Journal of Experimental Biology and Agricultural Sciences, April - 2022; Volume – 10(2) page 440 – 450

Journal of Experimental Biology and Agricultural Sciences

http://www.jebas.org

ISSN No. 2320 – 8694

Machine learning for the classification of breast cancer tumor: a comparative analysis

Madhumita Pal[1], Smita Parija[1,*], Ganapati Panda[1], Kuldeep Dhama[2], Ranjan K. Mohapatra[3,*]

[1]Electronics and communication Engineering, C. V. Raman Global University, Bidyanagar, Mahura, Janla, Bhubaneswar, Odisha 752054, India
[2]Division of Pathology, ICAR-Indian Veterinary Research Institute, Izatnagar, Bareilly-243122, Uttar Pradesh, India
[3]Department of Chemistry, Government College of Engineering, Keonjhar, Odisha 758002, India

Received – April 01, 2022; Revision – April 20, 2022; Accepted – April 29, 2022
Available Online – April 30, 2022

DOI: http://dx.doi.org/10.18006/2022.10(2).440.450

KEYWORDS

Breast cancer

Multilayer perceptron

K-NN

MLP

Random forest

ABSTRACT

The detection and diagnosis of Breast cancer at an early stage is a challenging task. With the increase in emerging technologies such as data mining tools, along with machine learning algorithms, new prospects in the medical field for automatic diagnosis have been developed, with which the prediction of a disease at an early stage is possible. Early detection of the disease may increase the survival rate of patients. The main purpose of the study was to predict breast cancer disease as benign or malignant by using supervised machine learning algorithms such as the K-nearest neighbor (K-NN), multilayer perceptron (MLP), and random forest (RF) and to compare their performance in terms of the accuracy, precision, F1 score, support, and AUC. The experimental results demonstrated that the MLP achieved a high prediction accuracy of 99.4%, followed by random forest (96.4%) and K-NN (76.3%). The diagnosis rates of the MLP, random forest and K-NN were 99.9%, 99.6%, and 73%, respectively. The study provides a clear idea of the accomplishments of classification algorithms in terms of their prediction ability, which can aid healthcare professionals in diagnosing chronic breast cancer efficiently.

* Corresponding author

E-mail: smita.parija@gmail.com (Dr. Smita Parija);
ranjank_mohapatra@yahoo.com (Dr. Ranjan K. Mohapatra)

Peer review under responsibility of Journal of Experimental Biology and Agricultural Sciences.

Production and Hosting by Horizon Publisher India [HPI] (http://www.horizonpublisherindia.in/).

1 Introduction

The World Cancer Research Fund recorded two million new cases of breast cancer in 2018, which resulted in approximately 626,679 deaths. Breast cancer, which is also known as "breast carcinoma," is the excessive growth of epithelial cells in the lining ducts and lobules of the breast. It is the second leading cause of death in women. As opposed to conducting several tests for the diagnosis of the disease, as recommended by an oncologist, machine learning algorithms can provide a better solution for the automatic prediction of breast cancer. Machine learning is a data mining technique that is used to design an automation system without human interference. The goal of supervised learning is to predict or classify an output based on prior information. Labeled data are required for training the machine learning model. The objective of unsupervised machine learning methods is to detect clusters within a heterogeneous data structure, without a labeled dataset. Thus, it can predict the output without a supervisor (Jiang et al. 2020). Medical applications are another significant area of machine learning. As healthcare applications contain large amounts of data, it is challenging to handle these data. Such data can be managed efficiently by using machine learning techniques. This can aid in the early detection of disease, reduce the cost of medicines, and increase the patient survival rate. Machine learning provides an automatic system approach that helps to diagnose the disease at an early stage and appropriate treatment can subsequently be provided to the patient at the correct time, which will reduce the death rate of patients suffering from this chronic disease (Sun et al. 2017). The proposed method provides an automated machine learning system for the early prediction of breast cancer and recommendations for proper treatment by the oncologist to the patient at the correct time.

Asri et al. (2016) presented a breast cancer analysis model using various machine learning techniques, namely the support vector machine (SVM), naïve Bayes, decision tree (C4.5), and K-nearest neighbor (K-NN). The experiment was performed on the Wisconsin Breast Cancer (original) datasets and simulated using the WEKA data mining tool. The performance analysis demonstrated that the SVM achieved the maximum accuracy (97.13%) with the lowest error rate among all of the algorithms. Muktevi (2020) proposed a breast cancer prediction system with the implementation of a machine learning algorithm. In this study, the author applied an SVM, naïve Bayes, random forest, and logistic regression to the Breast Cancer dataset to predict an accurate model. The experimental results showed that the random forest provided better results on the Breast Cancer dataset compared to other datasets. Moreover, the accuracy, precision, recall, and F1score were investigated and the components of the confusion matrix were discussed.

Rana et al. (2015) proposed a data mining model for the classification of tumors as malignant or benign. A model with the implementation of machine learning algorithms such as SVM, logistic regression, and K-NN was presented. In terms of accuracy, the false positive rate, sensitivity, and specificity of all algorithms were compared. In the study, the SVM exhibited the best prediction accuracy of 92.4%. Moreover, Islam et al. (2017) proposed a system using a 10-fold cross-validation method for predicting breast cancer. An SVM and K-NN were used for the detection of breast cancer. The experimental results demonstrated that the accuracies of the SVM and K-NN were 98.57% and 97.14%, respectively. Moreover, the SVM was better at predicting the disease in terms of accuracy.

Singh (2019) determined a biomarker for the prediction of breast cancer by using various machine learning algorithms. The experimental results showed that among nine attributes, glucose, age, and resistance were effective biomarkers for breast cancer prediction. Using these features for classification, the K-NN yielded a maximum classification accuracy of 92.11%, followed by the Gaussian SVM with a classification accuracy of 83.68%. Furthermore, García-Laencina et al. (2015) proposed a prediction model for five-year breast cancer survivability without imputation in 2015. The study showed that K-NN achieved the highest prediction accuracy of more than 81% and receiver operating characteristics of more than 0.78%. Wu et al. (2019) proposed a white-box machine learning model approach to predict the molecular subtypes of breast cancer based on BI-RADS features using MRI and mammography images. A 10-fold cross-validation method was applied to compute the performance (positive predictive value, accuracy, F1 score, and sensitivity) of the decision tree model. Moreover, Islam and coworkers have compared the machine learning techniques for the prediction of breast cancer by using the dataset retrieved from the UCI repository (Islam et al. 2020).

1.1 Breast cancer

Breast cancer is the most dangerous chronic disease that commonly occurs in women, normally at the age of above 40 years. Cancer does not cause any pain until it has spread to adjacent tissues. It begins as *in situ* carcinomas such as ductal carcinoma and lobular carcinoma (Sun et al. 2017). These occur on top of the ribs and pectoral muscles, and are divided into three main parts *viz*., (i) *Glandular tissue* - this tissue creates milk consisting of 15 to 20 lobules. Inside each lobule, grape-like structures known as alveoli are present, (ii) *Stroma* - this consists of adipose/fat tissue and contains the majority of the breast tissue. It contains Cooper's ligaments attached to the skin and pectoralis muscles, and (iii) *Lymphatic vessels* - these consist of drain lymph cells, which contain fluid that drains cellular waste and white blood cells. They mainly drain into the lymph nodes in the axilla/armpit. When the menopausal estrogenic hormones discharged by the ovaries stop, the alveolar cells die and the breast

tissue is reinstated. During the menstrual cycle, the production of estrogen and progesterone from the ovaries increases, and after menstruation, the flow is reduced.

Each menstrual cycle causes the alveolar cells to undergo demarcation and apoptosis. Every time a cell bisects, there is a change in the genetic mutation and this mutation causes tumor formation. Therefore, there is an increase in the occurrence of breast cancer with an increase in the number of menstrual cycles during the initial and late stages of menopause. Furthermore, medications containing estrogenic agents may maximize the possibility of breast cancer. Ionizing radiation such as CT scans and MRI may also increase the occurrence of breast cancer. Pregnancy at an early age and breastfeeding for a long period can decrease the risk of breast cancer.

1.2 Symptoms

The most common symptoms of breast cancer are (i) hard, painless lump or swelling which is normally found in the upper and outer part of the breast, (ii) swelling under the armpit, (iii) breast immovable, (iv) notching of skin, (v) fibrosis of Lactiferous ducts and suspensory ligaments and (vi) paget diseases

1.3 Diagnosis

Early-stage diagnosis of breast cancer is a difficult task, some common methods of breast cancer diagnosis are (i) feeling breast lump, (ii) with MAMMOGRAPHY, (iii) Imaging using ultrasound and MRI, and (iv) biopsy of swelling.

1.4 Treatment

Treatment of breast cancer depends on the type and stage of cancer, some commonly available treatment methods are (i) surgery, (ii) radiation therapy, (iii) chemotherapy, and (iv) hormonal therapy.

2 Materials and Method

To implement the algorithms of this research, we used the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the Kaggle site (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data). The K-NN, MLP, and RF algorithms were implemented on the dataset to obtain the classification results (Table 1). The code for each algorithm was written in Python and executed in the Jupyter Notebook. Several supervised data prediction techniques were used in the model and their performance was compared to provide a better quality of service for the healthcare system.

Figure 1 presents the architecture of the data mining model for the implementation of machine learning algorithms. It provides an understanding of how to load the dataset and how to extract the features from the dataset by using different stages. The first stage is the selection of a breast cancer dataset online, followed by the pre-processing and transformation of variables in the second stage, and the third stage is the implementation of various ML models. Finally, we have evaluated our model for the prediction of benign or malignant breast cancer using different metrics.

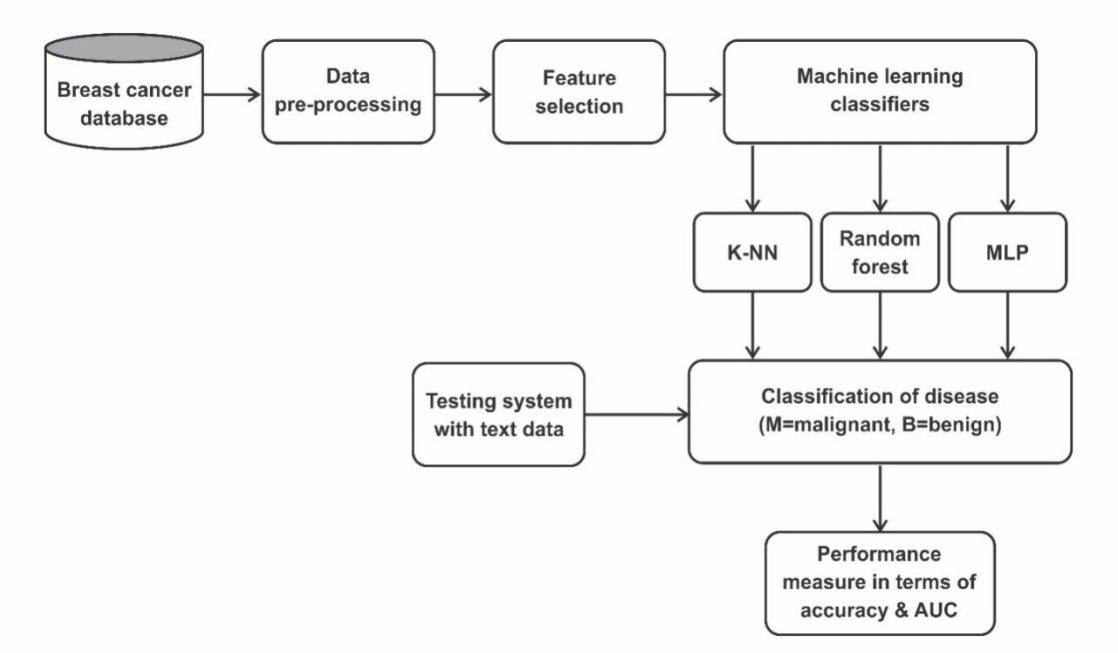

Figure 1 Architecture of data mining model for breast cancer classification

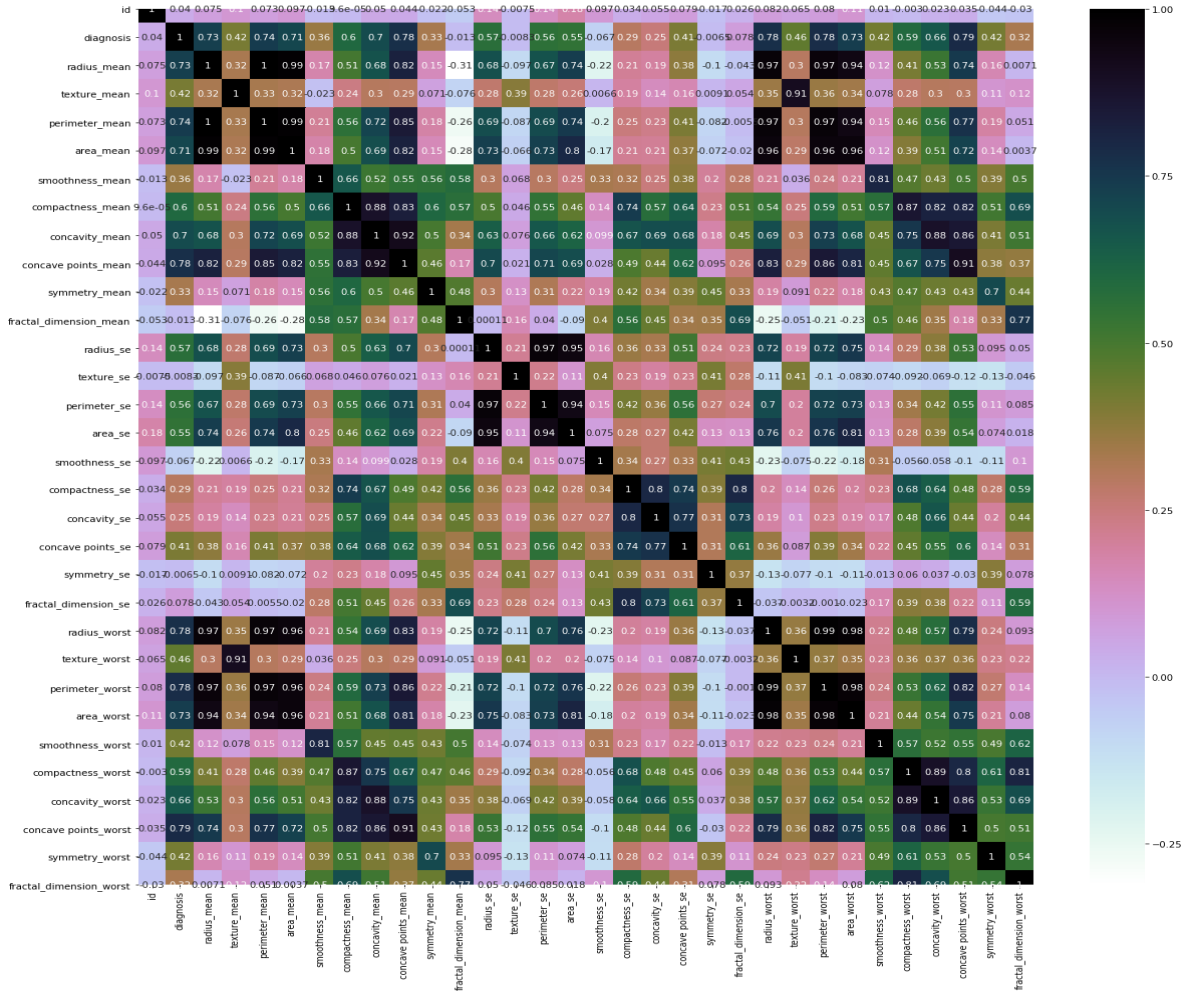

Figure 2 Correlation matrix of the given data set

2.1 Pre-processing and transformation

The breast cancer dataset was converted into the CSV format from the Excel file. Thereafter, by using explorative data analysis, we filtered the data by checking the null values; missing values existed in the dataset. We subsequently correlated the features using a correlation matrix, as illustrated in Figure 2.

2.2 Performance evaluation

To measure the performance of the various machine learning models, we used different metrics, including precision, recall, F1 score, and support. The formulae for calculating the metrics are presented in Table 1. We used confusion matrices for the different algorithms to evaluate the metrics.

Table 1 Formulae for calculation of different metrics

Metrics	Formula
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
F1_score	$\frac{2 * precision * recall}{precision + recall}$
Support	TN + FP, TP + FN

2.3 K-nearest neighbor (K-NN)

K-NN is a supervised ML algorithm that is used to classify an unknown object based on the nearest neighbor data point concept. By using different distance metric concepts, such as the Euclidean distance and Manhattan distance, the nearest neighbor data point can be determined. The K-NN algorithm is easy to implement, but it is inefficient for a large dimensional dataset. It is a nonparametric model that is used for solving classification

problems (Forsyth et al. 2018). The Euclidean distance for two points in Euclidean space is

$$U(a,b) = \sqrt{\sum_{j=1}^{n}(bj - aj)2} \quad (1)$$

Where, a, b are two points in n-space, and b_j and a_j are Euclidean vectors.

2.4 Random forest (RF)

RF is a supervised ML algorithm that is used for classification and regression. However, it is inefficient for regression problems in terms of accuracy (Verikas et al. 2011). It is a type of ensemble classifier that uses a decision tree algorithm in a randomization process. It consists of different decision trees of various sizes and shapes. In this case, random means the random sampling of the training tree while building the decision tree, and a random subset of input features is obtained when splitting at the node (Figure 3). One aspect that restricts decision trees as an optimal tool for predictive analysis is their inaccuracy. Decision trees cannot provide good classification results. RF combines the integrity of decision trees with docility, which causes a large accuracy rectification.

2.5 Multilayer perceptron (MLP)

The MLP is part of the feed-forward ANN technique. The MLP consists of three layers: the input layer, hidden layer, and output layer. Figure 4 depicts an MLP network with three input layers, four hidden layers, and one output layer. In this neural network, the information flows in the forward direction but not in the backward direction. As it is part of the feed-forward neural network, information is passed from input to output through the hidden layer in the forward direction. It uses different nonlinear activation functions at the hidden layer to inject nonlinear mapping from the input to the hidden layer and linear mapping from the hidden layer to the output layer. This method provides better accuracy for a large dimensional dataset. To train the model, a back-propagation learning algorithm is used to minimize the error. The network minimizes the sum of the square error by updating the weight at each layer in the backward pass (Costa et al. 2013).

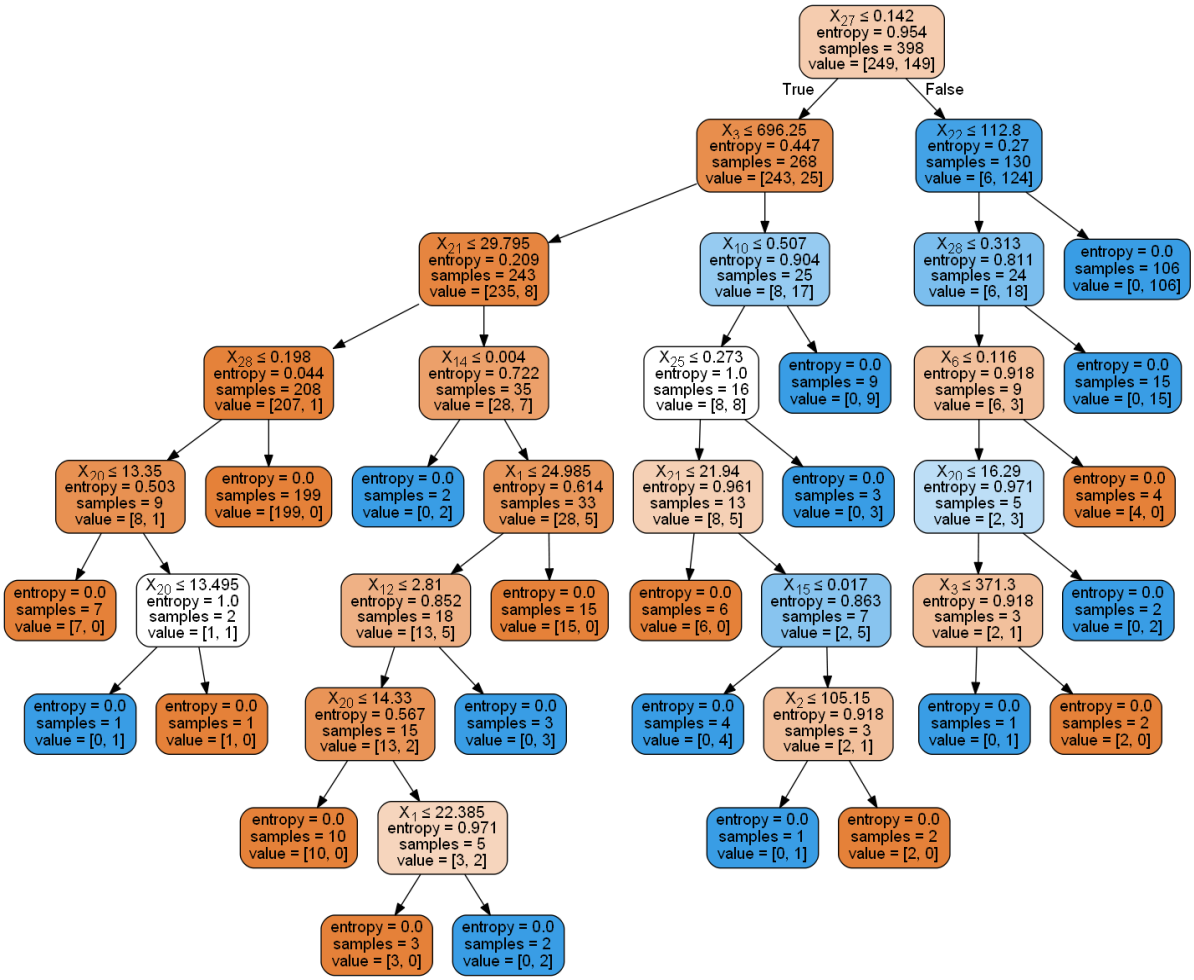

Figure 3 The obtained decision tree

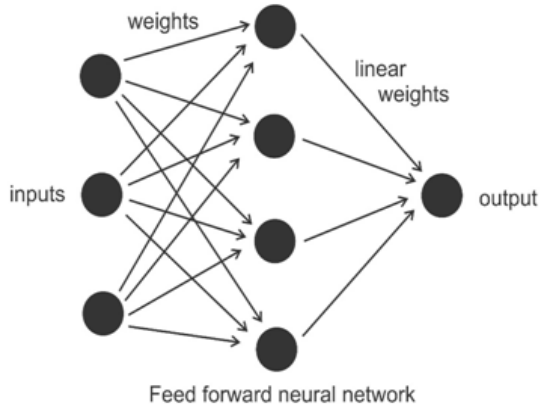

Figure 4 Diagram of feed-forward neural network

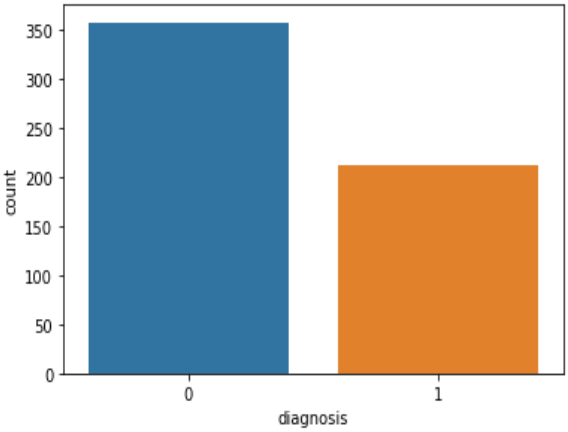

Figure 5 Count plot for classification of breast cancer patients

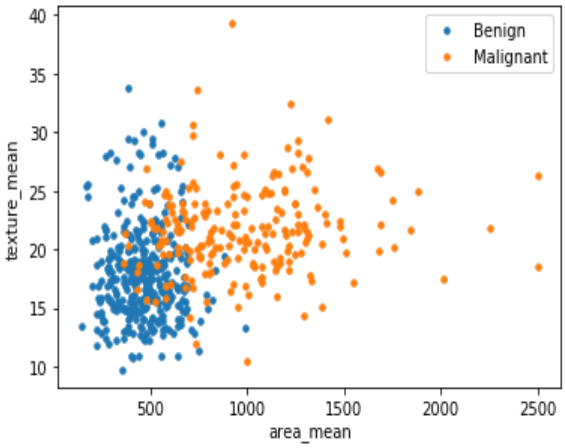

Figure 6 Data visualization plot for breast cancer features

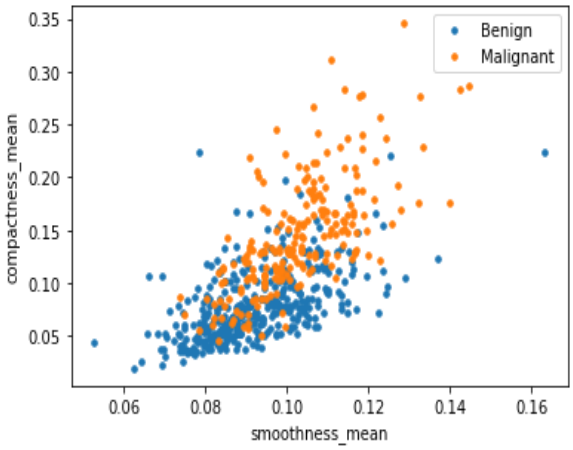

Figure 7 Correlation plot for breast cancer features

3 Results and Discussion

The dataset used in this study was the WDBC dataset, accessed from the Kaggle site. It consists of 569 samples and 32 attributes. The dataset was filtered using an explorative data analysis method. The attribute diagnosis indicates whether the breast cancer tumor is benign or malignant (M = malignant (1) and the tumor is cancerous; B = benign (0) and the tumor is noncancerous). Among the 569 samples, the diagnosis feature indicated that 357 samples were benign and 212 were malignant tumors as shown in Figure 5.

Thereafter, the K-NN, MLP, and random forest algorithms were implemented on the dataset using Python code, following which all of the codes were executed in the Jupyter Notebook. The experimental results demonstrated that the prediction accuracy obtained for classification using the K-NN algorithm was 76.3%, that when using the random forest algorithm was 96.4%, and that when using the MLP was 99.4%. In this section, the prediction results of the different machine learning algorithms are discussed. We have employed 10-fold cross-validation for splitting the dataset into numerous numbers of training and testing datasets for representing the prediction ability of the model more accurately. A total of 80% of the data samples were used for training the model and 20% of the data were used for testing the model. Subsequently, by pre-processing the data, we analyzed the data in terms of accuracy.

A correlation plot gives the correlation between different variables present in the dataset. It is mainly used for feature selection. If two features have a strong correlation value then one feature can be dropped. In this way, the number of input features of the given dataset can be reduced and the performance of the predictive model is increased. The data visualization plots (Figure 6 and Figure 7) show that with the increase in the area mean and compactness means chances of breast cancer tumors being malignant are more. Malignant observations are located in the right left corner and benign observations are located in the left lower corner. The blue color indicates the benign tumor and the orange color indicates the malignant tumor features. The mean of texture and smoothness does not show any impact on overdiagnosis.

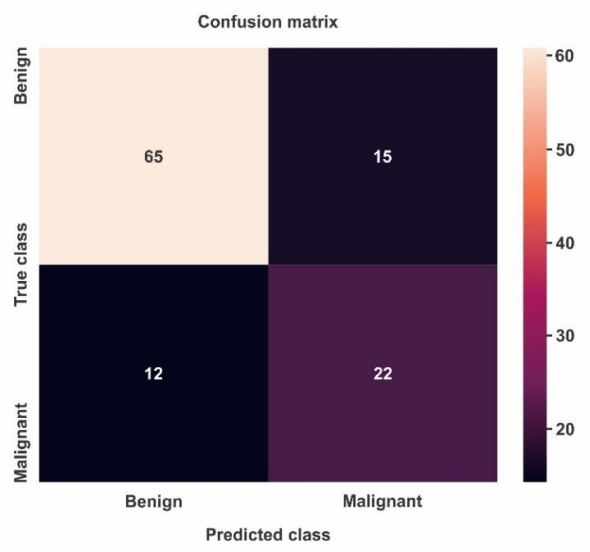

Figure 8 Confusion matrix of K-NN

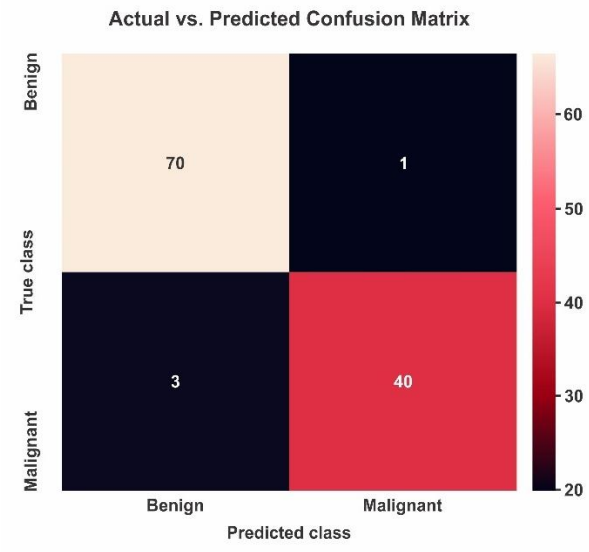

Figure 9 Confusion matrix of RF

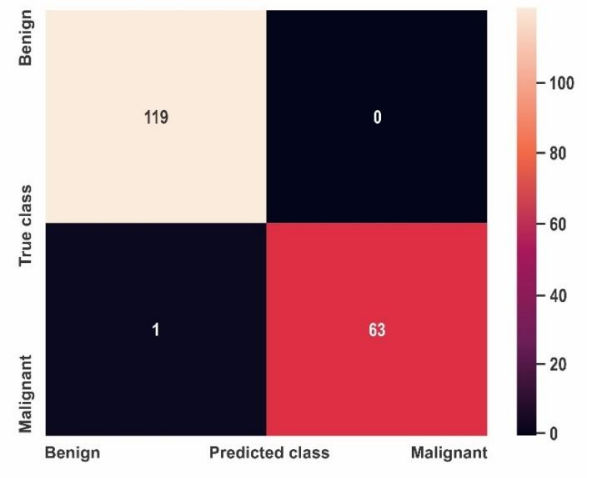

Figure 10 Confusion matrix of MLP

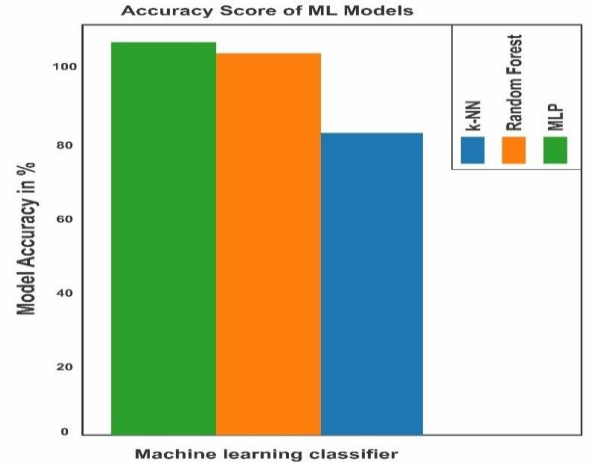

Figure 11 Accuracy comparisons of ML algorithms

3.1 Confusion matrix

In this research, different metrics such as accuracy, precision, recall, F1 score, and support were used for the accurate measurement. A confusion matrix is generally used for binary classification problems to measure the accuracy of the model. It consists of four parts *viz*., TP= true positive, TN= true negative, FP= false positive, and FN= false negative (here FN indicates that the predicted output is zero but the actual output is one; FP indicates that the predicted output is one but the actual output is zero; TP indicates that the predicted and actual outputs are equal; that is, one; and TN indicates that the predicted and actual outputs are equal; that is, zero).

When applying K-NN to the testing data from the confusion matrix (Figure 8), we obtained 65 TPs, namely patients with breast cancer that were correctly classified, and 22 TN patients without breast cancer that were correctly classified. The algorithm misclassified 12 patients with breast cancer by indicating that they did not (FN), and 15 patients who did not have breast cancer by indicating that they did (FP).

When applying the random forest algorithm to the testing data from the confusion matrix (Figure 9), we obtained 70 TPs, namely patients with breast cancer that were correctly classified, and 40 TN patients without breast cancer that were correctly classified. The algorithm misclassified three patients that did have breast cancer by indicating that they did not (FN) and one patient that did not have breast cancer by indicating that they did (FP). Similarly, when applying the MLP algorithm to the testing data from the confusion matrix (Figure 10), we obtained 119 TPs, namely patients with breast cancer that were correctly classified, and 63

Table 2 Performances measurements of K-NN, RF, and MLP algorithms

K-NN	Precision	Recall	F1 score	Support
0	0.84	0.81	0.83	80
1	0.59	0.65	0.62	34
Accuracy	-	-	0.76	114
Macro avg	0.72	0.73	0.72	114
Weighted avg	0.77	0.76	0.77	114
Random forest				
0	0.99	0.96	0.97	71
1	0.93	0.98	0.95	43
Accuracy	-	-	0.96	114
Macro avg	0.96	0.97	0.96	114
Weighted avg	0.97	0.96	0.96	114
MLP				
0	0.99	1.0	1.0	119
1	1.0	0.98	0.99	64
Accuracy	-	-	0.99	183
Macro avg	1.0	0.99	0.99	183
Weighted avg	0.99	0.99	0.99	183

TN patients without breast cancer disease were correctly classified. The algorithm misclassified one patient that had breast cancer by indicating that FN and the FP rate for the MLP algorithm were zero.

We can compare the RF confusion matrix to that of K-NN. It can be observed that K-NN was worse than RF in predicting patients with breast cancer (22 *vs*. 40) and worse at predicting patients without breast cancer (65 *vs*. 70). According to our study, RF should be selected between the two algorithms as it provides outstanding performance compared to K-NN for classifying breast cancer. Finally, we used the MLP on the testing dataset to create a confusion matrix. Moreover, RF was worse than the MLP in predicting patients with breast cancer (40 *vs*. 63) and worse in predicting patients without breast cancer (70 *vs*. 119). Therefore, among these three algorithms, the MLP provided an outstanding prediction accuracy of 99.4%. Moreover, our model gives excellent performance with prediction accuracy of 99.4% and AUC of 99.8% respectively using MLP algorithm (Asri et al. 2016).

According to Table 2, the precision, recall, F1 score, and support values obtained when using the K-NN algorithm were 0.84, 0.81, 0.83, and 80, respectively for the benign class and 0.59, 0.65, 0.62, and 34, respectively for the malignant class. The accuracy values in terms of the F1 score and support were 0.76 and 114, respectively, and the macro average values in terms of the precision, recall, F1 score, and support were 0.72, 0.73, 0.72, and 114, respectively. The weighted average values in terms of the precision, recall, F1 score, and support were 0.77, 0.76, 0.77, and 114, respectively.

When implementing the RF algorithm, the precision, recall, F1 score, and support values obtained were 0.99, 0.96, 0.97, and 71, respectively, for the benign class and 0.93, 0.98, 0.95, and 43, respectively, for the malignant class. The accuracy values for a random forest in terms of the F1 score and support were 0.96 and 114, respectively. The macro average values in terms of the precision, recall, F1 score, and support were 0.96, 0.97, 0.96, and 114, respectively. The weighted average values in terms of the precision, recall, F1 score, and support were 0.97, 0.96, 0.96, and 114, respectively.

With the MLP algorithm, the precision, recall, F1score, and support values were 0.99, 1.0, 1.0, and 119, respectively, for the benign class and 1.0, 0.98, 0.99, and 64, respectively, for the malignant class. The accuracy values in terms of the F1 score and the support were 0.99 and 183, respectively. The macro average values in terms of the precision, recall, F1 score, and support were 1.0, 0.99, 0.99, and 183, respectively. The weighted averages in terms of precision, recall, F1 score, and support were 0.99, 0.99, 0.99, and 183, respectively.

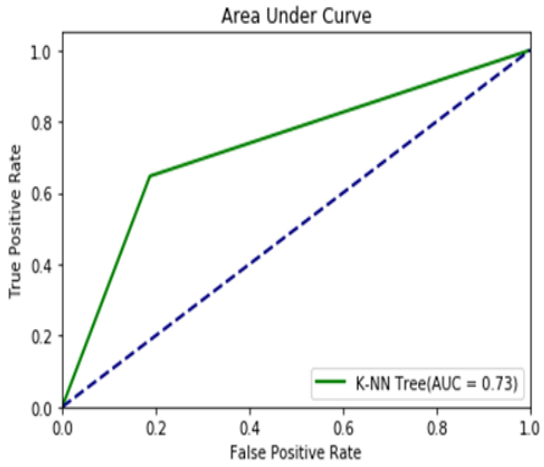

Figure 12 AUC of K-NN algorithm

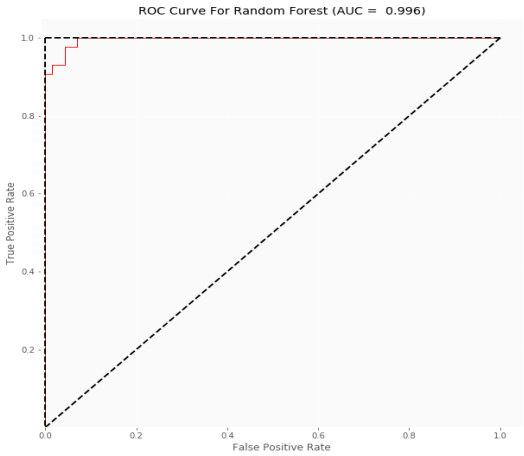

Figure 13 AUC of RF algorithm

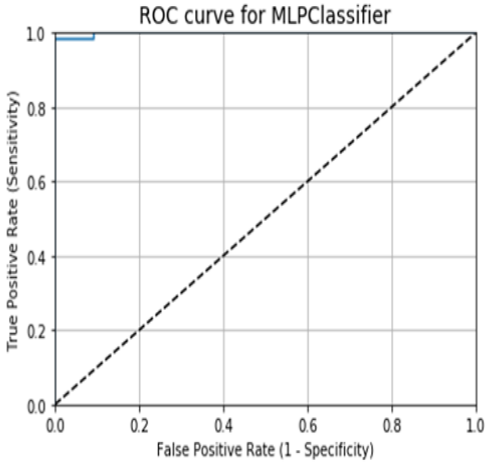

Figure 14 AUC of MLP algorithm

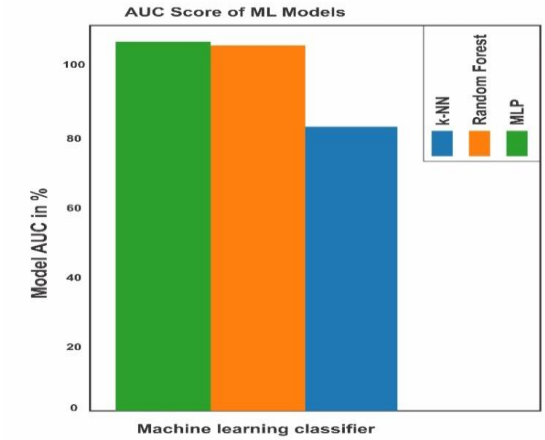

Figure 15 Comparison of AUC scores for different ML algorithms

It can be observed from Figure 11 that K-NN is a lazy learner and is not very active during the training process, unlike the other classifiers used to build the models. According to the graph, the accuracy obtained by the MLP (99.4%) was better than the accuracy obtained by the random forest (96.4%) and K-NN (76.3%) algorithms. Moreover, the MLP yielded the highest value of correctly classified instances and a lower value of incorrectly classified instances than the other classifiers.

The AUC is plotted with the TP rate (y-axis) against the FP rate (x-axis); TP rate =TP/TP+FN; FP rate =FP/TN+FP

A model provides excellent classification if the AUC value is nearly 1. In this case, it indicates how accurately the model predicts whether a patient suffers from breast cancer. According to Figure 12, the AUC value for the K-NN algorithm was 0.73, which means that there is a 73% chance that our model can accurately distinguish between the benign and malignant classes. Similarly, according to figure 13 for the RF algorithm, there is a 99.6% chance that the model can accurately predict breast cancer. It can be observed from Figure 14 that there is a 99.9% chance that the model can accurately classify the benign and malignant classes with the MLP algorithm. The highest diagnosis rate was obtained with the MLP algorithm using the AUC parameter, as illustrated in Figure 15.

The FP rate was the lowest (zero) when using the MLP classifiers. From the results, we may understand why the MLP has outperformed the other classifiers. According to the experimental results, the highest accuracy value (99.4%) was achieved by using the MLP algorithm in extracting the features of tumors (benign or malignant). It is observed that the performance of the MLP algorithm was better than those of the other classifiers in the accuracy, sensitivity, specificity, and precision. Moreover, a comparative analysis for the prediction of breast cancer disease is presented in Table 3.

Table 3 Performance comparison of proposed work with published works

Authors	Technique	Accuracy
Islam et al. (2020)	ANN	98.57
Latchoumi and Parthiban (2017)	WPSO-SSVM	98.42
Chaurasia et al. (2018)	Naive Bayes	97.36
Sakri et al. (2018)	Naïve Bayes	81.3
	Rep tree	80
	K-NNs	75
Banu and Subramanian (2018)	Bayes belief network	91.7
	Tree augmented naïve bayes	94.11
	Boosted augmented naïve bayes	91.7
Kaur et al. (2019)	RF	95.71
	MLP	96.42
Proposed work	RF	96.4
	MLP	99.4

Conclusions

This study has provided a clear idea of the accomplishment of classification algorithms in terms of their prediction ability, which can aid healthcare professionals to diagnose chronic disease (breast cancer) efficiently. Our goal was to achieve high accuracy for breast cancer classification by using a supervised machine learning algorithm. Our proposed model yielded the highest accuracy of 99.4% when using the MLP algorithm, followed by the random forest algorithm with 96.4% and the K-NN algorithm with 76.4%. The highest diagnosis rate was 99.8% achieved by the MLP algorithm, followed by random forest (99.6%) and K-NN (73%) when using the AUC parameter. We hope that our work will be very helpful in the determination of reliable biomarkers for the detection of breast cancer tumors with a larger dataset.

Acknowledgments

All authors acknowledge their respective Institutes and Universities.

Funding

No funding received.

Conflicts of interest

There are no conflicts to declare.

Data Availability

Not applicable

References

Asri, H., Mousannif, H., Moatassime, H.A., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016). *Procedia Computer Science*, *83*, 1064 – 1069.

Banu, A.B., & Subramanian, P.T. (2018). Comparison of Bayes classifers for breast cancer classifcation. *Asian Pacific Journal of Cancer Prevention*, *19*(10), 2917–20.

Chaurasia, V., Pal, S., & Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology*, 12(2), 119–26.

Costa, K., Ribeiro, P., Carmargo, A., Rossi, V., et al. (2013). Comparison of the techniques decision tee and MLP for data mining in SPAMs detection in computer networks. *Proceedings of the 3rd international conference on innovative computing Technology*, 344–348.

Forsyth, A.W., Barzilay, R., Hughes, K.S., Lui, D., et al. (2018). Machine Learning Methods to Extract Documentation of Breast Cancer Symptoms from Electronic Health Records. *Journal of Pain and Symptom Management*, *55*(6), 1492-1499.

García-Laencina, P.J., Abreu, P.H., Abreu, M.H., & Afonoso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in Biology and Medicine*, 59, 125–133.

Islam, M., Haque, R., Iqbal, H., Hasan, M., et al. (2020). Breast Cancer Prediction: A Comparative Study Using MachineLearning Techniques. *SN Computer Science*, *1*, 290.

Islam, M.M., Iqbal, H., Haque, M.R., & Hasan, M.K. (2017). Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbours 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, Bangladesh.

Jiang, T., Gradus, J.L., & Rosellini, A.J. (2020). Supervised machine learning: A brief primer. *Behavior Therapy*, *51*(5), 675-687.

Kaur, P., Kumar, R., & Kumar, M. (2019). A healthcare monitoring system using random forest and internet of things (IoT). *Multimedia Tools and Applications*, *78*, 19905–19916.

Latchoumi, T.P., & Parthiban, L. (2017). Abnormality detection using weighed particle swarm optimization and smooth support vector machine. *Biomedical Research*, *28*, 4749–51.

Muktevi, S. (2020). Prediction of Breast Cancer Disease using Machine Learning Algorithms. *International Journal of Innovative Technology and Exploring Engineering*, *9*(4), 2868-2878.

Rana, M., Chandorkar, P., & Dsouza, A. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. *International Journal of Research in Engineering and Technology*, *4*(4), 372-376.

Sakri, S.B., Rashid, N.B.A., & Zain, Z.M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, *6*, 29637–29647.

Singh, B.K. (2019). Determining relevant biomarkers for breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm. *Biocybernetics and Biomedical Engineering*, *39*, 393-409.

Sun, Y.S., Zhao, Z., Yang, Z.N., Xu, F., et al. (2017). Risk Factors and Preventions of Breast Cancer. *International Journal of Biological Sciences*, *13*(11):1387–1397.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forest: a survey and results of new tests. *Pattern Recognition*, *44*(2), 330–349.

Wu, M., Zhong, X., Peng, Q., Xu, M., et al. (2019). Prediction of Molecular Subtypes of Breast Cancer using BI-RADS Features Based on a "White Box" Machine Learning Approach in a Multi-modal Imaging Setting. *European Journal of Radiology*, *114*, 175-184.