BioGecko

Vol 12 Issue 02 2023
ISSN NO: 2230-5807

COMPARATIVE ANALYSIS OF SUPERVISED MACHINE LEARNING ALGORITHMS USED FOR PREDICTING BREAST CANCER

Varsha Narayanrao Ikhe
Department of Computer Science, Indira College of Commerce and Science, Pune, Maharashtra, India.
Email:varsha.ikhe@iccs.ac.in

Shilpa Suhas Pawale
Department of Computer Science, Indira College of Commerce and Science, Pune, Maharashtra, India.
Email:shilpa.pawale@iccs.ac.in

Shweta Nitin Banait
Department Computer Science Engg., DY Patil International University, Akurdi, Pune Maharashtra, India.
Email:shweta.banait@dypiu.ac.in

Abstract
Breast cancer is most frequently found in women all over the world. In the survey about two million women were infected annually. Without treatment, cancer can cause serious health problems and even loss of life. Early detection of cancer may reduce mortality and morbidity. Many of the after the report due to unavailable of expert doctor's consultation people from rural areas face fatal issues and the delaying treatment leads to last stage of the disease. At this stage the chances of survival is very less. There are different machine learning algorithms used for prediction of breast cancer in early stage in this this research paper we will explore the comparative study of breast cancer detection algorithm which gives more accuracy for detecting breast cancer.

Keywords: Breast cancer, Wisconsin Diagnosis Breast Cancer Detection(WBCD), Machine Learning, SVM, KNN, Logistic Regression, Random forest, Decision tree.

I. INTRODUCTION

One of the primary causes of women's deaths is breast cancer. After lung cancer, it is the second most lethal cancer. Like all cancers, breast cancer starts when healthy cells undergo a transformation and begin to grow erratically, generating a clump of cells known as a tumour. A tumour may be malignant or benign. Malignant tumours are cancerous growths that have the potential to expand and spread to other areas of the patient's body. A benign tumour is one that can develop in a specific area of the body but does not spread to other body areas. Several psychological, physical, and practical obstacles might be brought on by breast cancer. Thus, early breast cancer prediction becomes crucial. According to research, the risk of breast cancer has been linked to hormonal, behavioural and environmental factors. Following behavioural decisions and related measures have been shown to lower the risk of breast cancer [1]:

- Long-term breastfeeding
- Regular exercise and weight management
- Abstinence from dangerous alcohol use and protection from tobacco smoke
- Abstinence from protracted hormone usage and protection from excessive radiation exposure.

Here are a few pictures showing the distinction between a healthy breast and a breast with cancer [2]:

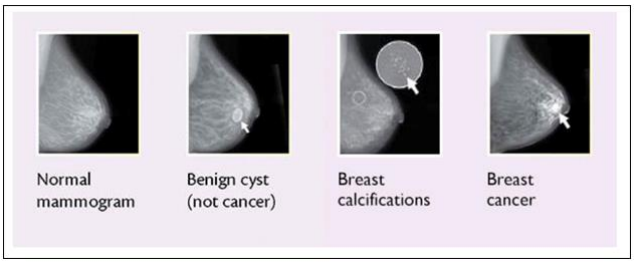

Fig.1distinction between a healthy breast and a breast with cancer

A tumour is an unnatural growth or mass of cells. It is benign when the tumour's cells are healthy. The tumour is considered malignant if the cells are aberrant, capable of uncontrollable growth, and generate a lump. The difference between Benign tumour and malignant tumour is depicted in following figure: [3]

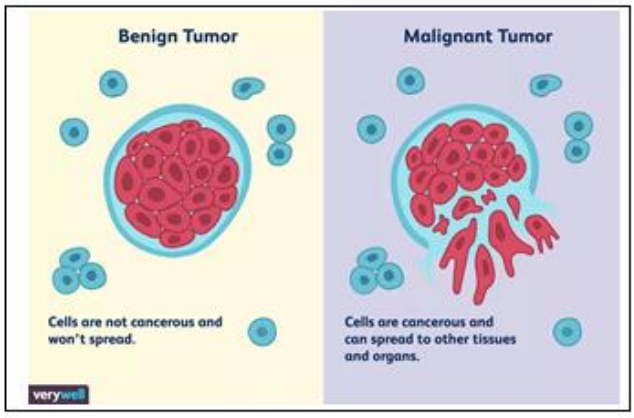

Fig. 2 Difference between Benign tumour and malignant tumour

II. PROPOSED METHODOLOGY

Breast cancer categorization categories malignant information based on how much or how little it has spread. Classification algorithms support the opposing attributes in the dataset by predicting one or more discrete variables. Breast cancer will be predicted using different machine learning algorithms. The behaviour among several algorithms needs to be compared in order to perform a comparative analysis. Methodology for "Brest Prediction using Supervised Machine Learning Algorithms" is executed as shown in Fig. 3:

Fig. 3 Supervised Machine Algorithms execution flowchart

A. Data Collection and data pre-processing

Download the dataset from the Wisconsin Diagnostic Breast Malignancy (WDBC) dataset. There are 569 records altogether, 357 of which are benign (noncancerous), and 212 of which are cancerous (Malignant). The Fig. 4 compares the affected (Malignant) and normal (Benign) cells in our data set. In some cases, the dataset may contain blank or missing values which needs to be handled and equipped and with an appropriate value.

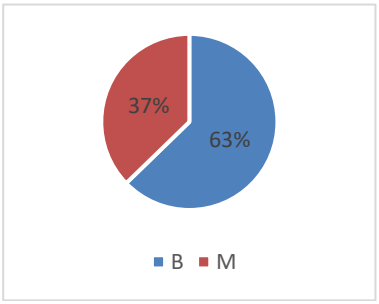

Fig. 4 Affected (Malignant) and normal (Benign) cells ration in data set

B. Algorithm selection

To build the model, labelled dataset is by selecting a suitable classification algorithm.

*1) Naive Bayes:*For classification tasks like text classification, the supervised machine learning method called Naive Bayes classifier is used. This algorithm models the input distribution of a certain class or category, so can belong to the family of generative learning algorithms. Naive Bayes algorithm depends on the principle of Bayes' Theorem.

2) Random Forest Algorithm: It is popular algorithm for classifying and predicting data which is a supervised machine learning technique. It constructs decision treesand uses their average for classification and majority vote for regression.

3) Logistic regression: It is supervised machine learning technique used for predicting the categorical dependent variable using a given set of independent variables. Output of this algorithm is probabilistic values lying between 0 and 1. Logistic regression is used for solving classification type problems.

*4) K-Nearest Neighbor (KNN):*The k-nearest neighbours algorithm, referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another.

*5) Decision Tree:*A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. Whereas Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches. The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions.

C. Training And Test Data

Our model will be trained to determine whether the cancer is benign or malignant by splitting the pre-processed data into training and testing sets in a ratio of 80:20, respectively.

III. Tool Used For Experiment

The dataset was analysed and predictions were made using WEKA, a popular tool. There are several algorithms used in the supervised learning area. The algorithms used in this research include Naive Bayes, Decision Tree, KNN, Logistic regression and random forest algorithm.

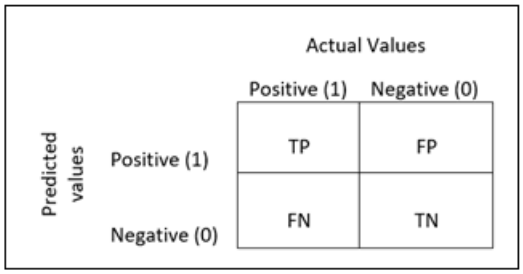

Fig. 5 Confusion Matrix to Be Used To CaptureResults[5]

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm. This matrix consists of True positive (TP): Observation is predicted positive and is positive. False positive (FP): Observation is predicted as positive and negative. True negative (TN): Observation is predicted negative and is negative. False negative (FN): Observation is predicted negative and is actually positive. Precision quantifies the number of positive class predictions that belong to the positive class. Recall quantifies the number of positive class predictions made from all positive examples in the dataset. F-Measure provides a single score that balances both the concerns of precision and recall in one number[5].

IV. Result Analysis

After executing the five algorithms, the results obtained are shown in confusion matrix as follows:

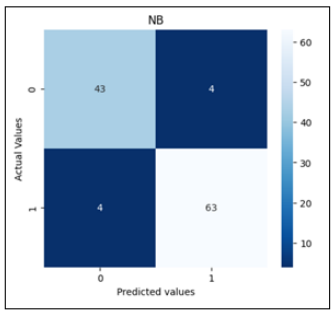

Fig. 6a Confusion Matrix using Naive Bayes

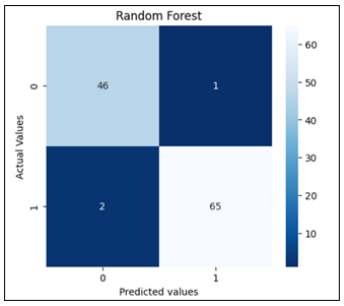

Fig. 6b Confusion Matrix using Random Forest

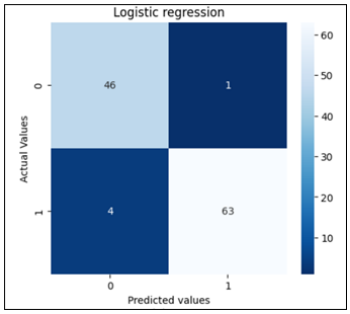

Fig. 6c Confusion Matrix using Logistic Regression

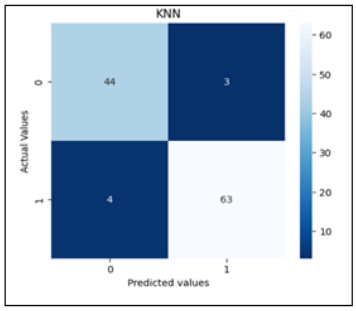

Fig. 6d Confusion Matrix using KNN

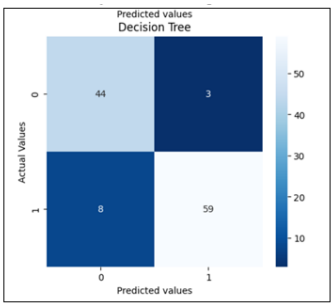

Fig. 6e Confusion Matrix using Decision Tree

The dataset was analysed for the aforesaid algorithms. For every algorithm, accuracy is the most important parameter which specifies how correctly the algorithm has classified the instances of the dataset. Apart from accuracy, Precision, Recall and F-score are also considered for the comparison of all five algorithms.

TABLE I COMPARISON AMONG ALGORITHMS

Classification Model	**TP**	**TN**	**FP**	**FN**
Random forest	46	65	1	2

Logistic Regression	46	63	1	4
Support Vector Machines	40	66	7	1
K Nearest Neighbour	44	63	3	4
Decision tree	44	59	3	8

V. CONCLUSION

The accuracy measures of five well-known classification models are analysed in this study based on their qualitative performance on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The best overall performer was found to be the random forest classifier. The final results are shown in Table 2, where it can be concluded that while the performance of the four chosen machine learning algorithms is closely competitive, Randomforest classifier is experimentally observed to be the best with accuracy of 97%. The four selected machine learning algorithms are implemented on the same data as input for a number of times, and the evaluated matrices are averaged in the final results.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/breast- cancer
[2] https://www.cancer.gov/types/breast/breast-changes
[3] https://www.verywellhealth.com/what-does-malignant-and-benign-mean 514240
[4] Annoy Chowdhury, Breast Cancer Detection and Prediction using Machine Learning, IJERTV9IS020280 ,19 June 2020.DOI: 10.13140/RG.2.2.23969.84320.
[5] www.sciencedirect.com