Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 3 (2023)

Breast Cancer Prediction using Machine Learning

Prof. Krishna Mohana A. J.[1], Prof. Pramod Kumar P. M.[2], Prof. Mohan A. R.[3], Prof. Raghavendra T. K.[4], Prof. Shrinidhi A.[5]

[1,2,3,4,5]*Assistant Professor, Dept. of Computer Science and Engineering, Vivekananda College of Engineering and Technology, Puttur, Karnataka, India.*

***Abstract*:** The primary identification and prediction of type of the cancer about to develop a compulsion in cancer study, in order to assist and supervise the patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research Logistic Regression, K-NN, SVM, Random Forest, Decision Tree has been proposed to predict the breast cancer. To produce deep predictions in a new environment on the breast cancer data. Besides this, this study predicts the best Model yielding high performance by evaluating dataset on various classifiers. In this paper Breast cancer dataset is collected from the UCI machine learning repository has 569 instances with 31 attributes. Data set is pre-processed first and fed to various classifiers like Logistic Regression, K-NN, SVM, Random Forest, Decision Tree. The algorithm with the best results will be used as the backend to the website and the model will then classify the cancer as benign or malignant.

***Keywords**: Breast cancer classification, Breast cancer prediction, benign, malignant, KNN, Support Vector Machine, Random Forest, Decision tree.*

1. Introduction

Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. In recent years, the incidence rate keeps increasing and data show that the survival rate is 88% after five years from diagnosis and 80% after 10 years from diagnosis. Normally binary classification that is benign and malignant can be seen frequently. Today, one in 5 people worldwide will develop cancer during their lifetime. Projections suggest that the number of people being diagnosed with cancer will increase still further in the coming years, and will be nearly 50% higher in 2040 than in 2020. The number of cancer deaths has also increased, from 6.2 million in 2000 to 10 million in 2020. More than one in six deaths is due to cancer. This reinforces the need to invest in both the fight against cancer and cancer prevention. The successful introduction of information and communication technologies (ICT) in medical practice is an important stake in the renovation of the health system and more precisely in cancer care Breast cancer (BC) is the most common cancer in women, affecting about 10% of all women at some stages of their life. A Classification algorithm, like Decision Tree are broadly used in the world of medicines to categorize the information for diagnosis. The process of Feature Selection helps in increasing the overall accuracy of the classifying model since it removes inappropriate attributes. There are many established ways of detecting and diagnosing cancer but they majorly depend on trained physicians, with the help of medical imaging, we notice certain indications that commonly seem to be visible in the advance stages of cancerous cells. Machine learning provides several probabilistic and statistical methods that let intelligent systems to learn from past knowledges which repeats to notice and recognize patterns from a dataset. But the limitations are that either they use faulty dataset or they don't wrangle the data correctly or select features properly. The aim of this very project is to guarantee that the benign and malignant classes of breast cancer are predicted and grouped accurately. This paper mainly gives a comparison between the performance of five classifiers: Support Vector Machine (SVM)), Random Forest, Logistic Regression, Decision tree and K-Nearest Neighbors (KNN Network) which according to research community are among the most influential data mining algorithms and among the top 10 data mining algorithms. Our objective is to predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity. The rest of this paper is organized as follows section 2 introduces methods and results of previous research on breast cancer diagnosis. Section 3 describes the proposed methodology for our work. Section 4 presents and explains in detail the experiments results. Section 5 concludes the paper.

2. Literature Survey

J. Sultana et al. from Majmaah University [1] worked on Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers.

The proposed approach makes use of Breast cancer dataset collected from the UCI machine learning repository. Data set is pre-processed first and fed to various classifiers like Simple Logistic-regression method, K-star, Multi-Layer Perceptron (MLP), Random Forest, Decision Trees (DT), PART, Multi-Class Classifiers and REP Tree. The results obtained are evaluated on various parameters. Result analysis reveals that among all the classifiers Simple Logistic Regression yields the best predictions.

Ahmad LG* et al. from Islamic Azad University of Tehran-Iran,[2] worked on using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence

They implemented machine learning techniques, i.e., Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN) to develop the predictive models. The main goal of their paper is to compare the performance of these three well-known algorithms on our data through sensitivity, specificity and accuracy. According to them SVM classification model predicts breast cancer recurrence with least error rate and highest accuracy. The predicted accuracy of the DT model is the lowest of all. The results are achieved using 10-fold cross-validation for measuring the unbiased prediction accuracy of each model.

Wenbin Yue et al. from Brunel University London [3] worked on Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis.

In this paper, they aim to review ML techniques and their applications in BC diagnosis and prognosis. Firstly, they provide an overview of ML techniques including artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbors (k-NNs). Then, they investigate their applications in Breast Cancer. Their primary data is drawn from the Wisconsin breast cancer database (WBCD) which is the benchmark database for comparing the results through different algorithms. Finally, a healthcare system model of their recent work is produced.

Nikita Rane et al. from Xavier Institute of Engineering, Mumbai [4] worked on Breast Cancer Classification and Prediction using Machine Learning.

This paper proposed now presents a comparison of six machine learning (ML) algorithms: Naive Bayes (NB), Random Forest (RT), Artificial Neural Networks (ANN), Nearest Neighbour (KNN), Support Vector Machine (SVM) and Decision Tree (DT) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset which is extracted from a digitised image of an MRI. For the implementation of the ML algorithms, the dataset was partitioned into the training phase and the testing phase. The algorithm with the best results will be used as the backend to the website and the model will then classify the cancer as benign or malignant.

Saria Eltalhi et al. from University of Benghazi, [5] worked on Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques.

The aim of this research is to review the role of machine learning and data mining techniques in breast cancer detection and diagnosis. Most of these studies concentrated on diagnoses and prognoses breast cancer using WEKA tool. Most of the studies compared different classification algorithms to breast cancer prediction such as Decision tree, Naïve Bayes, and Artificial Neural Network.

Rahul Raj Pandey from VIT Bhopal University, Madhya Pradesh, [6] worked on Analysis and Prediction of Breast Cancer using Machine Learning Techniques.

This paper presents an outline of evolved machine learning techniques in this medical field by applying machine learning algorithms on breast cancer dataset like Logistic regression, Random Forest, Decision Trees (DT) etc. The aim of this project is to guarantee that the benign and malignant classes of breast cancer are predicted and grouped accurately.

El Habib Benlahmard et al. from, University of Mons, Mons, Belgium [7] worked on Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis.

The main objective of this research paper is to predict and diagnosis breast cancer, using machine learning algorithms, and find out the most effective with respect to confusion matrix, accuracy and precision. It is observed that Support vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%).All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.

Proposed Method
This proposed system presents a comparison of machine learning algorithm such as SVM, KNN, Random Forest, Decision Tree, Logistic Regression. The dataset is obtained from UCI Repository. The algorithm that gives the best result will be given as model to the website for the users. The text extraction process takes place from the image and automatically submitted as input for model.

3. Methodology
In our project, we proposed the model that predicts the breast cancer is malign or benign. We have gathered datasets from different sources to train and test the model, and to test the accuracy. The highest accuracy attained model will be used for user's input to predict whether user has cancer or not.

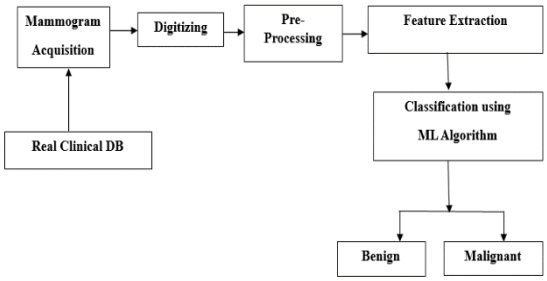

Figure (4.a) System Architecture

Phase 1: Pre-Processing Data
The first phase we do is to collect the data that we are interested in collecting for pre-processing and to apply classification and Regression methods. Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and lacking certain to contain many errors. Data pre-processing is a proven method of resolving such issues. For pre-processing we have used standardization method to pre-process the UCI dataset. This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be. In this case we collect the Breast Cancer samples which are Benign and Malignant. This will be our training data.

Phase 2: Features Selection
In machine learning and statistics, feature selection, also known as variable selection, attribute selection, is the process of selection a subset of relevant features for use in model construction. We have used Wrapper Method for Feature Selection. The important features found by the study are: 1. Concave points worst 2. Area worst 3. Area se 4. Texture worst 5. Texture mean 6. Smoothness worst 7. Smoothness mean 8. Radius mean 9. Symmetry means.

Phase 3: Model Selection
Supervised learning is the method in which the machine is trained on the data which the input and output are well labelled. The model can learn on the training data and can process the future data to predict outcome. They are grouped to Regression and Classification techniques. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B (Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning.

Phase 4: Prediction
Machine learning is using data to answer questions. So Prediction, is the step where we get to answer some questions. This is the point of all this work, where the value of machine learning is real.

Methods Used:
(1) Logistic Regression:
Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation like Linear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables.

The general workflow is:
1. get a dataset
2. train a classifier
3. make a prediction using such classifier

(2) k-Nearest Neighbour (k-NN):
K-Nearest Neighbour is a supervised machine learning algorithm as the data given to it is labelled. It is a nonparametric method as the classification of test data point relies upon the nearest training data points rather than considering the dimensions (parameters) of the dataset.
Algorithm:
1. Input the dataset and split it into a training and testing set.
2. Pick an instance from the testing sets and calculate its distance with the training set.
3. List distances in ascending order.
4. The class of the instance is the most common class of the 3 first trainings instances (k=3).

(3) Support Vector machine:
Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data. SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm. In an SVM model, each data item is represented as points in an n-dimensional space where n is the number of features where each feature is represented as the value of a coordinate in the n-dimensional space. Here's how a support vector machine algorithm model works:
1. First, it finds lines or boundaries that correctly classify the training dataset.
2. Then, from those lines or boundaries, it picks the one that has the maximum distance from the closest data points.

(4) Random Forest:
Random forest, like its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

(5) Decision Tree:
Decision Tree is a predictive modelling tool that can be applied across many areas. It can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions.

4. Results
On the Wisconsin Breast Cancer Diagnostic dataset (WBCD) we applied five main algorithms which are: SVM, Random Forests, Logistic Regression, Decision Tree, K-NN, calculate, compare and evaluate different results obtained based on confusion matrix, accuracy, sensitivity, precision, AUC to identify the best machine learning algorithm that are precise, reliable and find the higher accuracy After an accurate comparison between our models, we found that Random Forest Machine achieved a higher efficiency of 97.66%.

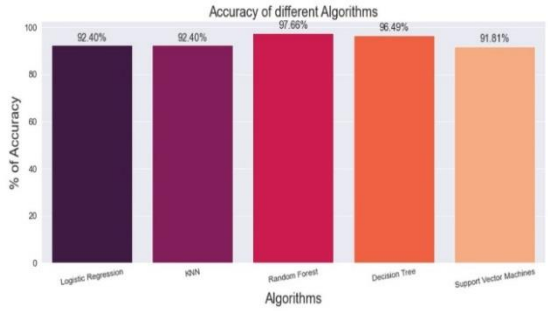

Figure (5.a) Accuracy of different models

5. Conclusion
Breast cancer if found at an early stage will help save lives of thousands of women or even men. These projects help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be

able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes.

Scope of Future Enhancement

For future works to apply these same algorithms and methods on other databases to confirm the results obtained via this database, as well as, in our future works, we plan to apply our and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

Acknowledgment

We would like to express my gratitude to internal guide Prof. Krishna Mohana A J who guided us throughout the project with his insight and knowledge. We would also like to thank our institute, for providing us an opportunity to conduct the project.

6. References

[1] Jabina Sultana,Abdul Khader Jilani, ”Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers”, International Journal of Engineering Research & Technology, 2018.

[2] Ahmad L G, Eshlaghy AT,Poorebrahimi A, Ebrahimi M and Razavi AR, ”Using three Machine Learning Techniques for Predicting Breast Cancer Recurrence”, Iran Health and Medical Informatic, 2013.

[3] Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis”, Brunel University London, 2019

[4] Nikita Rane, Jean Sunny, Rucha Kanade, ” Breast Cancer Classification and Prediction using Machine Learning”, Xavier Institute of Engineering, Mumbai, 2020.

[5] Saria Eltalhi, Huda Kutrani, “Breast Cancer Diagnosis and Prediction Using Machine Learning and Data Mining Techniques”, University of Benghazi, Libya, April 2019.

[6] Shakkeera L, Rahul Raj Pandey, Rahul Bhardwaj, Sidhya Virya Singh, “Analysis and Prediction of Breast Cancer using Machine Learning Techniques”, International Journal of Engineering and Advanced Technology (IJEAT) ,December 2020.

[7] Habib Benlahmard, Ait Abdelouha, Kawtar Aarika, “Machine Learning Algorithms for Breast Cancer Prediction And Diagnosis”, University of Mons,Aug 2021.