iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44676

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Detection of Breast Cancer Using Machine Learning Algorithms

Prithviraj Jain[1], Sanjana T S[2], Ranjitha P R[3], Vinuthana Chatra[4], Rachana Varma[5]
[1]*Assistant Professor,* [2, 3, 4, 5] *Student, Department of Computer Science and Engineering, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire, Karnataka, India*

Abstract: ***Breast Cancer is one of the most frequently occurring cancers in women. Cancer rates are increasing in almost every region around the world. Early detection of cancer is the most efficient way to prevent the deaths caused due to it. Currently, the most commonly used tests for the detection of cancer are Mammograms, Breast Ultrasound and Breast MRI. These techniques have their own disadvantages which include the risk of false detection or no guarantee that all cancers will be detected. It is crucial to use alternative methods that are easier to implement and produce more reliable results. The solution is to use various Machine Learning algorithms to overcome the disadvantages of traditional techniques. This paper aims to propose a prediction model using Machine Learning classifier algorithms like Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The performance of the various classifiers is compared in terms of accuracy, precision, and recall and the best classifier for the detection of cancer.***

Keywords: Machine learning, Classification, Accuracy, Breast Cancer, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbors

I. INTRODUCTION

The most commonly occurring type of cancer in women is breast cancer. It affects over two million women annually. According to the National Cancer Registry Programme [15], Breast cancer is the most common cancer in women in India and accounts for 14% of all cancers in women. There are no prevention techniques for breast cancer hence early detection and diagnosis are crucial in determining the chances of survival. Early detection of cancer is the most efficient way to prevent the deaths caused due to it. Currently, the most commonly used tests for the detection of cancer are Mammograms, Breast Ultrasound, and Breast MRI [13]. These techniques have their own disadvantages which include the risk of false detection or no guarantee that all cancers will be detected. It is crucial to use alternative methods that are easier to implement and produce more reliable results.

Machine learning (ML), an important branch of Artificial Intelligence, is a field concerned with algorithms that automatically improve through experience by the use of data. The field focuses on prediction, based on known properties learned from the training data. ML approaches are divided into three broad categories: Unsupervised Learning, Supervised Learning, and Reinforcement Learning Some of the fields that make use of ML are healthcare, agriculture, astronomy, finance, bioinformatics, and so on. Classification algorithms are a type of supervised learning technique. This algorithm learns from the observations or available data and classifies the new data or observations into categories or groups.

II. LITERATURE REVIEW

M. Tahmooresi, A et. Al. [1] examined the role of machine learning techniques in breast cancer detection. The authors proposed a hybrid model that combined several ML algorithms including Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), and Decision Tree (DT) for effective breast cancer detection. The findings of these researchers suggest that SVM is the most popular method and gives an accuracy of 99.8% but the drawback of this method is that they use the Image Testing dataset from Mammogram instead of a numerical dataset. This means that the prediction can be done only if the patient is having symptoms of cancer. Shubham Sharma, Archit Aggarwal, and Tanupriya Choudhury [2] mainly concentrated on the comparison of the widely popular machine learning algorithms, and also aim to be commonly used for breast cancer detection, namely Random Forest, Naïve Bayes, and KNN. It has been observed that among these algorithms KNN is the most effective in the detection of breast cancer because it has 95.9% accuracy and 90.47% Recall. Lal Hussain, Wajid Aziz, Sharjil Saeed et. Al. [3] explained how to distinguish cancer mammography from normal mammograms using strong machine learning classification approaches including Support Vector Machine (SVM) with linear, polynomial, Radial Base Function (RBF), Bayesian approach, Gaussian kernels, and Decision Tree.

©IJRASET: All Rights are Reserved

With the accuracy acquired using morphological, texture (TA=93.77%), and entropy characteristics, Bayes, SVM polynomial, and SVM Gaussian have the best performance in terms of specificity, sensitivity, PPV, NPV, TA, and AUC. One of the studies [4] evaluated a computer-aided diagnostic system with texture analysis to improve radiologists' accuracy in the identification of breast tumors as malignant or benign. Sri Hari Nallamala, Pragnyaban Mishra et. Al [5] studied in what way ensemble voting ML technique can be used for detecting breast cancer that considers only 16 features for analysis. Variance, range, and compactness were among the features extracted by Sailesh GC et. Al. [6]. To assess the performance, they employed SVM classification. Their findings revealed the highest variation (95%), range (94%), and compactness (86%) of any group. SVM might be deemed a suitable method for breast cancer detection based on their findings. Wenchao Xing and Yilin Bei [7] conducted an experimental study on the class-based weighting factors of the KNN classifier and analyzed real-time data sets and compared the performance. An improved KNN algorithm based on DBSCAN cluster denoising and density cropping was proposed. They used the cluster method to speed up the search speed of KNN and classified the efficiency of KNN and maintained accuracy. Shufen Ruan, Hongwei Li, Chaoqun Li, and Kunfang Song's [8] paper proposed a new class-specific deep feature weighting method for MNB text classifiers, which assigns each feature a specific weight for each class and also estimates the conditional probabilities of the text classifier by deeply computing feature weighted frequencies from training data. The authors employed a Friedman test, post-hoc Holm's test, and Wilcoxon test to analyze a set of algorithms, then compared the result with the proposed class-specific deep feature weighting approach. A special property of SVM is, that SVM simultaneously minimizes the empirical classification error and maximizes the geometric margin [9]. So SVM is called Maximum Margin Classifiers. SVM generally is capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. The LR model in the multivariable method for modeling the connection between several independent variables and a categorical dependent variable was examined and reviewed by Ernest Yeboah Boateng, and Daniel A. Abaye in a work focused on medical research [10]. They gave an excellent demonstration of how to use the LR model with data from a cohort of pregnant women and the factors that influence their decision to have a cesarean delivery or normal birth.

III. PROBLEM STATEMENT

In 2020, there were 2.3 million women diagnosed with breast cancer and 685000 deaths globally according to the World Health Organization (WHO) [11]. According to the National Cancer Registry Programme, Breast cancer is the most common cancer in women in India and accounts for 14% of all cancers in women. Breast cancer is detected after the symptoms appear and the most common symptom is a new lump or mass [12].

But, in many cases, there are no symptoms and there is a chance that cancer may not be detected in the initial stage. The imaging techniques currently being used like Mammograms or Ultrasound do not guarantee early detection of cancer. The better and more efficient solution for this problem is using techniques of Data Science and Machine Learning. The goal is to create a model to classify malignant and benign tumors.

IV. METHODOLOGY

The proposed solution is using ML algorithms on the dataset for the early detection of Breast Cancer. The next step is to load the dataset and explore the data. The dataset is available in UCI ML Breast Cancer Wisconsin (Diagnostic) datasets. The data is visualized using pair plots, heatmaps, and correlation bar plots. The next step is data processing where we need to clean the data. The cleaned dataset needs to be trained and tested with the 4 ML algorithms so that the best model for cancer detection in terms of accuracy can be found.

A. K-Nearest Neighbor (KNN) Algorithm

KNN is based on the Supervised Learning technique. The KNN algorithm at the training phase stores the dataset and when it receives new observation and it then classifies that data into a category that is much similar to the new data.

B. Naïve Bayes (NB) Classifier Algorithm

Naïve Bayes algorithm, a supervised learning technique is based on the Bayes theorem and is used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. This algorithm predicts on the basis of the probability of an object. NB is fast and performs well even with a small dataset.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. *Logistic Regression (LR) Algorithm*

Logistic regression is one of the most popular ML algorithms. It predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. The LR hypothesis tends to keep the cost function between 0 and 1.

D. *Support Vector Machine (SVM) Algorithm*

SVM is a popular algorithm under the Supervised Learning technique, which is used for Classification as well as Regression problems. The SVM algorithm's purpose is to create the optimum line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. The hyperplane is the name for this optimal choice boundary.

V. PROPOSED ARCHITECTURE

The proposed architecture of the system is shown in Figure 1. Supervised classification algorithms are used to train and test the data. The first step is to collect, clean, and pre-process the data set. The data set is then trained using the ML classification algorithms KNN, SVM, NB, and LR. The output variable is either benign or malignant. The accuracy is calculated for each model. The performance metrics to be measured are accuracy, precision, recall, and F1 score. The most accurate algorithm will be used to implement the web application.

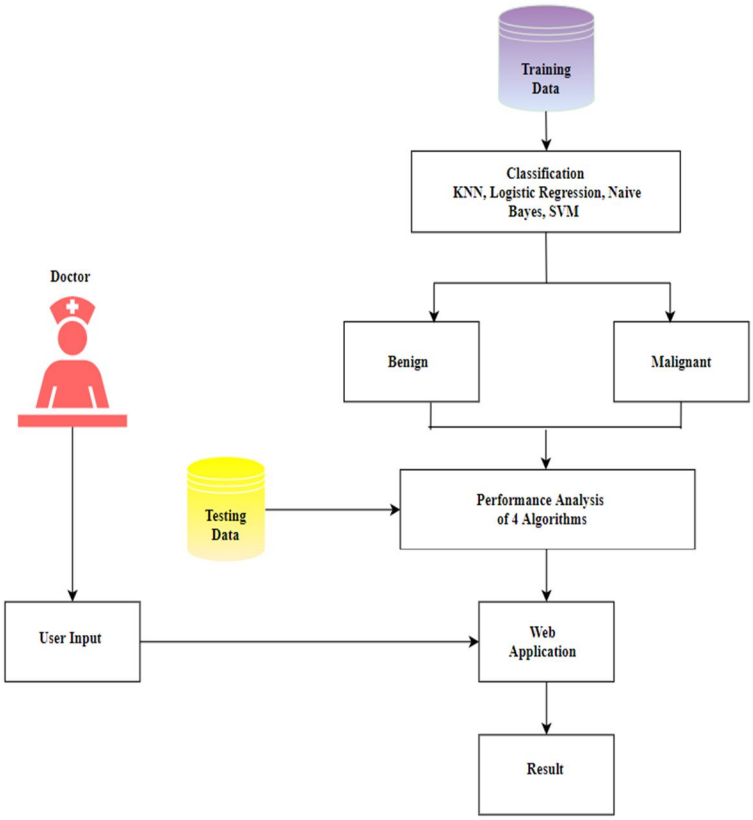


Figure 1: Proposed architecture diagram

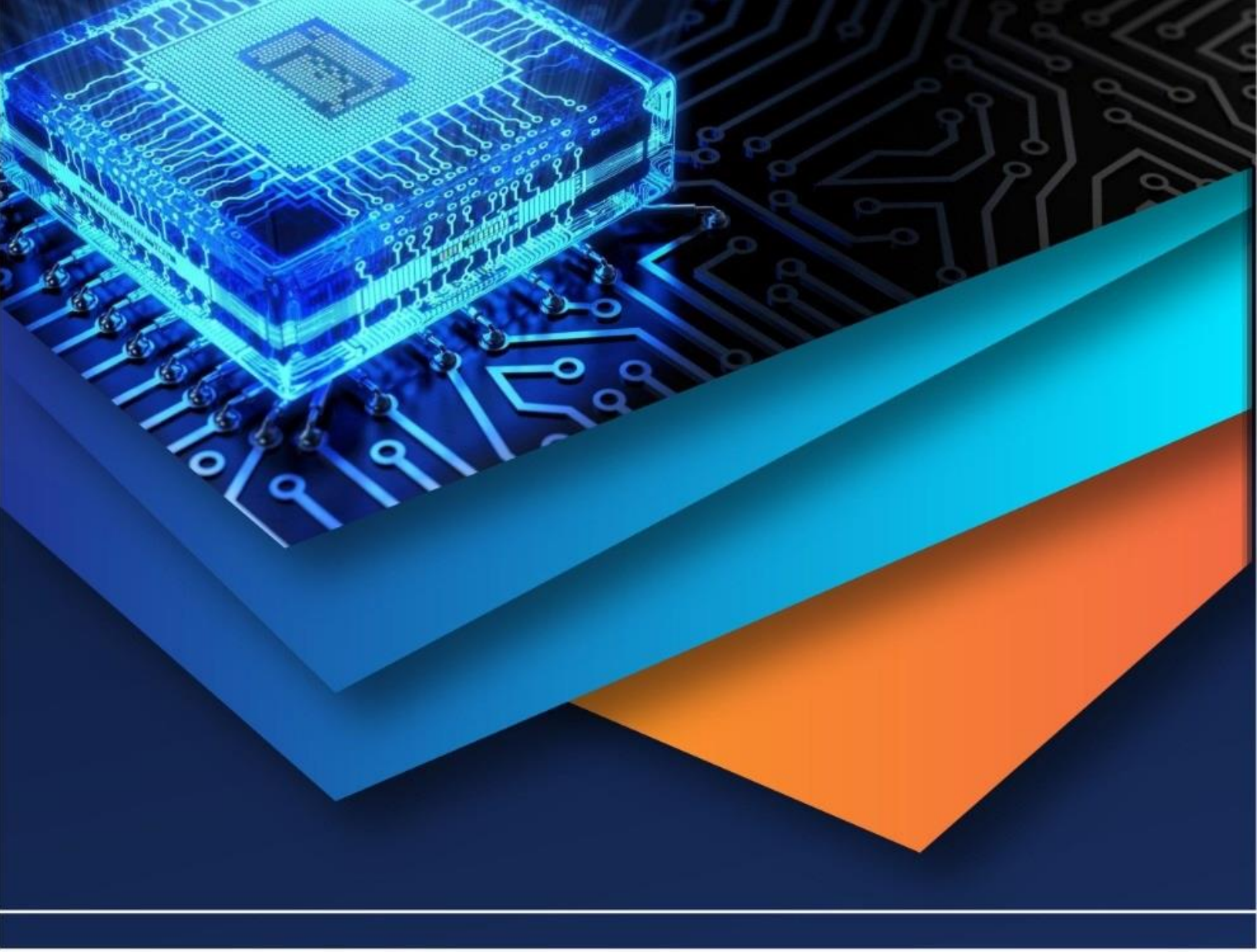
VI. CONCLUSION AND FUTURE SCOPE

Long-term mortality rates from breast cancer could be lowered by early cancer identification. In this paper, we have proposed to conduct a comparative analysis of four machine learning algorithms. Supervised machine learning algorithms will be highly helpful in cancer research. These algorithms are an efficient method for the early detection of cancer. In this study, the four algorithms that were identified were KNN, SVM, NB, and LR.

In the future using available analysis tools, the best algorithms must be identified. Based on the accuracy and precision the best algorithm will be chosen. Further, a web application will be implemented for the prediction.

REFERENCES

[1] M. Tahmooresi, A. Afshar, B. Bashari Rad, K. B. Nowshath, and M. A. Bamiah (2018). Early Detection of Breast Cancer Using Machine Learning Techniques, Journal of Telecommunication, Electronic and Computer Engineering,10(3-2), p. 21-27. 10.1109/TrustCom/BigDataSE.2018.00057
[2] Shubham Sharma, Archit Aggarwal, Tanupriya Choudhury (2018). Breast Cancer Detection Using Machine Learning Algorithms, International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), p. 114-118. 10.1109/CTEMS.2018.8769187
[3] Lal Hussain,Wajid Aziz ,Sharjil Saeed,Saima Rathore, Muhammad Rafigue (2018). Automated Breast Cancer Detection using Machine Learning Techniques by Extracting Different Feature Extracting Strategies, 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, p. 327-331.
[4] Ali Abbasian Ardakani, Akbar Gharbali (2015). Afshin Mohammadi. Classification of Breast Tumors Using Sonographic Texture Analysis, Journal of Ultrasound in Medicine,34(2), p. 225-23. 10.7863/ultra.34.2.225
[5] Sri Hari Nallamala, Pragnyaban Mishra, Suvarna Vani Koneru (2019). Breast Cancer Detection using Machine Learning Way, International Journal of Recent Technology and Engineering (IJRTE),8(2S3), p. 1402-1405. 10.35940/ijrte.B1260.0782S319
[6] Sailesh GC, Ravi Kasaudhan, Tae K. Heo, Hyung D. Choi (2015). Variability Measurement for Breast Cancer Classification of Mammographic Masses, Conference on research in adaptive and convergent systems, p. 177-182. https://doi.org/10.1145/2811411.2811505
[7] Wenchao Xing and Yilin Bei (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm, IEEE Access, 8, p. 28808-28819.10.1109/ACCESS.2019.2955754
[8] Shufen Ruan, Hongwei Li, Chaoqun Li, and Kunfang Song (2020). Class-Specific Deep Feature Weighting for Naïve Bayes Text Classifiers, IEEE Access, 8, p. 20151-20159.10.1109/ACCESS.2020.2968984
[9] Durgesh K. Srivastava and Lekha Bhambhu (2010). Data Classification using Support Vector Machine, Journal of Theoretical and Applied Information Technology, 2(1). http://www.jatit.org/volumes/research-papers/Vol12No1/1Vol12No1.pdf
[10] Ernest Yeboah Boateng, and Daniel A. Abaye (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research, Journal of Data Analysis and Information Processing, 7, p. 190-207.10.4236/jdaip.2019.74012
[11] World Health Organization (2020). Available: https://www.who.int/publications/m/item/cancer-ind-2020
[12] Center for Disease Control and Prevention. Available: https://www.cdc.gov/cancer/breast/basic_info/diagnosis.html
[13] American Cancer Society. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html
[14] National Cancer Registry Programme (2020). Available: https://www.ncdirindia.org/All_Reports/Report_2020/default.aspxs



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)