International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 02, February-2020

Breast Cancer Classification and Prediction using Machine Learning

Nikita Rane
Dept. of Information Technology,
Xavier Institute of Engineering,
Mumbai - 400016, India

Jean Sunny
Dept. of Information Technology
Xavier Institute of Engineering,
Mumbai - 400016, India

Rucha Kanade
Dept. of Information Technology,
Xavier Institute of Engineering,
Mumbai - 400016, India

Prof. Sulochana Devi
Dept. of Information Technology,
Xavier Institute of Engineering,
Mumbai - 400016,India.

***Abstract*—Breast cancer is a dominant cancer in women worldwide and is increasing in developing countries where the majority of cases are diagnosed in late stages. The projects that have already been proposed show a comparison of machine learning algorithms with the help of different techniques like the ensemble methods, data mining algorithms or using blood analysis etc. 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.**

Keywords - Breast cancer classification, Breast cancer prediction, benign, malignant, Naïve Bayes, KNN, Support Vector Machine, Artificial Neural Network, Random Forest, Decision tree, SQLAlchemy.

I.INTRODUCTION

This research paper has gathered information from ten different papers based on breast cancer using machine learning and other techniques such as ultrasonography, blood analysis etc. The project by S. Gokhale, is using the ultrasonography(USG) technique which is a powerful method in detecting details about the breast mass that usually cannot be detected even by mammography.

Another project by Pragya Chauhan and Amit Swami, which is based on the ensemble method usually used to increase the prediction accuracy of breast cancer. A Genetic algorithm based weighted average method that includes crossover and mutation is used for the prediction of multiple models.

Further more, a project by Abien Fred M. Agarap uses different methods like GRU-SVM, NN, multilayer perceptron (MLP), softmax regression to classify the dataset into benign or malignant. A project by Priyanka Gupta shows the comparison of the lesser invasive techniques such as Classification and Regression Trees (CART), random forest, nearest neighbour and boosted trees. These four classification models are chosen to extract the most accurate model for predicting cancer survivability rate.

Another project by Muhammet Fatih Aslan, Yunus Celik , and Kadir Sabanci, Akif Durdu that uses
the blood analysis dataset from UCI. It draws results that are from methods like Extreme Learning Machine (ELM), ANN etc. it also has an added MATLAB GUI environment that for classification with ANN.

Further more, a project by Yixuan Li and Zixuan Chen shows a performance evaluation using three indicators including prediction accuracy values, F-measure metric and AUC values are used to compare the performance of these five classification models. Other experiments show that random forest model can achieve better performance and adaptation than other four methods. A project by Mumine Kaya Keles, which is a comparative study of data mining classification algorithms. Another project by Sang Won Yoon and Haifeng Wang that uses four data mining models are applied in this paper, i.e., support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree. Furthermore, feature space is highly deliberated in this paper due to its high imapct on the efficiency and effectiveness of the learning process. Lastly a project by Wenbin Yue and Zidong Wang that shows the algorithms that helped them with the diagnosis and prognosis of their dataset.

II. REVIEW OF LITERATURE

Ultrasound characterisation of breast masses by S. Gokhale written by proposed a system where they found that doctors have known and experienced that breast cancer occurs when some breast cells begin to grow abnormally. These cells divide more briskly and disperse faster than healthy cells do and continue to accumulate, form- ing a lump or mass that the may start causing pain. Cells may spread rapidly through your breast to your lymph nodes or to other parts of your body. Some women can be at a higher risk for breast cancer
because of their family history, lifestyle, obesity, radiation, and reproductive factors. In the case of cancer, if the diagnosis occurs quickly, the patient can be saved as there have been advances in cancer treatment. In this study we use four machine learning classifiers which are Naive Bayesian Classifier, k-Nearest Neighbour, Support Vector Machine, Artificial Neural Network and random forest.

Harmonic imaging and real-time compounding has been shown to enhance image resolution and lesion characterisation. More recently, USG elastography seems to be quite ncouraging. Initial results show that it can improve the specificity and positive predictive value of USG within the characterisation of breast masses. The reason why any lesion is visible on mammography or USG is that the relative difference within the density and acoustic resistance of the lesion, respectively, as compared to the encompassing breast tissue. [1]

Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach written by Pragya Chauhan and Amit Swami proposed a system where they found that Breast cancer prediction is an open area of research. In this paper different machine learning algorithms are used for detection of Breast Cancer Prediction. Decision tree, random forest, support vector machine, neural network, linear model, adabost, naive bayes methods are used for prediction.

An ensemble method is used to increase the prediction accuracy of breast cancer. New technique is implemented which is GA based weighted average ensemble method of classification dataset which over- came the limitations of the classical weighted average method. Genetic algorithm based weighted average method is used for the prediction of multiple models. The comparison between Particle swarm optimisation(PSO), Differential evolution(DE) and Genetic algorithm(GA) and it is concluded that the genetic algorithm outperforms for weighted average methods. One more comparison between classical ensemble method and GA based weighted average method and it is concluded that GA based weighted average method outperforms. [2]

On Breast Cancer Detection: An Application of Machine Learn ing Algorithms on the Wisconsin Diagnostic Dataset by the Abien Fred M. Agarap. In this paper, six machine learning algorithms are used for detection of cancer . GRU-SVM model is used for the diagnosis of breast cancer GRU-SVM, Linear Regression, Multilayer Perceptron (MLP), Nearest Neighbour (NN) search, Softmax Regression, and Support Vector Ma- chine (SVM) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by measuring their classification test accuracy, and their sensitivity and specificity values. The said dataset consists of features which were computed from digitised images of FNA tests on a breast mass. For the implementation of the ML algorithms, the dataset was partitioned in the following fashion 70 percent for training phase, and 30 percent for the testing phase. Their results were that all presented ML algorithms exhibited high performance on the binary classification of carcinoma, i.e. determining whether benign tumour or malignant tumour. Therefore, the statistical measures on the classification problem were also satisfactory. To further corroborate the results of this study, a CV technique such as k-fold cross-validation should be used. The appliance of such a way won't only provide a more accurate measure of model prediction performance, but it'll also assist in determining the foremost optimal hyper-parameters for the ML algorithms. [3]

Analysis of Machine Learning Techniques for Breast Cancer Prediction by the Priyanka Gandhi and Prof. Shalini L of VIT university, vellore. In this paper, ML techniques are explored in order to boost the accuracy of diagnosis. Methods such as CART, Random Forest, K-Nearest Neighbours are compared. The dataset used is acquired from UC Irvine Machine Learning Repository. It is found that KNN algorithm has much better performance than the other techniques used in comparison. The most accurate model was K-Nearest Neighbour. The classification model such as Random Forest and Boosted Trees showed the similar accuracy. Therefore, the most accurate classifier can be used to detect the tumour so that the cure can be found in early stage. [4]

Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data by the Muhammet Fatih Aslan, Yunus Celik, Kadir Sabanci, and Akif Durdu for carcinoma early diagnosis. During this paper, four different machine learning algorithms are used for the early detection of carcinoma. The aim of this project is to process the results of routine blood analysis with different ML methods. Methods used are Artificial Neural Network (ANN), Extreme Learning Machine (ELM), Support Vector Machine (SVM) and Nearest Neighbor (k-NN). Dataset is taken from the UCI library. In this dataset age, BMI, glucose, insulin, homeostasis model assessment (HOMA), leptin, adiponectin, resistin, and chemokine monocyte chemoattractant protein (MCP1) attributes were used. Parameters that have the best accuracy values were found by using four different Machine Learning techniques. This dataset includes age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin and MCP1 features that can be acquired in routine blood analysis. The significance of these data in breast cancer detection was investigated by ML methods. The analysis was performed with four different ML methods. k-NN and SVM methods are determined using Hyperparameter optimization technique. The highest accuracy and lowest training time were given by ELM which was 80%. and 0.42 seconds. [5]

Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction by Yixuan Li and Zixuan Chen used two datasets in the study. The study firstly collects the data of the BCCD dataset which contains 116 volunteers with 9 attributes and data of WBCD dataset which contains 699 volunteers and 11 attributes. Then we preprocesses the raw data of WBCD dataset and obtained the info that contains 683 volunteers with nine attributes and therefore the index indicating whether the volunteer has the malignant tumour. After comparing the accuracy, F-measure metric and ROC curve of 5 classification models, the result has shown that RF is chosen as the primary classification model during this study. Therefore, the results of this study provide a reference for experts to distinguish the character of carcinoma .In this study, there are still some limitations that ought to be solved in further work. For instance, though there also exist some indices people haven't found yet, this study only collects the info of 10 attributes during this experiment. The limited data has an impact on the accuracy of results. additionally , the RF can also be combined with other data mining technologies to get more accurate and efficient results in the longer term work. [6]

The purpose of the paper "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study” by Mumine Kaya Keles was to predict and detect breast cancer early even if the tumour size is petite with non-invasive and painless methods that use data mining classification algorithms. Therefore, a comparison of data mining classification algorithms was made with the Weka tool. In this paper, the Weka data mining software was applied to an antenna dataset so as to examine the efficacy of data mining methods in the detection of breast cancer. The dataset that was created had 6006 rows/values, 5405 of which were used as the training dataset, while 601 were used as the test data set. The dataset was then converted to the arff format, which is the file type used by the Weka tool. The 10-fold cross-validation was then used to obtain the most authentic results using the Knowledge Extraction based on Evolutionary Learning data mining software tool. Random forest performed the best during the 10 fold cross-validation giving an average accuracy of 92.2 percent. [7]

The project “Breast Cancer Prediction Using Data Mining Method” by Haifeng Wang and Sang Won Yoon is used to test the influence of feature space reduction, a hybrid between principal component analysis (PCA) and related data mining models is proposed, which applies a principle component analysis method to reduce the feature space. To evaluate the performance of these models, two widely used test data sets are used, Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995). 10-fold cross-validation method is implemented to estimate the test error of each model. PCs-SVM is the highest for WBC data that is a 97.47 percent, and PCi-ANN is the best considering accuracy for WDBC data that is 99.63%. The reason for better results from PCA preprocessing is because the principal components only represent a large part of the information in the complete data space, which to some extent can reduce data noise, as a result, feature space is enriched (elite effect). [8]

“Machine Learning with Applications in breast cancer Diagnosis and Prognosis” by Wenbin Yue and Zidong Wang In this paper, they provided explanations of various ML approaches and their applications in BC diagnosis and prognosis wont to analyse the info within the benchmark database WBCD. ML techniques have shown their remarkable ability to enhance classification and prediction accuracy. Although many algorithms have achieved very high accuracy in WBCD, the event of improved algorithms remains necessary. Classification accuracy may be a vital assessment criteria but it's not the sole one. Different algorithms consider different aspects, and have different mechanisms. Although for several decades ANNs have dominated BC diagnosis and prognosis, it's clear that more recently alternative ML methods are applied to intelligent healthcare systems to supply a spread of options to physicians. [9]

III. BACKGROUND

A. Breast Cancer Classification

Breast cancer classification divides carcinoma into categories depending on how they have spread or if they have spread at all. Classification algorithms predict one or more discrete variables, supported the opposite attributes within the dataset. data processing software is required to run the classification algorithms. the aim of classification is to pick the simplest treatment. Classification is vital because it allows scientists to spot, group, and properly name organisms via a uniform system. Classification and clustering are two widely used methods in data processing . Clustering methods aim to extract information from a knowledge set to get groups or clusters and describe the info set itself. Classification, also referred to as supervised learning in machine learning, aims to classify unknown situations supported learning existing patterns and categories from the info set and subsequently predict future situations. The training

set, which is employed to create the classifying structure, and therefore the test set, which tends to assess the classifier, are commonly mentioned in classification tasks Classification may be a quite complex optimisation problem. Many ML techniques are applied by researchers in solving this classification problem. The most famous algorithm that is used for breast cancer classification or prediction is an artificial neural network, random forest, support vector machine, etc. Scientists strive to seek out the simplest algorithm to realise the foremost accurate classification result, however, data of variable quality also will influence the classification result. Further, the rarity of knowledge will influence the number of algorithm applications also. If the carcinoma is found early, there are more treatment options and a far better chance for survival. Women whose carcinoma is detected at an early stage have a 93 percent or higher survival rate within the first five years. Getting checked regularly can put your mind comfortable. Finding cancer early can also save your life.

B. Machine learning algorithms

Machine learning is an application of AI (AI) that gives systems the power to automatically learn and improve from experience without being manually programmed. Machine learning focuses and depends on the event of computer programs that will access the data provided and use it to learn for themselves. The method of learning begins with data or datasets, examples, experiences, or instructions, so they can then figure out a pattern and or improve them in the near future, if necessary.

1. Naive Bayes :

A Naive Bayes classifier may be a probabilistic machine learn- ing model that’s used for the classification tasks. The crux of the classifier is predicated on the Bayes theorem. Using Bayes theorem, we will find the probability of an event, as long as B has occurred. Here, B is that the evidence and A is that the hypothesis. the idea made here is that the predictors/ features are independent. that's the presence of one particular feature that doesn't affect the opposite. Hence it's called naive.

2. Random forest :

Random forests also known as random decision forests creates a large number of trees that achieve their output through ensemble learning methods for classification, regression. Bagging and feature randomness are the features it uses to construct those trees. The random forest has an advantage over the decision tree which, is that it does not overfit the data.

3. Artificial neural network :

Artificial neural networks (ANN) or neural network systems are computing systems that mimic the functioning of a human brain. The main aim of the algorithm is to provide a faster result with more accuracy than an old or traditional system. if the algorithm has been given the data or an image

about a particular object then the algorithm will quickly be able to identify or categorise images that do not contain the said object.

4. Support vector machine :

In machine learning, support vector machines are supervised models. A support vector machine creates a hyperplane when classifying the objects. A hyperplane is a line on a plane that distinguishes the two classes. Given a group of coaching examples, each marked as belonging to at least one or the opposite of two categories, an SVM training algorithm builds a model that assigns new examples to at least one category or the opposite, making it a non-probabilistic binary linear classifier (although methods like Platt scaling exist to use SVM during a probabilistic classification setting). New examples are then mapped into that very same space and predicted to belong to a category supported the side of the gap on which they fall.

5. K nearest neighbours:

KNN (K- Nearest Neighbours) is one among many supervised learning algorithms utilised in data processing and machine learning, it's a classifier algorithm where the training is predicated "how similar" may be a data from other. It is a lazy algorithm. KNN works by finding the distances between a point and all the examples within the data, selecting the required number examples (K) closest to the point, then votes for the leading frequent label.

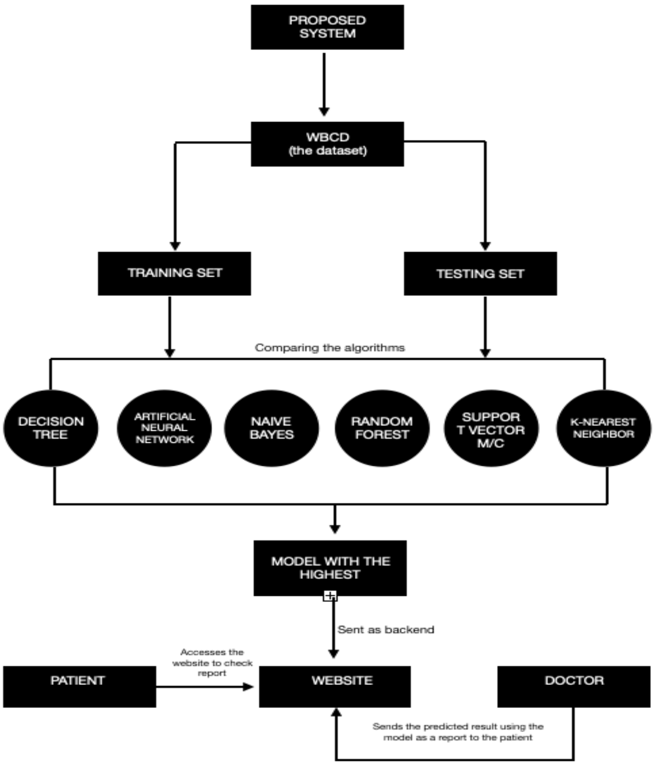
6. Decision Tree :

DTs apply a top-down approach to data so that given a knowledge set, they struggle to group and label observations that are similar between them, and appearance for the simplest rules that split the observations that are not the same between them until they reach a certain degree of similarity. They use a layered splitting process, where at each layer they struggle to separate the info into two or more groups, in order that data that fall under an equivalent group is most similar to every other (homogeneity), and groups are as die-rent as possible from one another.

IV. PROPOSED SYSTEM

This proposed system presents a comparison of machine learning (ML) algorithms: Support machine vector(SVM), Decision Tree(DT), Random Forest(RT), Artificial Neural Net- works(ANN), Naive Bayes (NB), Nearest Neighbour (NN) search. The data-set used is obtained from the Wisconsin datasets.For the implementation of the ML algorithms, the dataset was partitioned into the training set and testing set. A comparison between all the six algorithms will be made. The algorithm that gives the best results will be supplied as a model to the website. The website will be made from a python framework, called flask. And it will host the database on Xampp or Firebase or inbuilt Python and flask libraries. This data set is available on the UCI Machine Learning Repository. It consists of 32 real world attributes which are multivariate. The total number of instances is 569 and there are no missing values in this data set. The process of the proposed system is as follows,

1. The patient books an appointment through our website.
2. The patient will then meet the doctor offline for the respective appointment.
3. The doctor will first check the patient manually, then perform a breast mammogram or an ultrasound. That ultra sound will show an image of the breast consisting the lumps or not.
4. If the lumps are detected, a biopsy will be performed. The digitised image of the Fine Needle Aspirate (FNA) is what forms the features of the dataset.
5. Those numbers will be provided to the system by the doctor and the model will detect if it's a benign or a malignant cancer.
6. The report will be then forwarded to the patient on their respective account.



7. Fig 1. Block diagram

V. 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.

VI. REFERENCES

[1] ”Ultrasound characterisation of breast masses”, The Indian journal of radiology imaging by S. Gokhale., Vol. 19, pp. 242-249, 2009. K. Elissa, “Title of paper if known,” unpublished.

[2] “Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach” by Pragya Chauhan and Amit Swami, 18 October 2018

[3] “On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset” by Abien Fred M. Agarap, 7 February 2019

[4] “Analysis of Machine Learning Techniques for Breast Cancer Prediction” by the Priyanka Gupta and Prof. Shalini L of VIT university, vellore, 5 May 2018.

[5] “Breast Cancer Diagnosis by Dierent Machine Learning Methods Using Blood Analysis Data“ by the Muhammet Fatih Aslan, Yunus Celik , Kadir Sabanci and Akif Durdu, 31 December, 2018

[6] ”Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction”, by Yixuan Li, Zixuan Chen October 18, 2018

[7] “Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study” by Mumine Kaya Keles, Feb 2019

[8] “Breast Cancer Prediction Using Data Mining Method ” by Haifeng Wang and Sang Won Yoon, Department of Systems Science and Industrial Engineering State University of New York at Binghamton Binghamton, May 2015.

[9] “Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis” by Wenbin Yue , Zidong Wang, 9 May 2018