive attributes and the class

Attribute Information:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the threelargest values) ofthese features were computed for each image,resulting in 30 features. For instance, field 3 is Mean Radius, field13 is Radius SE, field 23 is Worst Radius.

- class:
  - WDBC-Malignant
  - WDBC-Benign

Summary Statistics:

| | Min | Max |
|---|---|---|
| radius (mean): | 6.981 | 28.11 |
| texture (mean): | 9.71 | 39.28 |
| perimeter (mean): | 43.79 | 188.5 |
| area (mean): | 143.5 | 2501.0 |
| smoothness (mean): | 0.053 | 0.163 |
| compactness (mean): | 0.019 | 0.345 |
| concavity (mean): | 0.0 | 0.427 |
| concave points (mean): | 0.0 | 0.201 |
| symmetry (mean): | 0.106 | 0.304 |
| fractal dimension (mean): | 0.05 | 0.097 |
| radius (standard error): | 0.112 | 2.873 |
| texture (standard error): | 0.36 | 4.885 |
| perimeter (standard error): | 0.757 | 21.98 |
| area (standard error): | 6.802 | 542.2 |
| smoothness (standard error): | 0.002 | 0.031 |
| compactness (standard error): | 0.002 | 0.135 |
| concavity (standard error): | 0.0 | 0.396 |
| concave points (standard error): | 0.0 | 0.053 |
| symmetry (standard error): | 0.008 | 0.079 |
| fractal dimension (standard error): | 0.001 | 0.03 |
| radius (worst): | 7.93 | 36.04 |
| texture (worst): | 12.02 | 49.54 |
| perimeter (worst): | 50.41 | 251.2 |
| area (worst): | 185.2 | 4254.0 |
| smoothness (worst): | 0.071 | 0.223 |
| compactness (worst): | 0.027 | 1.058 |
| concavity (worst): | 0.0 | 1.252 |
| concave points (worst): | 0.0 | 0.291 |
| symmetry (worst): | 0.156 | 0.664 |
| fractal dimension (worst): | 0.055 | 0.208 |

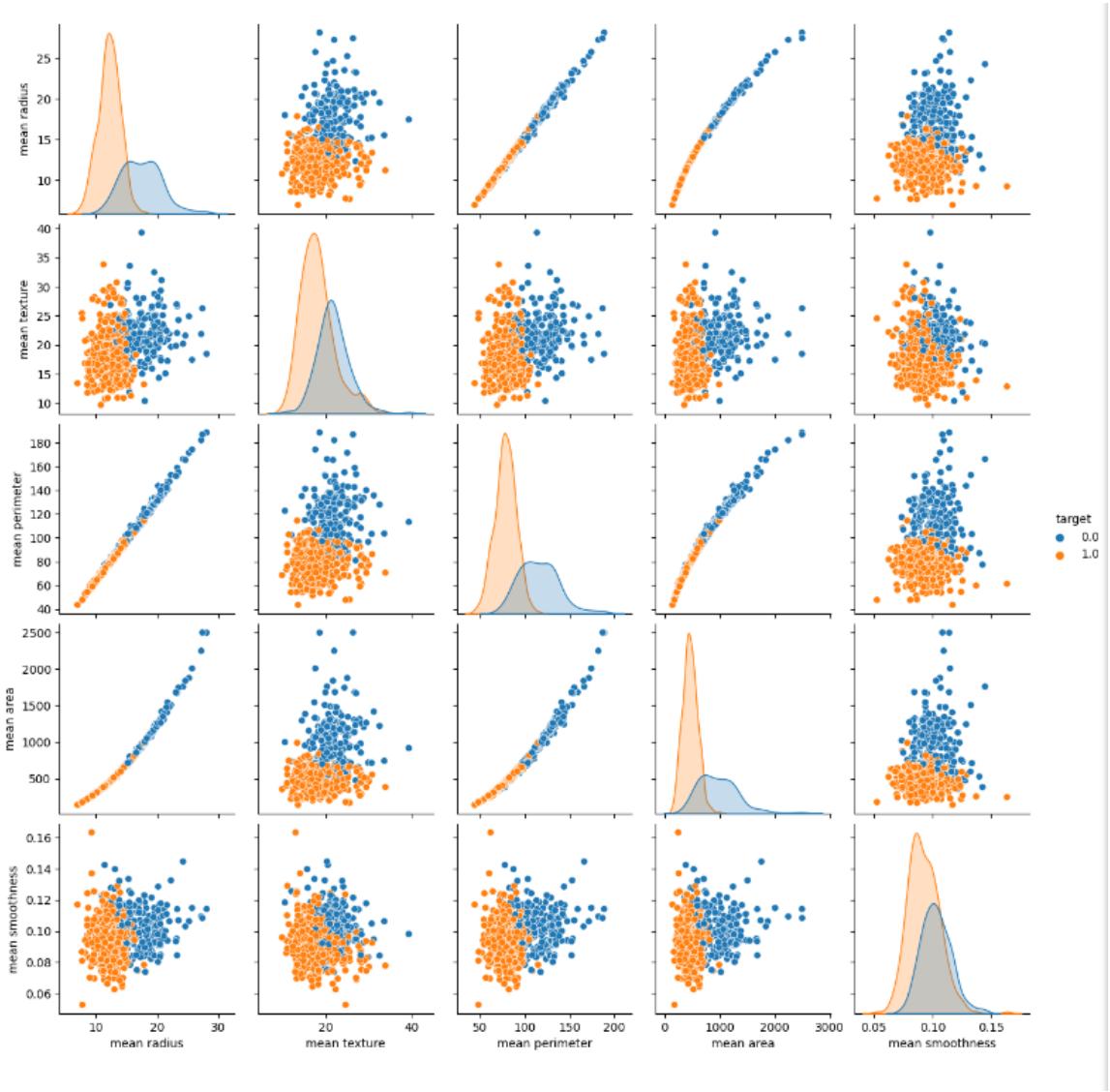
**B. Data pre-processing:**

It is a technique which is used to convert primary data or raw data into a clean dataset to make it suitable for a building and training machine learning models.

**C. Visualization:**

It is a graphical representation of the given data by using various visual elements such as charts, graph or another visual format.

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com




**D. Machine learning algorithm:**

Machine learning algorithms are tools used to transform a dataset into a model. In this study, we evaluated different algorithms like logistic regression, random forest, naive Bayes, and decision trees to determine whether a patient has breast cancer or not.

**E. Prediction**

Prediction gives result of an algorithm. Prediction takes the data as input and predict future value of data.

**MACHINE LEARNING ALGORITHMS**

Machine Learning is a process that machines (computers) are trained with data to make the decision for similar cases. ML is employed in various applications, such as object recognition, network, security, and healthcare. There are two ML types i.e. single and hybrid methods like ANN, SVM, K-Nearest Neighbor (KNN) etc.

Following are the used ML algorithms:

**A. SUPPORT VECTOR MACHINE (SVM)**

Support Vector Machine (SVM) is a supervised pattern classification model utilized as a training algorithm to learn classification and regression rules from collected data. Its objective is to segregate data until a hyperplane with a significant minimum distance is identified. SVM can be applied to classify two or more data types and encompasses various single or hybrid models like Standard SVM (St-SVM), Proximal Support Vector Machine (PSVM), Newton Support Vector Machine (NSVM), Lagrangian Support Vector Machines (LSVM), Linear Programming Support

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com

Vector Machines (LPSVM), and Smooth Support Vector Machine (SSVM).

**B. K-NEAREST NEIGHBORS(KNN):**

K-NN , it is a Supervised type of M L Algorithm which is widly used for Classification and also for Regression problems and mainly used for the Classification problems. It also helps to stores all the avail data. It has good algorithms, which stores all available data and it will also classifing their new data point with based on their similarity. When we insert new records it can also find the result for the new records based on their similarity of the old records. It does not makesit own assumption but instead it will take decision based on their underlying data.

**C.DECISION TREE ALGORITHM (DT):**

Decision Tree (DT) is a data mining method employed for the early identification of breast cancer. It represents classifications or regressions in a tree-like structure. The model divides the dataset into smaller sub-data, further breaking them down into smaller components. This process leads to the development of a tree structure, with the final outcome revealed at the last level. In this tree structure, the leaves represent class labels, while the branches represent combinations of features that lead to these labels. Consequently, DT is robust against noise.

**D.RANDOM FOREST ALGORITHM(RF):**

Random forest , it is Supervised type of ML Algorithm which is used widely in the Classification and also the Regression problems. One of the main part of these Algorithm is that it can also handle the data set containing the variables which is continuous and also as in the case of regression. It performs best for classification problems. It is also like the concept of the ensemble type learning, this is the process of combining the multiple classifiers for solving the complex problem.

**E. DECISION TREE:**

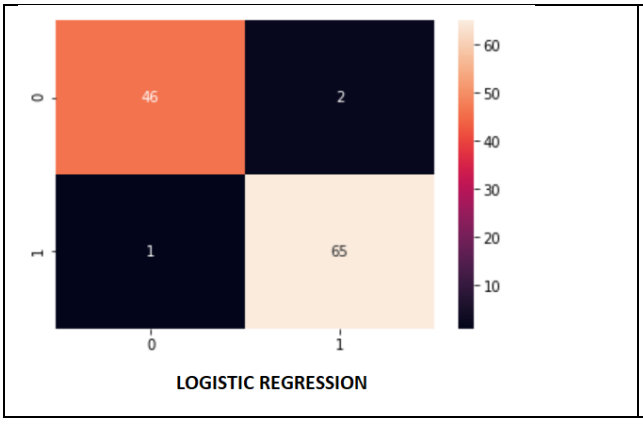
Decision Tree, it is a Supervised type machine learning technique, it also used for both the classifications and the Regressions problem, but mainly it prefers for solving the Classification problems. It is the tree-structured classifier, which has internal nodes which represent the feature of the data, branches represents the decisions rules and then each the leaf represents the output/outcome. The test or the decision is performed on their feature basis from the given datasets. It is like a graph representation, it gets all their possible solutions for the problem/decisions based on their given condition.

**F. LOGISTIC REGRESSION:**

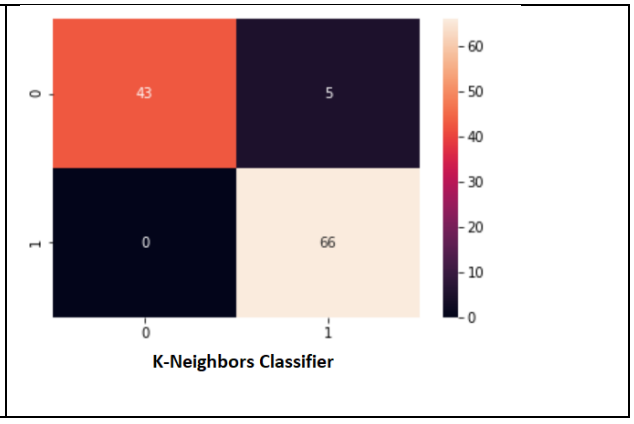
Logistic Regression, it is the popular ML Algorithm, comes under supervised Learning Technique. It is used in order to predict the categorial dependent variables with the use of independent variables. It also predict the outputs for the categorical dependent variables. In my project it is helpful inorder to predict the result for the attribute , which is patient has cancer or not. But the output can be a categorical value or discrete value like yes/no ,0 or 1. Mostly recommended for solving classification based problems.

## III. Experimental Results

Following visualizations are heatmap and results of machine learning algorithms like KNN, SVM, Random Forest Classifier, Decision Tree etc. applied to Kaggle dataset.

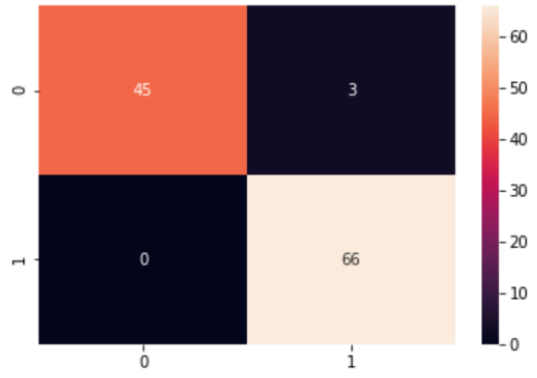


LOGISTIC REGRESSION



K-Neighbors Classifier

International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com




SVM

Accuracy: 0.9736842105263158

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 1.00 | 0.94 | 0.97 | 48 |
| 1.0 | 0.96 | 1.00 | 0.98 | 66 |
| micro avg | 0.97 | 0.97 | 0.97 | 114 |
| macro avg | 0.98 | 0.97 | 0.97 | 114 |
| weighted avg | 0.97 | 0.97 | 0.97 | 114 |

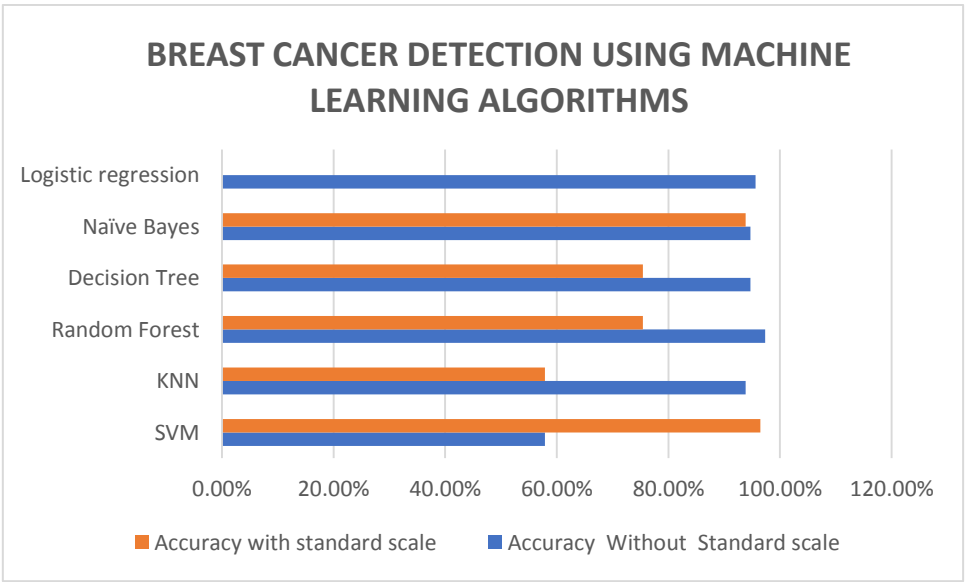
Random Forest Classifier

Decision Tree Accuracy: 0.956140350877193

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.98 | 0.92 | 0.95 | 48 |
| 1.0 | 0.94 | 0.98 | 0.96 | 66 |
| micro avg | 0.96 | 0.96 | 0.96 | 114 |
| macro avg | 0.96 | 0.95 | 0.95 | 114 |
| weighted avg | 0.96 | 0.96 | 0.96 | 114 |

## IV. Experimental analysis

In our paper, we used Jupyter Notebooks and Python. We explored various supervised learning algorithms, including Support Vector Classifier (SVM), Random Forest, Naïve Bayes, Decision Tree, KNN, and Logistic Regression. The dataset had features with different units and magnitudes, so we standardized them using Standard Scaling in SKLearn. Model selection is crucial in Machine Learning, and we focused exclusively on supervised learning. We applied these methods to predict outcomes and recorded their accuracies.



International Journal of Engineering, Management and Humanities (IJEMH)
Volume 5, Issue 3, May.-June, 2024 pp: 70-76 ISSN: 2584-2145
www.ijemh.com

## V. CONCLUSION:

This paper involves in comparing among different algorithm on machine learning (SVM, KNN, Random-forest, decision tree, Naïve bayes, logistic regression) for breast cancer detection. Among those algorithm, SVM algorithm gives the best accurate results to other comparing algorithms.

**FUTURE WORKS:**

Algorithms such as CNN, linear regression, k-means clustering may be used in this dataset in future. In this dataset we can apply many algorithms and better outcome will be detected. So, we can give accurate information to user. Therefore, for people and the government this will be helpful. But in future various classification algorithms can be used for detection purposes.

## References:

[1]. Sweta Bhise, Simran Bepari, Shrutika Gadekar, Deepmala Kale, Aishwarya Singh Gaur, Dr. Shailendra Aswale ," Breast Cancer Detection using Machine Learning Techniques",(2021),

[2]. Harinishree M. S., Aditya C. R., Sachin D. N. "Detection of Breast Cancer using Machine Learning Algorithms– A Survey",2021

[3]. Dana Bazazeh and Raed Shubair , "Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", 2016

[4]. Viswanatha Reddy Allugunt, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms",2022

[5]. Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury," Breast Cancer Detection Using Machine Learning Algorithms", 2018

[6]. Fiddin Yusfida A'la, Adhistya Ema Pennanasari, Noor Akhmad,"A Comparative Analysis ofTree-based Machine Learning Algorithms for Breast Cancer Detection", 2019

[7]. R. Chtihrakkannan, P.Kavitha, T.Mangayarkarasi, R.Karthikeyan, "BreastCancer Detection using Machine Learning", 2019

[8]. Karthikeyan B, Sujith Gollamudi, Harsha Vardhan Singamsetty, Pavan Kumar Gade , Sai Yeshwanth Mekala, "Breast Cancer Detection Using Machine Learning", 2020

[9]. Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru , "BreastCancer Detection using Machine Learning Way",2019

[10]. Youness Khourdifi, Mohamed Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification",2018