Available online @ https://jjem.jnnce.ac.in
https: www.doi.org/10.37314/JJEM.SP0244
Indexed in International Scientific Indiexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 08 December 2024

Machine Learning Technique to Identify the Breast Cancer

Bindu M G[1]*, Dr. Raghavendra S P[2]
[1*]Student Department of MCA, [2]Assistant Professor Department of MCA
JNN College of Engineering, Shivamogga
bindumg914@gmail.com, raghusp@jnnce.ac.in

Abstract

Breast cancer is still among the leading diseases affecting women and there is a high mortality rate for the disease. In this work, the following model of Logistic regression was implemented for diagnosing benign vs malignant tumors of breast cancer using Wisconsin diagnostic dataset. The variables used this dataset's associated with tumor characteristics, such as tumor's perimeter, smoothness, texture and radius. Hypotheses are examined with the assistance of statistical measures and features are analyzed and depicted in detail to identify significant relationships. As a measure of testing the model, the dataset is divided into the training and the testing dataset. The proposed model (LR) is trained utilizing the training set, and its output is assessed by means of such measures as accuracy, precision and etc. The significance of proper classification of breast cancer is demonstrated in this research while establishing the practicality of the logistic regression technique in achieving it. The outcome exhibit that the proposed system that is logistic Regression can beneficial and reliable diagnostic tool in medical science and beneficial for the improvement of the health status of the individuals suffering from breast tumor additionally for the management of breast tumor.

Keywords: *Breast cancer, logistic regression, benign, malignant, Wisconsin diagnostic dataset, testing, training*.

1. Introduction

Breast tumor is the more extensively spread and danger health concern which affects a considerable number of females worldwide. It is very difficult to detect the breast tumor at the early stage for the treatment and curing of the tumor. So here we use the logistic regression model which classifies the tumor as either cancerous or non-cancerous, we can use the several kinds of machine learning algorithms to predict this tumor, in that the logistic regression plays a vital role and give more accuracy. The logistic regression is a common method of binary classification that aims at finding a probability that an instance belongs to a given class known as logistic regression. Applying input parameters such as size and shape of the tumor, its texture and color, and age of the patient, and others, logistic regression may be allocated as an excellent method for categorizing breast tumors, in particular, benign and malignant ones. It's essential for classify breast tumors accurately since it enables doctors to cultivate right decisions about what to do for handle the patients, order for further tests or just manage them. If malignant tumors are accurately diagnosed, physicians may begin early measures such as radiation, chemotherapy, surgery, and others to prevent the disease's progression and enhance patients' quality of life. When benign tumors are accurately diagnosed however, simpler procedures can be undertaken, thus making the patients feel more comfortable. So, within this study, logistic regression was employed to classify the cases of breast cancer. This section also describes how a proposed methodology was developed starting with the data set used, training, and testing the model, and the limitations of this approach. Moreover, as mentioned in the paper, logistic regression is presented here as the great instrument that can help in achieving the aim of accurate categorization of breast cancer at the initial stage of diagnosis and treatment.

2. Literature survey

Given that the proposed system few of the related works are here i.e., Ch. Srividya et.al (2023) [4] this research demonstrates that utilizing the logistic regression can correctly spot breast cancer 92.98% of the time. This method is good because it avoids too many unwanted medical steps. It helps in finding the illness early and planning how to treat it.

S. Sathyavathi et.al (2020) [13] This research was completed on using Logistic Regression to predict The Wisconsin Breast Cancer dataset contains instances of breast cancer. with a 93.63% precision. A high early-stage detection value which is being highlighted by the paper as well as medical imaging progress and computer-aided design machine learning applications is the core of the paper.

Anoy Chowdhury (2020) [1] Machine learning is used in this study to sort the breast tumors as either cancerous or non-cancerous using Wisconsin Diagnostic dataset. It decreases unnecessary treatments that further enhance patient outcomes since accurate diagnostic procedures are conducted. Thus, the article concentrates on the possibilities of enhancing the machine learning datasets and algorithms to arrive at a higher diagnostic accuracy and better treatment planning.

Raghavendra R et.al (2021) [9] this paper discusses the value of early breast tumor screening. It evaluates machine learning methods and comes up with the probability that the tumors are the most probable to be cancerous or non-cancerous using Logistic Regression with 94% accuracy. Some of the results of this study indicate that employing machine learning enhances the diagnosis accuracy compared to conventional methods.

Sweta Bhise et.al (2021) [15] This paper aims at exploring the utilization of Machine learning in breast cancer detection, specially on the survival aspect for the diseases. Namely, it examines how the features selection methods of five classifiers – namely, SVM, Random Forest, KNN, Logistic Regression and Naïve Bayes are performed in both the CNN and RFE classifiers. From the BreaKHis 400X dataset experiments, it could be clearly noted that here the CNN surpasses other techniques regarding accuracy and precision.

Priyanka et.al (2021) [8] this study examine that the breast tumor is a familiar disease for so many people and, especially in females and reduction of mortality rates is contingent on the identification of signs of cancer. In specific, deep learning systems are gaining supremacy over several conventional classifiers on large scales of data and on images. Here CNN is used.

Shirin Raut et.al (2023) [11] This paper aims towards outlining the requirements for robots to accurately detect and predict the cases of breast tumor using the Wisconsin dataset. It measures its efficiency accordance with the percentage of absolute sickness to the sum and the ratio when the sickness is none. In more detail, Identities were identified with high accuracy.

Hua Chen et.al (2023) [6] This study provides a summary of the breast tumor recognition here to handle data imbalance, The data provided is splitted, then normalization is allocated to preprocessing the data, the attributes are chosen according to Pearson correlation and Logistic regression, Random Forest, XGBoost, and K-Nearest Neighbors are evaluated. XGBoost attained the greatest accuracy (0.94). It highlights the importance of memory in diagnostic practice and the demand for improved prediction algorithms.

Juhi Seth et.al (2020) [7] This research discusses how often people get cancer, and more focusing on breast tumor. It uses a math method called logistic regression to guess cancer stages based on what disease experts see in genes and get the accuracy of 95%. The study points out how key it is to be able to spot cancer early because most cancer is caused by gene issues or the way people live.

Apoorva V et.al (2021) [2] This study states that breast tumors are still the largest killer with tremendous numbers of death throughout the world. It needs early identification. So, his research uses CNN algorithm and numerate

method like K-Nearest Neighbors (KNN), Decision Tree (CART), Support Vector Machine (SVM) and Naive Bayes. In predicting image-based research founds that CNN are better, while in predicting numerical data SVM give a better result.

S K Ahmed Mohiddin et.al (2022) [12] This research focus on the study has cumulatively examined the six supervised algorithms such as SVM, KNN, Random Forests, Decision Trees, Naive Bayes and Logistic Regression with the objective of using machine intelligence (ML) to aid in the recognition of breast tumor. Random Forest on validation using K-Fold obtaining accuracy over 90% depicts its usefulness in analyzing medical data.

3. Methodology

The design and methodology for the purpose of identifying the breast tumor utilizing the logistic regression involves different steps:

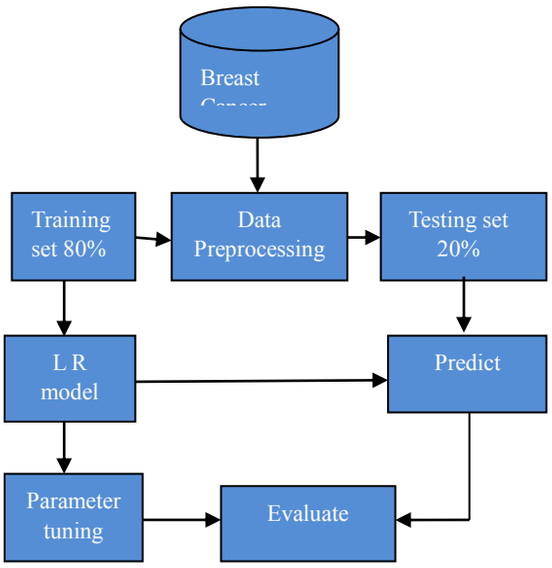

Figure 1: Block diagram of logistic regression model

as it is mentioned in fig1. including data loading, preprocessing, model training, evaluation, and prediction. The following steps are described in detail below:

3.1. Data loading and preprocessing:

The initial step in the suggested approach is to import the several libraries that will be needed. Then, load the Wisconsin Diagnostic dataset using scikit-learn.

So, it is represented as,

$$X \in R^{n*d}$$

Here, X represents the matrix contains n different samples and d different features.

$$y\epsilon\{0,1\}^n$$

Here, y represents the vector with length n and 0 symbolizes the benign tumor and 1 represents the malignant tumor. The problem type is a binary-classification. The data is loaded up and then transformed into pandas DataFrame so, it might be readily managed and analyzed on surface level through data insight aiming towards complete understanding in depth.

3.2. Data exploration and visualization:

Exploratory data analysis is performed to learn more about a dataset. Inspecting of DataFrame's shape helps to find out the count of rows and columns. The mathematical aspect of exploratory data analysis for the WDBCDS encapsulated in the below manner:

$$shape(D) = (a, b)$$

Where, shape inside the dataset is D together with a rows and b columns.

In the visualization, as mentioned in fig 2, the pair plots are employed to figure out the correlation between cancerous and normal tumors using the two different colors that is described as follows:

$$Colors = \begin{cases} color_1 \; if \; y = 0 \; (benign) \ldots\ldots (1) \\ color_2 \; if \; y = 1 \; (malignant) \end{cases}$$

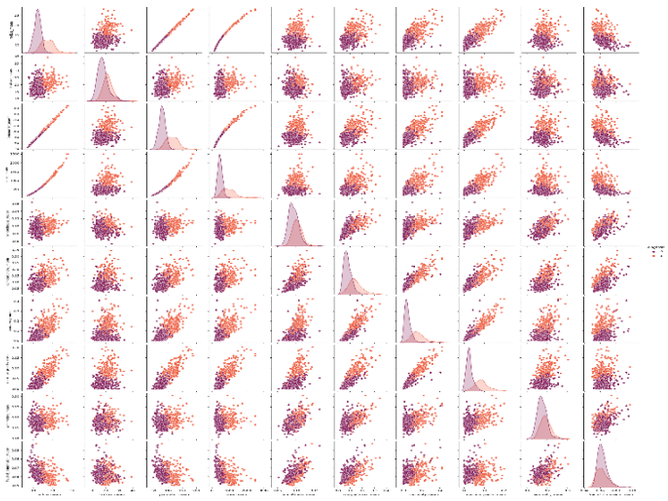

Figure 2: Pair Plot

3.3. Data splitting

The collection of data is divided into training and testing sets using the Scikit-learn train-test split () method.

$D_{train}, D_{test} = train\_test\_split(D, test\_size = 0.2)$

Here, it splits the dataset into 80% training and 20% testing data.

3.4. Logistic regression model training

The LogisticRegression() class from scikit-learn is accustomed to instantiate a logistic regression model. This model is then trained on the training aggregation of data with the fit () function so that it can estimate the ideal parameters that increase the likelihood of observed data.

$$h_0(x) = \frac{1}{1+e^{-0Tx}} \ldots\ldots (2)$$

3.5. Model evaluation and prediction

The trained model using logistic regression may be evaluated using few performance metrics. The accuracy score() function checks the truthful ness of both the test and training sets of information by comparing expected labels with actual labels. On Accuracy metric, one often sees what percentage of samples was correctly classified.

Forecast derived from the threshold is:

$$\hat{y}_i = \begin{cases} 1 \; if \; h_0(x_i) \geq 0.5 \\ 0 \; if \; h_0(x_i) < 0.5 \end{cases} \ldots\ldots (3)$$

This would classify the input as 1 (malignant) if the predicted probability is ≥0.5, otherwise 0 (benign).

The performance matrix, accuracy is as shown:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i = y_i) \ldots\ldots (4)$$

This is proportion of correctly classified instances.

$$Precision = \frac{TP}{TP+FP} \ldots\ldots (5)$$

$$Recall = \frac{TP}{TP+FN} \ldots\ldots (6)$$

Where, TP is true positive, TN is true negatives, FP is false positive, FN is false negative.

The general methodology and approach of the project consist of the outlined steps: data loading, preprocessing, exploratory analysis, model training, evaluation, and prediction. Executing these procedures will lead to development of a approach to the detection of breast tumour using logistic regression modeling that will issue the accurate predictions based on input data and help in making well-informed therapeutic decisions.

4. Result and Discussion

The machine learning for the diagnosis of breast tumor is examine with the help many different classifiers, including Random Forest, KNN, Decision Tree, and Logistic Regression, using Wisconsin dataset (569) samples and which has 31 different features. As mentioned in the fig 3, here every classifier yields the different accuracy. Among these the proposed methodology i.e. logistic regression will yield 96% accuracy i.e. out of 569 samples 547 will used to get the 96% accuracy, which is greater than all the other classifiers. So, as the proposed methodology is a problem of binary classification that is benign or malignant the logistic regression will match well with this problem. Hence the classification of the tumor easily done using logistic regression in the proposed methodology.

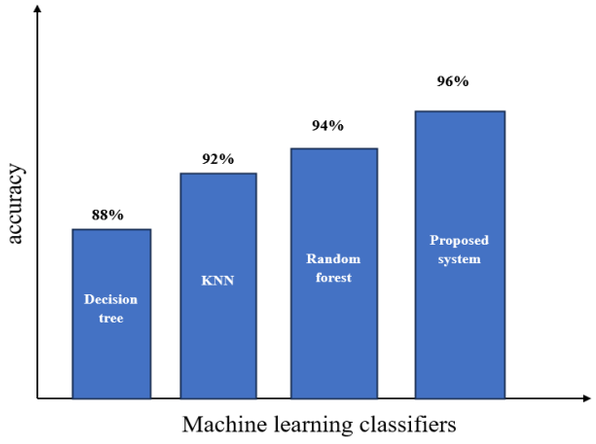

Figure 3: Comparative Analysis

Success rate for Logistic Regression = 96%

Conclusion

Logistic regression breast cancer classification exactly pinpoints whether a breast tumor is

likely to be cancerous or non-cancerous for this project. In light of this implementation using Wisconsin Diagnostic dataset, it proved that the logistic regression has the ability in handling the data builds efficiently. These stages included the exploratory data analysis and data visualization, the evolution of the construction model, the verification model, and lastly the model prediction. Hence, the high accuracy of the logistic regression model showed the importance of early recognition of the breast tumor for proper planning and management. It is less computationally intensive and model interpretability when performing a LR here, and generally it would greatly benefit the healthcare practitioners in their decision-making along with patient management.

References

1. Anoy Chowdhury. Breast Cancer Detection and Prevention using Machine Learning. University of Engineering & Management, Kolkata, India (2020).
2. Apoorva V. et.al. Breast Cancer Prediction using Machine Learning Techniques, Atlantis Highlights in Computer Science volume 4 Proceedings of the 3rd International Conference on Integrated Intelligent Computing Communication & Security (ICIIC 2021).
3. Arslan Khalid. et.al. Breast Cancer Detection and Prevention using Machine Learning, Diagnostics MDPI, 2023.
4. Ch. Srividya et.al. Breast Cancer Detection using Logistic Regression, Computer Science and Engineering, Institute of Aeronautical Engineering, Eur. Chem. Bull, 2023.
5. Chaurasia V et al. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 2018.
6. Hua Chen. et.al. Research Article Classification Prediction of Breast Cancer Based on Machine Learning, Hindawi Computational Intelligence and Neuroscience, 11 January 2023.
7. Juhi Seth. et.al. Detection of Breast Cancer by Logistic Regression using Machine Learning, International Journal of Innovative Science and Research Technology, 5 May 2020.
8. Priyanka. et.al. A Review Paper on Breast Cancer Detection using Deep Learning, IOP Conf. Series: Material Science and Engineering, 2020.
9. Raghavendra R. et.al. Breast Cancer Detection and Prediction using Machine Learning, Internation Journal of Scientific Development and Research (IJSDR) 2021.
10. R. Shastri. et.al. Breast Cancer Detection using Machine Learning Algorithms, International Journal of Research in Industrial Engineering, 2020.
11. Shirin Raut. et.al. Breast Cancer Detection using Machine Learning, International Journal of Scientific Research in Engineering and Management (IJSREM), May-2023.
12. Sk. Ahmed Mohiddin. et.al. Breast Cancer Prediction using Machine Learning, Dogo Rangsang Research Journal, 2022.
13. S. Sathyavathi. et.al. Breast Cancer Identification using Logistic Regression, Bioscience Biotechnology Research Communications (BBRC), 2020.
14. Sunday Samuel Olofintuyi, Breast Cancer Detection with Machine Learning Approach, FUDMA Journal of Science (FJS), 2, April 2023.
15. Sweta Bhise et.al. Breast Cancer Detection using Machine Learning Techniques, International Journal of Engineering Research & Technology (IJERT), July-2021.