International Journal of Reliable and Quality E-Healthcare
Volume 12 • Issue 1

Classifying Malignant and Benign Tumors of Breast Cancer:

A Comparative Investigation Using Machine Learning Techniques

Meshwa Rameshbhai Savalia, Institute of Technology, Nirma University, India
Jaiprakash Vinodkumar Verma, Institute of Technology, Nirma University, India*
https://orcid.org/0000-0001-6116-1383

ABSTRACT

Breast cancer is the second major cause of cancer deaths in women. Machine learning classification techniques can be used to increase the precision of diagnosis and bring it closer to 100%, thus saving the lives of many people. This paper proposed four different models, built using different combinations of selected features and applying five ML classification techniques to all the models to identify the best model with the highest accuracy. It analyzes five machine learning techniques, namely logistic regression (LR), support vector machines (SVM), naive bayes (NB), decision trees (DT), and k-nearest neighbor (KNN), for prediction of breast cancer using the Wisconsin Diagnostic Breast Cancer Dataset on these four models. The objective of the paper is to find the best ML algorithm that can most accurately predict breast cancer for a particular model. The outcome of this paper helps the doctors to improvise the diagnosis by knowing the effect of combinations of symptoms with the growth of breast cancer.

KEYWORDS

Breast Cancer Diagnosis, Classification, Decision Trees, K-Nearest Neighbor, Logistic Regression, Naive Bayes, Support Vector Machines, Wisconsin Diagnostic Breast Cancer Dataset

INTRODUCTION

Breast cancer is one of the major causes of death around the world. One in every ten women is affected by breast cancer (Ilbawi & Velazquez-Berumen, 2018). It is essential to diagnose and predict dreadful tumors as early as possible to save a woman's life. We need to improve efficiency and simplify the testing and treatment processes. Hence medical records in the form of images as well as numerical data are required for this purpose which is already stored digitally in repositories. These repositories are publicly available

DOI: 10.4018/IJRQEH.318483 *Corresponding Author

for research to improve the diagnosis process. As per WHO, there were 9.6 million deaths due to cancer in 2018, making it the second-largest cause of death in the world (Ilbawi & Velazquez-Berumen, 2018). Globally, about 1 in 6 deaths is due to cancer. As per the American Cancer Society, 1,762,450 new cancer cases and 606,880 cancer deaths are estimated to occur in the United States in 2019 (Siegel et al., 2019). According to (Bray et al., 2018), the risk of dying from cancer before the age of 75 years is 7.34% in males and 6.28% in females. Breast cancer is one of the most chronic and dreadful diseases and one of the most common types of cancer found in women in the world. It accounts for 14% of all cancers in women. Overall, 1 in 28 women is likely to develop breast cancer during their lifetime. There were about 2.09 million cases of breast cancer in 2018. Chances of survival can be improved by early detection. Chances of survival can be increased by 98% if the cancer is diagnosed early (Ilbawi & Velazquez-Berumen, 2018). The average accuracy of manually diagnosing breast cancer by a human being from Fine Needle aspiration cytology (FNAC) is only 90%. This percentage can be optimized by applying machine learning techniques on digitized images of breast cells. It is important to correctly detect and diagnose the patients as early as possible. AI can be used for better and accurate detection and diagnosis of breast cancer.

Machine learning employs a variety of statistical, probabilistic, and optimization techniques. It allows the machine to "learn" from past examples and detect hard-to-discern patterns from large, noisy, or complex datasets (Cruz & Wishart, 2007). It can be used in medical applications, especially those that depend on complex proteomic and genomic measurements. Recently, researchers have been using machine learning for cancer diagnosis as well as prognosis. There is also a growing trend of personalized predictive medicine by using artificial intelligence. Plenty of research has been done which implants Machine Learning Techniques on the medical diagnosis of breast cancer using the Wisconsin Breast Cancer Diagnosis Dataset (WDBC). (Meraliyev et al., 2017) applied K nearest neighbor (KNN), SVM, ANN, Logistic regression, and decision tree (DT) model to predict breast cancer from the WDBC dataset. It uses K-fold cross-validation techniques to find evaluation measures for the model such as accuracy, sensitivity, specificity, etc. It claims that ANN, DTC, and logistic regression give 98% accuracy whereas KNN gives 99% accuracy and finally SVM can give 100% accuracy. (Kathija & Nisha, 2016) applied SVM and Naïve Bayes techniques for breast cancer data classification. This paper finds the smallest subset of features from the Wisconsin Diagnosis Breast cancer (WDBC) dataset by applying a 5-fold cross-validation method and confusion matrix accuracy so that it can ensure a highly accurate ensemble classification of breast cancer. This paper suggests that the naive Bayes model gives the highest accuracy of 95.65%. (Borges, 2015) presents a detailed description of the WDBC dataset. In addition, he applies the NB algorithm and JV8 algorithm for classification which has 97.80% and 96.05% accuracy respectively. Pre-processing is done using tools available in Weka 3.6. This paper proposed a comparative analysis of five machine learning techniques namely Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (DT), and K-Nearest Neighbor (KNN) for the prediction of breast cancer. We have used the Wisconsin Breast Cancer Diagnostic dataset (WDBC) (Dua & Graff, 2019) for the classification of benign and malignant tumors for breast cancer. This paper applies various machine learning classification techniques to the dataset to identify the best methodology for the classification task that gives the most accurate and reliable results.

The rest of the paper's organization is as follows: Section 2 shows a comparative study of the related literature on the different research done. Section 3 presents the proposed system for the proposed research work presented in this paper. Section 4 presents the methodology and concepts applied to achieve defined objectives. Section 5 and 6 presents performance analysis by describing the experiments and analysis of the experimental results. Section 7 concludes the paper and discusses future works.

RELATED WORK

This section describes the study of different Machine Learning (ML) approaches proposed or implemented by researchers in the area of cancer diagnosis. (Padma & Sowmiya, 2018) presents