comien

International Journal on Computational Engineering

Journal Homepage www.comien.org/index.php/comien

Predictive Model Using Machine Leaning Approach for the Detection of Breast Cancer

Kuldeep Pathoee [a], Deepesh Rawat [b,*], Dilip K.J.B Saini [c]

[a] Mechanical Engineering, Swami Rama Himalayan University, Jollygrant, Dehradun, 248160, India
[b] Electronics & Communication Engineering, Swami Rama Himalayan University, Jollygrant, Dehradun, 248160, India
[c] Computer Science and Engineering, Swami Rama Himalayan University, Jollygrant, Dehradun, 248160, India

Corresponding author: *deepeshrawat@gmail.com

Abstract*—One of the most common cancer among the women, which is diagnosed and increasing rapidly worldwide, is Breast cancer. Every year the percentage of women diagnoses by this invasive cancer is increasing. It is the major cause of death in women globally. It is critical for a healthy life to predict and diagnose cancer at an early stage. Early detection of breast cancer can considerably improve the prognosis and increase the likelihood of a patient's survival, since it allows for timely clinical treatment. As a result, fast analytics and feature extraction methods are required for high-accuracy cancer prediction, which can be accomplished utilizing Machine learning. In our research work which we present in this paper, we compare various machine learning (ML) algorithms including i) Random Forests ii) Logistic Regression, iii) Decision Tree and iv) Support Vector Machine. We evaluate and analyze the performance of these entire algorithms using area under the receiver operating characteristic (AUROC) curve, and confusion metrics and find the best machine learning model for prediction of breast cancer. The findings are calculated using the evaluation criteria of Precision, Recall, Accuracy, and Specificity. Confusion matrix based on evaluation parameters that put a greater emphasis on predicted cases. A performance evaluation is computed for various machine learning models. For simulation, we used the Wisconsin Dataset of Breast Cancer (WDBC) in our research. After simulation, the SVM model obtained 98.24% accuracy on testing test with an AUC of 0.993, while the logistic regression achieved 94.5% accuracy with an AUC of 0.998. With their mathematical models, these algorithms can be further tweaked to improve breast cancer prediction.

***Keywords*—Supports vector machine; random forests; decision tree; logistic regression; area under ROC curve; receiver operating characteristics.**

Manuscript received 16 Jan. 2024; revised 19 Feb. 2024; accepted 7 Jun. 2024. Date of publication 30 Jun. 2024.

I. INTRODUCTION

According to International Agency of Research Cancer, in 2020, approx 19.3 million new cases of cancer were diagnosed and large number fatalities of about 10 million are expected because of cancer [1]. Now breast cancer is most frequent cancer among women and has surpassed lung cancer and prostate cancer as the most invasive cancer in the world. It is estimated about 2.3 million (11.7% of the total cancer cases) diagnosed as breast cancer in 2020, which means ones in every 8 cancers case is breast cancer. Breast cancer is expected to kill 685000 people in 2020, major of these fatalities occurs in low-resource areas [1]. According to experts, the number of people diagnosed with cancer will double by 2040.As a result, there is a need to develop or invest in novel cancer diagnosis tools and methods. The successful integration of artificial intelligence and data analysts into medical practice has the potential to revolutionize the health-care system, as well as cancer treatment by examining a vast amount of health-care information

This research proposed, compare and demonstrate four machine learning models for detecting breast cancer that are i) random forest ii) decision tree iii)logistic regression and iv) support vector machine. The WDBC dataset (Wisconsin-Breast Diagnostic Cancer) from UCI machine learning repository is used in our research work. The aim of this research is to demonstrate that machine learning techniques like SVM, random forests, decision trees, and logistic regression can be used to solve classification problems. This research also offers the framework for a comparative analysis of the various approaches and helps in identifying the optimal machine learning method for creating a machine-learning model. The work in this paper is demonstrated as follows: The second section illustrate literature survey of

previous similar efforts as well as their outcomes. In Section 3, the methodology, data, and experimental setting are all discussed. In Section 4 we demonstrate our results and analyze table and graph of our work. In section 5 we conclude our work that contains the experimental data and result comments. The next paragraphs discuss over these sections in details.

In recent years, machine-learning algorithms have been increasingly applied in the prediction and diagnosis of breast cancer. To improve categorization, prediction, and detection operations, machine learning approaches rely on computer models and information gathered from prior and earlier data. Researchers employed mammography scans, SEER data, WDBC data, and data from a number of hospitals to diagnose and forecast cancer using algorithms such as Random forest, K-nearest Neighbour(KNN), Support vector machine, and others. Using data from the Iranian centre, L. G. Ahmad and Eshlaghy [2] investigated the performance of decision tree (C4.5), SVM, and ANN for the diagnosis of breast cancer.

Another study [3] was conducted. S. Nayak and D. Gope utilised 3D images and various ML algorithms to diagnose breast cancer, demonstrating that the Support vector machine (SVM) algorithm performs the best overall.

M. H. Memon, Jian Ping Li[4] analyzed the dataset from Wisconsin Breast(Diagnostics) Cancer (WBC) database and use elimination technique (recursive feature) for enhancing the Support vector machine model. To check the accuracy rate performance matrix was designed and demonstrate that on linear kernel SVM achieved accuracy of 99%, whereas on RBF and Polynomial SVM achieve accuracy of 98% and 97% respectively.

S. Alghunaim and H. Al-Baity[5] analyzed Gene Expression and DNA methylation data on Weka and Spark tools, and find the tumors using various machine learning algorithm with the accuracy of SVM 99.8 % and 98.03% on spark and weeka tool respectively.

H. Asri and H. Mousannif[6] demonstrate that SVM model achieve best performance in term of low error rate and precision and demonstrate the accuracy with 97.13%

B. Gayathri and C. Sumathi [7] compare the RVM algorithm with other ML algorithms for diagnosing breast cancer and use linear discriminant technique to minimise features. In their research work, they use WBC dataset with RVM method to classify it, resulting 96 percent accuracy. The simulation data yielded a sensitivity and specificity of 98 percent and 94 percent, respectively.

II. Material and Method

In our research work, we used machine-learning classifiers on the WBCD dataset to present one of the most successful and predictive machine-learning model for the detection of breast cancer. Different machine learning classifiers like Random forest, decision tree, logistic regression, and support vector machine are applied, and the results are evaluated using area under receiver operating characteristic (AUROC) curve to determine which model is the best and provides the highest accuracy for the prediction of breast cancer.

The proposed architecture of our research work is shown in Figure 2 where whole architecture is carried out in four steps. Figure 2.a shows loading of WDBC dataset which is first step of our work. After loading of dataset step 2 includes, pre-processing of obtained data is performed in our methodology includes: data cleansing, followed by parameters selection, set target and feature extraction as show in Figure 2.b . All the steps are discuss below in details

The proposed model's classification accuracy is assessed using the validation dataset (Wisconsin breast cancer diagnosis). This is accomplished by dividing the data into two halves, the first half of the data, which is about 70%, is used to create a machine learning (ML) model, referred as training data. Rest of the 30% of the data is test data, which is used to demonstrate how well the model works. For evaluating and assessing the multiple ML models, the experiment is carried out in Python programming with numerical and scientific libraries.

A. Acquisition of data

The collection of data from various sources for experiments is known as data acquisition. The Wisconsin-Breast (Diagnostic) Cancer (WBC) dataset utilized in this study came from the open-source machine learning repository at the University of California, Irvine (UCI).

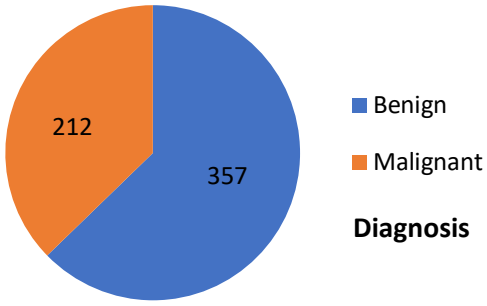

Fig. 1 Wisconsin Breast (Diagnostics) Cancer Datasets

This dataset as shown in Figure 1, contains 569 instances (diagnosis, benign 357 cases, and malignant 212 cases) with 32 attributes, including two class attribute labels. Ten real-valued features are computed for each cell nucleus: (radius, texture, perimeter, area, smoothness', compactness, concavity, concave points, symmetry, and fractal dimension) [8], and an ID number. The features mentioned above are derived from a digitised image of a fine needle aspiration method, which is performed on a breast mass and are used to define the properties of the cell nuclei in the image.

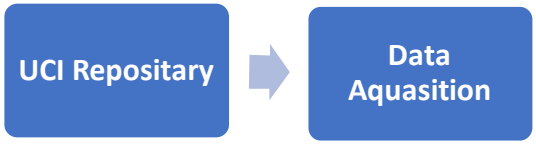

Fig. 2a Step 1 Loading of Dataset

B. Preprocessing of data

The data acquired for the research may contain noise and inconsistencies, so data preparation is used to enhance the dataset quality and produce data that is free of any type of

contamination, so that it can utilize for modeling [9]. Data preprocessing involves a number of steps, including data cleansing, feature selection, extraction among others. The data is first divided into two datasets during data preprocessing: i) training and ii) test. The training dataset is made up of 399 observations with 31 variables that are utilized to train the machine learning method. The dataset kept for validation contains 170 instances across 31 features which is used during the prediction step. Centering and scaling are two further preprocessing measures, which are conducted on dataset.

Fig. 2b Step 2 Pre Processing

The association between several characteristics has been examined in the above figure 2. Correlation between two characteristics has a value ranging from -1 to 1. They are inversely proportional if they are -1, and they are directly proportional if they are 1. The closer the value is to the two extremes, the clearer the association. In the diagram above, 12 features are compared to one another, their association with diagnostic features is examined, and all results are shown as percentages. The colour bar is used to visually distinguish between characteristics that are strongly connected and those that are substantially unrelated. The black grid represents a feature that is highly unrelated, while the darker grey grid represents a feature that is highly correlated. The same correlation matrix was plotted for all of the features, and the beast feature was chosen based on it. The correlation matrix aids in identifying the characteristics that trigger breast cancer in women.

C. Machine Learning (ML) algorithms

After preprocessing, the data is further classified using machine learning algorithms. In this section the machine learning model is built after processing of data. The Machine Learning classification categories are SVM with RBF, decision tree, logistic regression, and random forest with same random state. The training data is utilized to train the models to differentiate between benign and malignant tumors'. Furthermore, the data dimensions are reduced by the feature selection and extraction method to train the models

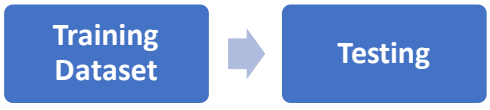

Fig. 2c Step 3 Build Model

D. Performance evaluation

Testing of suggested model(s) generated is part of evaluating the performance of a machine learning algorithm. The evaluation in this study is carried out by comparing the outputs of the model with the actual values of data. To evaluate the performance of the models the test dataset is used in distinguishing benign and malignant cancers throughout this phase.

By comparing the actual and expected results, the confusion matrix is formed. The performance of the classifier is computed using the data in the matrix [12]. Accuracy, AUROC, precision, recall, sensitivity and other performance metrics can be used to measure and evaluate ML model performance.

Fig. 2d Step 4 Result Computation

III. Results and Discussion

For minimizing the dimension of processed data feature selection and extraction procedures are applied, as a result relatively smaller versions of the original dataset are produced. In order to train the datasets, SVM, random forest, decision tree, and logistic approaches were used. To analyze, compare and determine the best algorithm for breast cancer prediction, we use the Confusion matrix, accuracy, and precision, as well as sensitivity as a performance matrix. The confusion matrix is used to assess classification performance. The most common method to identify the performance of metric for the classification of various algorithms is "accuracy". It is the proportion of correct predictions to total predictions. Other parameters that are used to measure the performance of machine learning models are sensitivity and precession. The precision of any model is described as the number of correct documents returned by machine learning model whereas sensitivity in machine learning is defines as number of positives returned by model. Accuracy percentage, which we get for various machine leaning models using breast cancer diagnostic dataset, is shown in Table 1.

TABLE I
ACCURACY ON TESTING AND TRAINING DATASET

Algorithm	Training Set Accuracy	Testing Set Accuracy
Support Vector Machine	98.49	98.24
Random Forest	99.49	96.49
Logistic Regression	98.74	94.15
Decision Tree	100	96.49

From the above table it is clear that accuracy for support vector machine is about 98.24%, which is highest as compare to other model, while accuracy for logistic regression is about 94.14 %, lowest among all.

TABLE II
PREDICTION OF MALIGNANT AND BENIGN TESTING DATASET, CONFUSION MATRIX

Algorithm	Malignant	Benign
SVM	113	55
Random forest	110	55
Logistic regression	111	54
Decision Tree	107	54

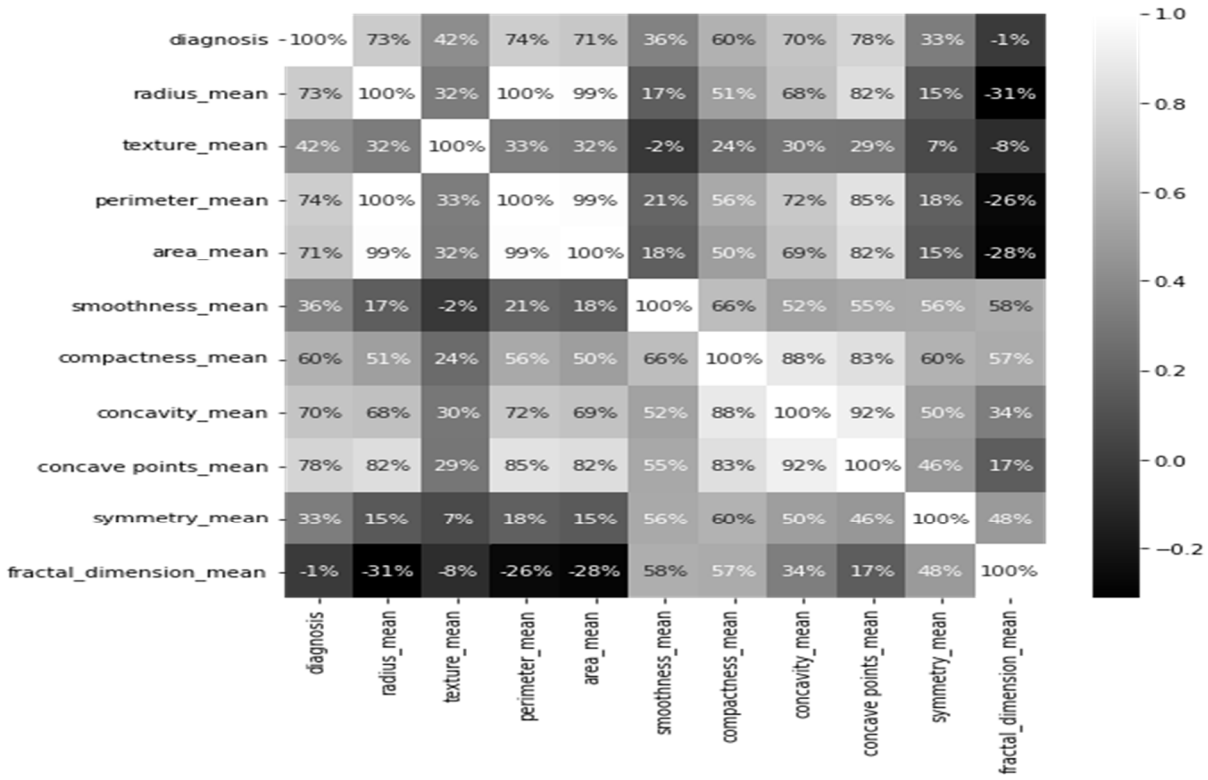

Fig. 2e Correlation Matrix Between Different Features

Table 2. shows that the SVM correctly identify168 of the 171 cases on test data. There are 113 cases that are genuinely cancer, 55 cases that are actually benign, and three cases that were mistakenly diagnosed. With the same datasets, SVM outperformed other algorithms.

TABLE III
CLASSIFIERS PERFORMANCE (IN PERCENTAGE)

Algorithm	Precision	Sensitivity	F Measure	Class
SVM	99	98	99	Malignant
	96	98	97	Benign
Random	99	96	97	Malignant
forest	92	98	95	Benign
Logistic	98	97	97	Malignant
regression	93	96	95	Benign
Decision	98	93	96	Malignant
Tree	87	96	82	Benign

From the table we can analyses that for support vector machine, precision is 99, sensitivity 98, F measure 99 which is higher as compare to any other classifier. In comparison to other model Support Vector Machine model always shows best result in performance for two classes malignant and benign in WDBC dataset. We also demonstrate our result with the help of ROC curve to demonstrate the diagnostic ability of a machine learning models. The area under the receiver operating characteristics (AUROC) curve is a graphical plot and used to represent the outcomes of the machine learning model in which maximum outcome is represent by values approach to one.

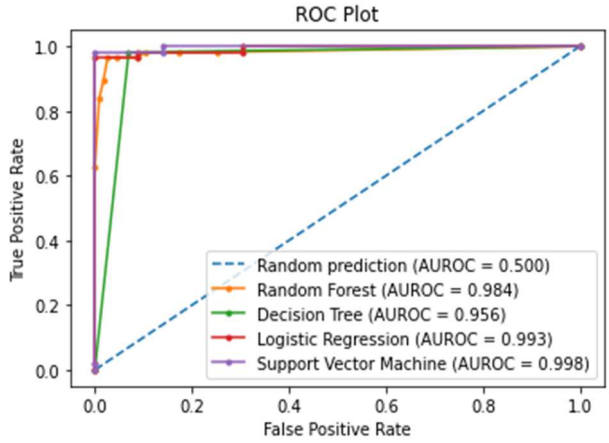

Fig. 3 ROC curve comparison

The receiver operating characteristics graph is used to evaluate the performance of a classification model at all classification thresholds. This curve shows two parameter: i) True positive rate(TPR), ii) False positive rate(FPR).
The TPR, which is also referred as sensitivity, is calculated as the ratio number of true positives and the sum of the number of true positives and the number of false negatives. TPRdefines how good the model is at predicting the positive class when the actual outcome is positive.

$$TPR\ (Sensitivity) = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \quad (1)$$

The FPR also referredas the inverted specificity, is calculated as the ratio of number of false positives and the sum of the number of false positives and the number of true

negatives.It define how often a positive class is predicted when the actual outcome is negative.

$$FPR = \frac{False\ Positives}{(False\ Positives + True\ Negatives)} \quad (2)$$

ROC curve shows the true positive rate, on the y-axis and is plotted against the false positive rate represent on the x axis [10]. The values of both x and y-axis are spread from Zero to One. The graph is generated from measuring true positive rate and false positive rate for each feasible classifier threshold value[11]. Figure 1 shows the receiver operating curves of each machine-learning algorithm. AUROC, which is,stands for "Area under the ROC Curve." measures the entire two-dimensional area cover by ROC curve. The area under this curve is calculated in which larger area covered by respective classifier represent the better the performance. In our work as shown in above fig. 1 the Support vector machine (SVM) have greatest AUC (Area Under the ROC) Curve score of 99.8%, followed by Logistic regression, which had an AUC of 99.3%. Also it is shown in Fig. 1, that the Decision tree has the lowest AUC of 95.6 percent.

IV. Conclusion

In this work we observed WDBC dataset and use various classifiers classify malignant and benign tumors. The results are comparing, calculate and evaluate based on confusion matrix, precision, sensitivity and accuracy[12]. The experiment is set up in Python, using NumPy library, pandas, SciKit learn, Matplot Lib. After simulating the program and comparing the different models it is found that SVM has demonstrated its accuracy and achieve best performance in prediction of breast cancer. SVM achieved a highest accuracy of 98.24% with AUC 0.998, 99 precession in Malignant and 96 in Benign, which is better than all other algorithm. All the results are obtained by using WDBC dataset, same algorithm and model can be use for other dataset in future to get better result. In future, we will work on latest dataset with more disease classes to obtain higher accuracy with another machine learning.

References

[1] International Agency for Research on Cancer, Press release December 2020, https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action.

[2] L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A.R. Razavi, “Using three machine learning techniques for predicting breast cancer recurrence,” (2013), J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124

[3] S. Nayak, D. Gope “Comparison of supervised learning algorithms for RF-based breast cancer detection,”, 2017 Computing and Electromagnetics International Workshop (CEM), Barcelona (2017)

[4] M. H. Memon, J. P. Li, A. U. Haq, M. H. Memon, and W. Zhou, ‘‘Breast cancer detection in the IOT health environment using modified recursive feature selection,’’ Wireless Commun. Mobile Comput., vol. 2019, pp. 1–19, Nov. 2019

[5] S. Alghunaim and H. H. Al-Baity, ‘‘On the scalability of machinelearning algorithms for breast cancer prediction in big data context,’’ IEEE Access, vol. 7, pp. 91535–91546, 2019.

[6] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

[7] B. M. Gayathri and C. P. Sumathi, “Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer,” 2016 IEEE Int. Conf. on Computational Intelligence and Computing Research (ICCIC), pp 1-5, IEEE, 2016

[8] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic) Accessed [August] [2021].

[9] Poonam Pandey and Radhika Prabhakar, “An analysis of machine learning techniques (J48 & Ada Boost) - for classification,” 2016 1st India Int. Conf. on Information Processing (IICIP), PP 1- 6, IEEE, 2016, India.

[10] L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proenca, “Digital signature of network segment for healthcare environments support,” Irbm, vol. 35, no. 6, pp. 299-309, 2014.

[11] Dana Bazazeh and Raed Shubair. “Comparative study of machine learning algorithms for breast cancer detection and diagnosis,” 2016 5th Int. Conf. on Electronic Devices, Systems and Applications (ICEDSA), 6-8 December 2016, Ras Al Khaimah, UAE.

[12] Zahra Nematzadeh, Roliana Ibrahim and Ali Selamat, “Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques,” Proc. in 2015 10th Asian Control Conf. (ASCC), pp 1-6, IEEE, 2015.