International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

Breast Cancer Prediction

Shreya S Kshirasagar[1], Mr. Ramesh K[2]

[1]*Student, Dept of Electronics and Communication Engineering, University BDT College of Engineering, Karnataka, India.*
[2]*Professor, Dept of Electronics and Communication Engineering, University BDT College of Engineering, Karnataka, India.*

Abstract - *Nowadays detecting breast cancer is extremely important in the medical world. One of the major malignancies that may afflict women is breast cancer, which can be quite harmful. Breast cancer (BC) includes two types: benign (non-cancerous) and malignant (cancerous). Malignant is listed as a type of cancer that can be cured, whereas benign is listed as a disease that cannot be cured. BC symptoms include changed genes, excruciating pain, size and shape, variations in the color (redness) of the breasts, and changes in the texture of the skin. For the prediction, machine learning methods are employed. Six different classification techniques, which includes Decision Tree, Logistic Regression, Support Vector Machine (SVM), Random Forest, K Nearest Neighbor (KNN), Naive Bayes, and others, are used to identify breast cancer. These algorithms fall under category of supervised machine learning. These methods are used to predict the development of breast cancer. These algorithms' accuracy results are assessed. When compared to the other algorithms utilized, it is found that Logistic Regression and Random Forest have the highest accuracy rate, reaching up to 96.49 percent.*

Key Words: **Breast Cancer, Machine Learning, Logistic Regression, SVM, Decision Tree, KNN, Naïve Bayes.**

1. INTRODUCTION

Changes or abnormalities in the genes that support cell development are what cause the condition known as cancer. These alterations enable the cells to proliferate and multiply in an erratic and disorderly fashion. These modifications provide the cells the ability to replicate and proliferate in an irregular and disordered manner. These abnormal cells eventually develop into a tumor. Even if the body doesn't require tumor's, they don't die like other cells do. One type of cancer that develops in the breast cells is breast cancer. This kind of cancer may show up in breast lobules or ducts.

In addition, the fatty tissue and fibrous connective tissue of the breast can develop into cancer. These cancer cells become uncontrolled as they grow, invade other healthy breast tissues, and have been known to invade the lymph nodes beneath the arms. There are two types of cancer: malignant and benign. Cancers that are malignant are cancers. These cells keep dividing uncontrollably and start affecting other cells and tissues in the body. This form of cancer is challenging to treat since it has spread to every other region of the body [1].

These tumors' can be treated with chemotherapy, radiation therapy, and immunotherapy, among other types of therapies. Since benign cancer is not carcinogenic it does not grow to other parts of body, it is far less harmful than malignant cancer. Such tumors frequently don't actually need to be treated. Women over 40 are the ones who get diagnosed with breast cancer most frequently. But any age of women can contract this illness. It may also happen if breast cancer runs in the family. According to data, breast cancer alone holds around 25% of women diagnosed with cancer and 15% of women's death due to the cancer globally, and it has historically had a high mortality rate [2]. Since scientists have long been aware of its risks, extensive study has been done in an effort to identify the best cure.

A machine learns progressively on its own through a process called machine learning (ML). The ML model serves as a mathematical tool for artificial intelligence. Artificial intelligence is a machine that thinks for itself and emulates human intelligence. The machine improves at its job as it gains more "experience," just like a human does. Machine learning is described as "the mechanism by which a computer operates more correctly as it acquires and learns from the data given" when viewed as a methodology.

Machine learning algorithm are classified into supervised and unsupervised algorithms. In this work classification algorithms are used which comes under supervised type of learning. With the help of the algorithms, it is able to predict the result.

2. LITERATURE REVIEW

The author of [1] proposed that the breast cancer can be predicted by using dataset extracted from Wisconsin Breast Cancer repository. The data set has 569 data points with 30 Attributes. The accuracy obtained by LogisticRegression is about 96.5%.

In [2] the author gives the comparison of ML algorithms for breast cancer prediction. The paper employs the decision tree and logistic regression machine learning methods. In this paper WDBC dataset is used which contains 570 rows and 32 columns. The paper lists out that the logistic

regression gave 94.4%Accuracy whereas Decision Tree gave about 95.14% accuracy, hence decision tree algorithm was chosento make the predictions more accurately.

In [3] the author proposed work on identifying breast cancer risk factor with the help of Machine learning algorithms. The Support Vector Machine classifieris used in comparison with naïve Bayes classifier.

For the breast cancer prediction, the Wisconsin diagnosis breast cancer dataset was utilized. The SVM algorithm performed Excellently, exhibiting accuracy up to97.91% as compared to NB algorithm which gave 95.6% accuracy.

In [4] the author proposed work on Breast Cancer analysis using K Nearest Neighbor algorithm. The author used KNN technique for making the prediction of breast cancer. By using Manhattan distance with K = 1, yields an accuracy around 98.40% whereas Euclidean distance with K= 1, yields a high accuracy of about 98.70%.

In [5] the author worked on an intelligent system employing SVM based classifier for predictive breast cancer detection and prognosis. Support vector machines (SVMs)-based classifiers outperform Bayesian classifiers and artificial neural networks for the diagnosis and prognosis of breast cancer sickness. The enhanced SVM method performed admirably, displaying high values for great significant to 96.91, specificity up to 97.67 percent, and sensitivity up to 97.84 percent.

3. PROPOSED WORK

3.1 Objective

Breast cancer is a disease which we hear about a lot nowadays. It is one of the most widespread diseases. It is important to identify the disease so that women may start treatment as soon as possible. It is best for a correct and early diagnosis. The primary goal of this work is to help pathologist to predict the type of cancer at a faster rate.

As machine are capable of calculating the results faster than humans do and it can also repeat itself thousands of times without being exhausted. So whenever huge amount of data arrives machine can predict more than ten thousand iterations per second, which reduces the amount of time required for pathologist to analyze the biopsy report. Also results obtained by using machine learning algorithms are more accurate.

3.2 Methodology

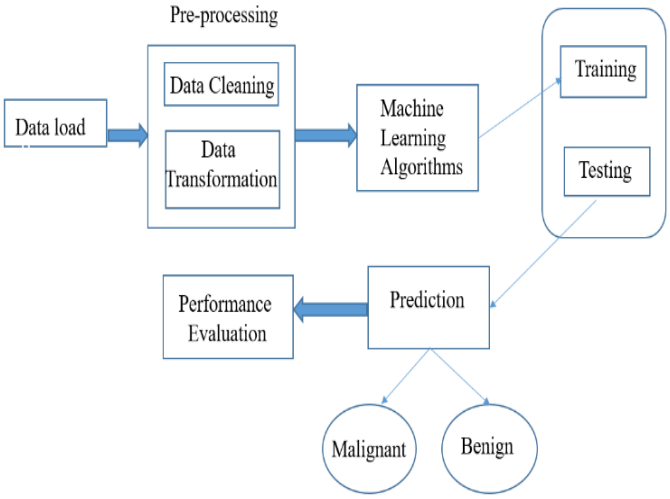


Figure 1: Block Diagram of proposed work

The figure 1 illustrates the block diagram of the proposed work. The data is loaded, run over machine learning algorithms, the data is split into training set and test set. Finally, the test set data is used for making the predictions.

A. Data Collection

The WBCD dataset utilized in this study was created by Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, in United States.

The dataset contains 357 benign and 212 malignant breast cancer patients, respectively. The dataset comprises 32 columns, with the ID number being the first column and the diagnosis outcome (0-benign and 1-malignant) being the second column. worst) of ten features.

These features represent the shape and size of the target cancer cell nucleus. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure in biopsy test. For each cell nucleus, these features are determined by analyzing under a microscope in a pathology laboratory.All values of the features are stored up to four significant digits. There were no null entries in the dataset. The 10 real-value features are described in the Table 1.

Feature Name	Feature Description
Radius	Average of distance from center to circumference points.
Texture	Standard deviation of gray scale value.
Perimeter	Gross distance between the snake points.
Area	Total number of pixels on the inside of the snake along with one half of the pixels in the circumference.
Smoothness	Local variance in length of radius, quantified by calculating the length difference
Compactness	Perimeter ^2/ Area
Concavity	Intensity of the contour concave points
Concave points	The number of contour concavities.
Symmetry	The difference in length between lines perpendicular to the major axis in both directions to cell boundary.
Fractal Dimension	Coastline estimation. A higher value leads to a less normal contour representing a higher risk of malignancy.

Table 1: Feature Description

B. Data Preprocessing

When it comes to creating a machine learning model, data pre-processing is the first step marking the initiation of the process. Typically, real world data is incomplete, inconsistent, inaccurate (contains errors and outliers), and often lacks specific attribute values/trends. This is where the data pre-processing enters the scenario- it helps to calm, format, and organize the raw data, thereby making it ready-to-go for machine learning models.

3.3 Algorithm

1. Logistic Regression

Logistic Regression is a mathematical function which is utilized to convert expected values into probabilities. Any true worth between 0 and 1 can be changed into another value. The value of the logistic regression must be between 0 and 1, and it cannot be more than this value. As a result, it takes shape of a "S" curve. The S-form curve is also known as the logistic function or sigmoid function. In logistic regression, we use the threshold value notion, which determines the probability of either 0 or 1.

2. Decision Tree

Decision Tree (DT) is a powerful machine learning algorithm used for both classification as well as regression. DT can be a tree type of structure where each internal node is a test condition for the vector to move further and the terminal nodes represent the class or the prediction value to be predicted. DT is good for the classification of a few class labels but do not produce proper results if there are many classes and less training observations. And moreover, DTs can be expensive to train computationally.

3. K-Nearest Neighbor

K-Nearest Neighbour (KNN) is said to be the simplest and the most straightforward classification algorithm. Like most machine learning algorithms, K-NN does not learn anything from the provided dataset and its attributes, but simply use the points from the training data and finds the K number of nearest neighbors to that data point using Euclidean Distance and classify it to the class which has the first K neighbors closest to it.

4. Random Forest:

Random Forest is a very well machine learning algorithm that is used in the supervised learning methodology. The Random Forest classifier aggregates the outcomes from several decision trees applied to distinct subgroups of the input dataset in order to increase the predicted efficiency of the input dataset. Rather of depending on a decision tree classifier, the random forest utilizes predictions from all of the trees to forecast the final result based on votes of the majority of predictions.

5. Support Vector Machine:

Support vector machine (SVM) is a quite simple classification algorithm. This classifier is named so because it takes the help of vectors in the feature space to classify the class of a new vector. The Maximum Margin Hyper-plane (MMH) decides whether the new vector belongs to class one or class two. If the data point lies beyond the negative hyper-plane or to the left of MMH then it belongs to the class one, else it belongs to the class two, where class one and two are two different classes in a given situation. SVMs can also be used if there are more than two classes.

6. Naïve Bayes:

Naïve Bayes (NB) theorem is a machine learning algorithm that works on the probability concepts.

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

Where P(A) is the Prior Probability, P(B) is the Marginal Likelihood, P(B|A) is the Likelihood and P(A|B) is the Posterior Probability. The NB algorithm follows the above equation for the determination of the class of a data point. The posterior probability is calculated based on the position of the vector in the feature space and then the data point is assigned to the class with greater posterior probability.

3.4 Training the Model

Overfitting is a problem that arises frequently during model training. This issue arises when a model performs remarkably well on the data, we used to train it but struggles to generalize effectively to new, undiscovered data points. Contrarily, under fitting happens when the model performs poorly even when tested on the training set of data.

The most common approach for identifying these types of issues is to develop a number of data samples for the model's training phase and testing phase. Following the analysis, we will train the machine with the 80% of the data. In this context, "training" refers to instructing the computer with data.

3.5 Testing the Model

After training the machine on the first 80% of the data, the remaining 20% of the data points are utilized to check its efficiency. In other words, we may also quantify how much specific process knowledge the machine acquired.

Diagnosis:

The patient's test results or the doctor's evaluation of the patient's medical history are both used to make the diagnosis of breast cancer. In order to identify whether the patient has normal and cancerous cancer, here 32 characteristics are utilized. Where, B= 357 AND M= 212.

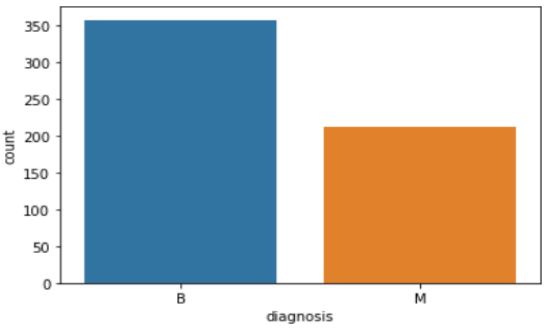


Figure 2: Count of Benign and Malignant tumor

Heat map:

The heat map of the correlation between the WBCD dataset's feature sets is displayed in the figure 20 as shown below. The value of the first dimension is seen as the heat map's row, while the value of the second dimension is seen as its column. This creates a two-dimensional correlation matrix between the two discrete dimensions.

3.6 PREDICTION:

After Machine learning model is fit, the model can predict whether the patient has Malignant type that is patient is suffering from cancer or Benign type that is patient does not have cancer by implementing six different machine learning algorithms of available dataset.

Accuracy:

A confusion matrix is a table that is often used to describe the performance of a classification model or a classifier on a set of test data for which the true values are known.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Figure 4: Confusion matrix

$$ACCURACY = \frac{TP + TF}{TP + TN + FP + FN}$$

Where,

TP= True Positive

TN= True Negative

FP= False Positive

FN= False Negative

A confusion matrix made up of TP, FP, TN, & FN is created for the true and projected results in order to determine the accuracy of each method utilised.

4. RESULT AND DISCUSSION

Following the implementation of machine learning algorithm, the whole dataset is divided into training set and test set. Where 80% of dataset that is about 456 samples of data is given for training the machine and remaining 20% of data that is about 113 samples of data is given for training, which is further used to predict the outcome.

Machine learning algorithms' performance is evaluated and contrasted. The result represented in the table demonstrate the predicted accuracy of training and testing dataset for Logistic Regression, Decision Tree, KNN, Random Forest, SVM and Naïve Bayes.

Algorithm	Training accuracy	Testing accuracy
Logistic Regression	94.28%	96.49%
Decision Tree	100%	92.98%
K Nearest Neighbor	94.72%	92.98%
Random Forest	99.78%	96.49%
SVM	97.14%	95.61%
Naïve Bayes	94.06%	92.10%

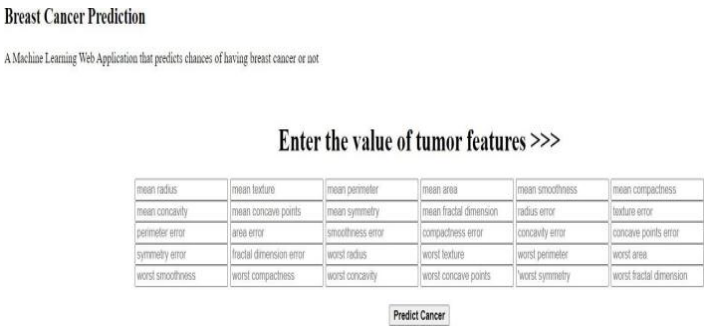
Table 2: Comparison between training and testing accuracy

From the table 2, it is observed that Logistic Regression and Random Forest algorithms has achieved higher accuracy that is about 96.49% for the prediction of breast cancer.

Web Application used to predict Breast Cancer can be done using FLASK, Graphic User Interface (GUI) can be created using HTML.

INPUT:

The web application for prediction of breast cancer is created.The user need to enter the values of the 30 Features as included inthe dataset extracted from Wisconsin breast cancer repository.



Snapshot 1: Input of the model

5. RESULT:

The values entered by the user is given to the web application which is used to predict the result that is whether the patient has breast cancer or patient does not have breast cancer.

Snapshot 2: Output of the model

6. CONCLUSION

Breast cancer is currently one of the most dangerous illnesses that affects women. It is the leading cause of death for women. This study utilized the Wisconsin breast cancer dataset, and several machine learning algorithms were used to incorporate the efficiency and usefulness of the algorithms in order to identify the categorization of normal and cancerous breast cancer that had the best accuracy.

7. FURURE SCOPE

In future it is able achieve greater efficiency by selecting the best attribute from the dataset and by also increasing the number of dataset. Artificial Neural Networks can be applied to make the predictions because it uses the hidden layer to make better and smarter predictions

REFERENCES

1] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," *2021 International Conference on Artificial Intelligence (ICAI), 2021.*

2] A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and

diagnosis," *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), 2018.*

3] Sultana, Jabeen, Abdul Khader Jilani, & "Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers." International Journal of Engineering & Technology [Online], 7.4.20 (2018): 22-26. Web. 30 Nov. 2019.

4] P.Sathiyanarayanan, S.Pavithra, M.Sai Saranya, and M.Makeswari, " Identification of breast cancer using The Decision Tree Algorithm," *2019 IEEE International Conference on System, Computing, Automation and Networking(ICSCAN), 2019.*

5] M.D. Bakthavachalam, Dr.S. Albert, Antony Raj, "Breast Cancer Analysis using K-NearestNeighbor Algorithm", *2020 International Conference on Artificial Intelligence(ICCS), 2020.*

6] Puneet Yadav et al. "Diagnosis of Breast Cancer using Decision Tree Models and SVM", International Research Journal of Engineering and Technology, Vol. 5, Issue 3, Mar 2018.

7] Medjahed, Seyyid Ahmed, Tamazouzt Ait Saadi, and Abdelkader Benyettou. "Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules." International Journal of Computer Applications 62.1 (2013).

[8] Kriti Jain et al. "Breast Cancer Diagnosis Using Machine learning Techniques", International Journal of Innovative Science, Engineering & Technology, Vol. 5, Issue 5, May 2018.

[9] Zheng, Bichen, Sang Won Yoon, and Sarah S. Lam. "Breast cancer diagnosis based on feature extraction using a hybrid of Kmeans and support vector machine algorithms." Expert Systems with Applications 41.4 (2014): 1476-1482.

[10] Puneet Yadav et al. "Diagnosis of Breast Cancer using Decision ree Models and SVM", International Research Journal of Engineering and Technology, Vol. 5, Issue 3, Mar 2018.

[11] Medjahed, Seyyid Ahmed, Tamazouzt Ait Saadi, and Abdelkader Benyettou. "Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules." International Journal of Computer Applications 62.1 (2013).

[12] Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." Journal of Algorithms & Computational Technology 12.2 (2018): 119-126.