SN Computer Science (2020) 1:290
https://doi.org/10.1007/s42979-020-00305-w

ORIGINAL RESEARCH

Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques

Md. Milon Islam[1] · Md. Rezwanul Haque[1] · Hasib Iqbal[1] · Md. Munirul Hasan[2] · Mahmudul Hasan[3] · Muhammad Nomani Kabir[2]

Received: 11 August 2020 / Accepted: 18 August 2020

Abstract

Early detection of disease has become a crucial problem due to rapid population growth in medical research in recent times. With the rapid population growth, the risk of death incurred by breast cancer is rising exponentially. Breast cancer is the second most severe cancer among all of the cancers already unveiled. An automatic disease detection system aids medical staffs in disease diagnosis and offers reliable, effective, and rapid response as well as decreases the risk of death. In this paper, we compare five supervised machine learning techniques named support vector machine (SVM), K-nearest neighbors, random forests, artificial neural networks (ANNs) and logistic regression. The Wisconsin Breast Cancer dataset is obtained from a prominent machine learning database named UCI machine learning database. The performance of the study is measured with respect to accuracy, sensitivity, specificity, precision, negative predictive value, false-negative rate, false-positive rate, F1 score, and Matthews Correlation Coefficient. Additionally, these techniques were appraised on precision–recall area under curve and receiver operating characteristic curve. The results reveal that the ANNs obtained the highest accuracy, precision, and F1 score of 98.57%, 97.82%, and 0.9890, respectively, whereas 97.14%, 95.65%, and 0.9777 accuracy, precision, and F1 score are obtained by SVM, respectively.

Keywords Breast cancer prediction · Cancer dataset · Machine learning · Support vector machine · Random forests · Artificial neural networks · K-nearest neighbors · Logistic regression

Introduction

A significant issue in the field of bioinformatics or medical science [1] is the accurate diagnosis of certain important information. The diagnosis of the disease is an energetic and tricky job in medicine domain. There is a huge amount of medical diagnosis data available in many diagnostic centers, hospitals, and research centers as well as on numerous

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications" guest edited by Bhanu Prakash K N and M. Shivakumar.

✉ Md. Milon Islam
milonislam@cse.kuet.ac.bd

Md. Rezwanul Haque
rezwanh001@gmail.com

Hasib Iqbal
pranto00250@gmail.com

Md. Munirul Hasan
monirul.iiuc@gmail.com

Mahmudul Hasan
mahmudul.hasan@stonybrook.edu
https://sites.google.com/view/shauqi/home

Muhammad Nomani Kabir
nomanikabir@ump.edu.my

[1] Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna 9203, Bangladesh

[2] Faculty of Computing, Universiti Malaysia Pahang, 26300 Gambang, Kuantan, Malaysia

[3] Department of Computer Science, Stony Brook University, Stony Brook, NY 11794-2424, USA

Published online: 01 September 2020

websites. It is hardly necessary to classify them to make the system automated and quick diagnosis of diseases. The disease diagnosis is usually based on the knowledge and skill of the medical planning officer in the medical field. Because of this, there are circumstances of errors, unwanted biases, and also needs a long time for exact diagnosis of disease.

Conferring to the American Cancer Society [2], the ladies are affected by breast cancer in comparison to all other cancers already introduced. Estimation shows that the ladies will be affected with intrusive breast cancer approximately 252,710 and around 63,410 females will be detected within situ breast cancer in the United States in 2017. Men also have a greater chance of breast cancer. An estimation for men is that they will be affected by this cancer approximately 2470 in the United States in 2017. Another estimation shows that about 41,070 persons will die from this cancer in 2017. Recent statistics in the UK reports that 41,000 women are affected by breast cancer every year whereas only 300 men are affected by this disease.

Breast cancer is the leading cancer in females all over the world. Breast cancer is caused due to the abnormal growth of some cells in the breast. Several techniques have been introduced for the correct diagnosis of breast cancer. Breast screening or mammography [3] is a technique to diagnose breast cancer. It is used to check the nipple status of women through X-rays. Generally, it is almost impossible to detect breast cancer at the initial stage due to the small size of the cancer cell seen from outside. It is possible to diagnose cancer at the early stage through mammography, and this test takes just a few minutes.

Ultrasound [4] is a familiar technique for the diagnosis of breast cancer in which the sound wave is sent inside the body to observe the condition inside. A transducer that emits sound waves is positioned on the skin and the echoes of the tissues of the body are captured with the bounce of sound waves. The echoes are transformed into a gray scale, i.e., a binary value which is represented in a computer. Positron emission tomography (PET) [5] imaging by means of F-fluorodeoxyglucose permits doctors to realize the position of a tumor in the human body. It is constructed on the recognition of radiolabeled cancer-specific tracers. Dynamic MRI [6] has developed the detection procedure for breast distortions. The modality predicts the speed of contrast enhancement by increasing the angiogenesis in cancer. Magnetic reasoning imaging associates with metastasis on contrast enhancement in breast cancer-affected people. Elastography [7] is a newly developed technique based on imaging technology. This technique is applicable when breast cancer tissue is more substantial than the adjacent regular parenchyma. The benign and malignant types are differentiated by a color map of probe compression in this approach.

In very recent years, various machine learning [8–11], deep learning [12, 13], and bio-inspired computing [14] techniques are used in several medical prognoses. Though a number of modalities have been demonstrated, none of the modalities are able to provide a correct and consistent result. In mammography, the doctors should read a high volume of imaging data which reduces the accuracy. This procedure is also time-consuming, and in some worse case, detects the disease with the wrong outcome. This paper compares some machine learning techniques to detect the disease from the input features. Five supervised machine learning approaches have been used to diagnose the disease with proper outcome.

The remaining part of the paper is organized as follows. The next section outlines the current review of the state of the art in this field followed by which the methods and materials used for the study are illustrated. The theoretical concept of each machine learning technique is illustrated in the subsequent section. Then the performance measurement parameters are described. The experimental setup and result analysis are investigated before the final section. The final section draws a conclusion.

Related Works

With the evolution of medical research, numerous new systems have been developed for the detection of breast cancer. The research associated with this area is outlined in brief as follows.

Sakri et al. [15] focused on the enhancement of the accuracy value using a feature selection algorithm named as particle swarm optimization (PSO) along with machine learning algorithms K-NNs, Naive Bayes (NB) and reduced error pruning (REP) tree. Their work perspective holds the Saudi Arabian women's breast cancer problem, and according to their report, it is one of the major problems in Saudi Arabia. Their reports suggest that women with age range greater than 46 are the main victim of this malicious disease. Holding this sentiment, authors of [15] implemented five phase-based data analysis techniques on the WBCD dataset. They reported a comparative analysis between classification without feature selection method and classification with a feature selection method. They have acquired 70%, 76.3%, and 66.3% accuracy for NB, RepTree, and K-NNs, respectively. They used Weka tool for their data analysis purpose. With PSO implemented, they have found four features that are best for this classification task. For NB, RepTree, and K-NNs with PSO, they obtained 81.3%, 80%, and 75% accuracy values, respectively. Kapil and Rana [16] proposed a modified decision tree technique as a weight improved decision tree and implemented it on WBCD and another breast cancer dataset which is retrieved from the UCI repository. Using the Chi-square test, they have found that they have ranked each feature and kept the relevant features for this classification task. For the WBCD dataset, their proposed technique acquired

approximately 99% accuracy, while for the breast cancer dataset, it acquired approximately 85–90% accuracy.

Yue et al. [17] mainly demonstrated comprehensive reviews on SVM, K-NNs, ANNs, and Decision Tree techniques in the application of predicting breast cancer on benchmark Wisconsin Breast Cancer Diagnosis (WBCD) dataset. According to the authors, deep belief networks (DBNs) approach with ANN architecture (DBNs-ANNs) has given the more accurate result. This architecture obtained 99.68% accuracy, whereas for the SVM method, the two-step clustering algorithm alongside the SVM technique has achieved 99.10% classification accuracy. They also reviewed the ensemble technique where SVM, Naive Bayes, and J48 were implemented using the voting technique. The ensemble method acquired 97.13% accuracy. Banu and Subramanian [18] have emphasized Naive Bayes techniques on breast cancer prediction and described a comparison study on Tree Augmented Naive Bayes (TAN), Boosted Augmented Naive Bayes (BAN) and Bayes Belief Network (BBN). They used SAS-EM (Statistical Analytical Software Enterprise Miner) for the implementation of the models. The same popular WBCD dataset is used in their work. According to their findings with the help of gradient boosting 91.7%, 91.7%, and 94.11% accuracy have been achieved for BBN, BAN, and TAN, respectively. Hence, their research suggests that TAN is the best classifier among Naive Bayes techniques for this dataset. Chaurasia et al. [19] implemented Naive Bayes, RBF network, and J48 Decision Tree techniques on WBCD dataset. For their purpose of research, they used the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.9 as a tool of analysis. For Naive Bayes, they obtained 97.36% accuracy which is greater than 96.77% and 93.41% accuracy values resulted from the RBF network and J48 Decision Tree, respectively.

Azar et al. [20] introduced a method for the prediction of breast cancer using the variants of decision tree. The modalities used in this technique are the single decision tree (SDT), boosted decision tree (BDT), and decision tree forest (DTF). The decision is taken by training the data set and after that testing. The outcomes presented that the accuracy obtained by SDT and BDT is 97.07% and 98.83%, respectively, in the training phase which clarifies that BDT performed better than SDT. Decision tree forest obtained an accuracy of 97.51% whereas SDT 95.75% in the testing phase. The dataset was trained by a ten-fold cross-validation fashion. In [21], the authors demonstrated a procedure for the detection of breast cancer. The experiments that have been done for detecting the disease are discussed here using local linear wavelet neural network (LLWNN), and recursive least square (RLS) to enhance the performance of the system. The LLWNN-RLS is providing the maximum values of average Correct Classification Rate (CCR) 0.897 and 0.972 for 2 and 3 predictors, respectively, with a few calculation times. It also provides the lowest value of minimum description length (MDL) and average squared classification error (ASCE) with much lesser time.

Senapati et al. [22] proposed a hybrid system for the detection of breast cancer using KPSO and RLS for RBFNN. The centers, as well as variances of RBFNN, are adjusted using K-particle swarm optimization and adjusted using back-propagation. The classification accuracy achieved by RBFNN-KPSO and RBFNN-extended Kalman filter is 97.85% and 96.4235%, respectively, whereas the coverage time is 8.38 s and 4.27 s, respectively. Hasan et al. [23] developed a mathematical model for the prediction of breast cancer based on the symbolic regression of Multigene Genetic Programming. The ten-fold technique is used to avoid overfitting here. A comparative study is also illustrated. The stopping criteria for the model were generated but the generation level did not reach zero. The highest accuracy obtained by the model is 99.28% with 99.26% precision. A variant of SVM [24] is introduced for the diagnosis of breast cancer. Here six kinds of SVM are explained and used for performance evaluation. The standard SVM results are compared with other types of SVM. Four-fold cross-validation is used for training and testing. The highest accuracy, specificity, and sensitivity achieved by St-SVM are 97.71%, 98.9%, and 97.08%, respectively, in the training phase. The highest accuracy, sensitivity, and specificity obtained by NSVM, LPSVM, SSVM, and LPSVM are 96.5517% 98.2456%, 96.5517%, and 97.1429% individually in the testing phase.

The authors in [25] presented an efficient method for the detection of breast cancer by categorizing the features of breast cancer data utilizing the inductive logic programming technique. A comparison study with a propositional classifier is also drawn. Kappa statistics, F-measure, area under the ROC curve, true-positive rate, etc. are calculated as a performance measure. The system simulated in two platforms named Aleph and WEKA. Jhajharia et al. [26] appraised variants of decision tree algorithms for the diagnosis of breast cancer. The system used the most common decision tree algorithms named CART and C4.5 which are simulated in the WEKA platform using Matlab and Python. The CART implemented in Python achieved the highest accuracy 97.4% and the highest sensitivity 98.9% is obtained in the CART which is implemented in Matlab, and 95.3% specificity is acquired by CART and C4.5, respectively, which are simulated in WEKA. Some of the smart healthcare systems [27, 28] are developed in the IoT environment for the initial treatment of such types of diseases.

Methods and Materials

Data Set Description

The breast cancer dataset was retrieved from the UCI machine learning repository [29]. There are 699 instances in

this dataset, where the cases are either benign or malignant. For such cases, 458 (65.50%) are benign, and 241 (34.50%) are malignant. The class in the dataset is partitioned into 2 or 4, wherever 2 corresponds to the benign case, and 4 corresponds to the malignant case. The attributes consist in the dataset which is in Table 1 excluding sample code number and class level.

The benign cases are identified as a positive class, and the malignant cases are identified as a negative class in our research. Linear correlation refers to straight-line relationships between two variables which can range between − 1 and + 1, where − 1 refers to the perfect negative relationship and + 1 refers to the perfect positive relationship. The relationship among nine attributes of benign and malignant classes is determined that depicts the Pearson correlation amongst positive and negative classes which are shown in Figs. 1 and 2. Figure 1 shows that (x_1, x_9) and (x_5, x_9) and (x_7, x_9) are negatively correlated.

Table 1 Attributes of the dataset

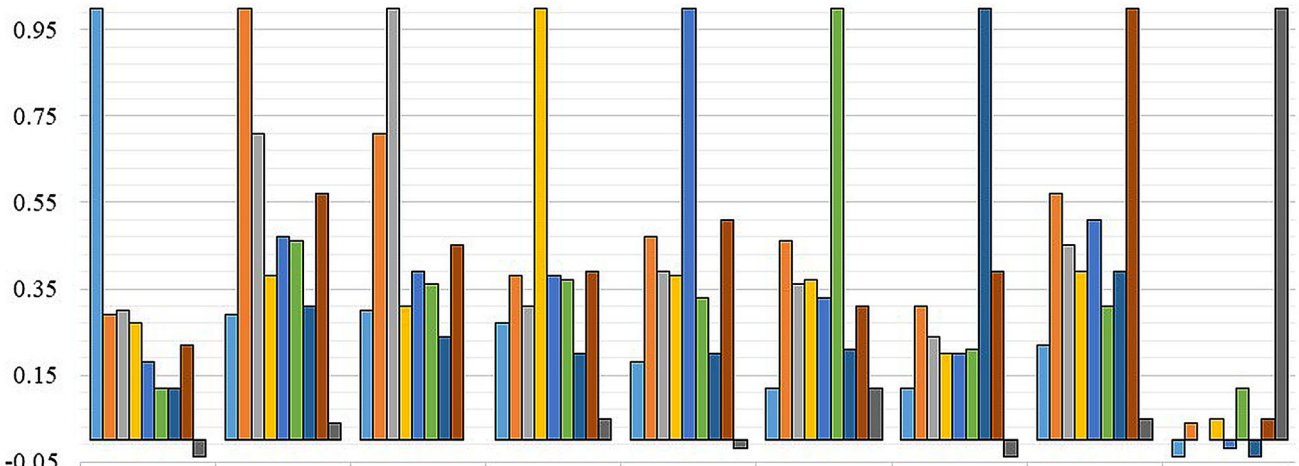
Attributes	Domain	Symbol
Clump thickness	1–10	x_1
Uniformity of cell size	1–10	x_2
Uniformity of cell shape	1–10	x_3
Marginal adhesion	1–10	x_4
Single epithelial cell size	1–10	x_5
Bare nuclei	1–10	x_6
Bland chromatin	1–10	x_7
Normal nuclei	1–10	x_8
Mitoses	1–10	x_9

Data Preprocessing

Data preprocessing is used to complement missing values, identify or remove outliers, and solve self-contradiction. The sample code number is reduced from the dataset as it has no impact on diseases. There are 16 absent values of traits in the dataset. The mean replaces the absent traits for that class. Additionally, the dataset is employed random selection to confirm the proper circulation of the data.

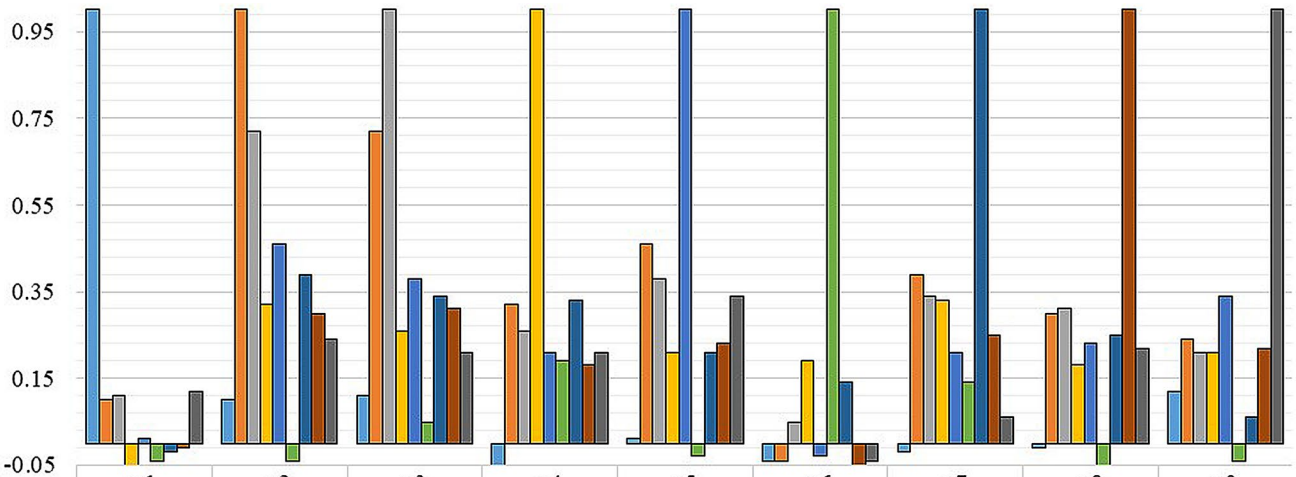
Training and Testing Phase

The training phase extracts the features from the dataset and the testing phase is used to determine how the appropriate model behaves for prediction. The dataset is divided into



	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	1.00	0.29	0.30	0.27	0.18	0.12	0.12	0.22	-0.04
x2	0.29	1.00	0.71	0.38	0.47	0.46	0.31	0.57	0.04
x3	0.30	0.71	1.00	0.31	0.39	0.36	0.24	0.45	0.00
x4	0.27	0.38	0.31	1.00	0.38	0.37	0.20	0.39	0.05
x5	0.18	0.47	0.39	0.38	1.00	0.33	0.20	0.51	-0.02
x6	0.12	0.46	0.36	0.37	0.33	1.00	0.21	0.31	0.12
x7	0.12	0.31	0.24	0.20	0.20	0.21	1.00	0.39	-0.04
x8	0.22	0.57	0.45	0.39	0.51	0.31	0.39	1.00	0.05
x9	-0.04	0.04	0.00	0.05	-0.02	0.12	-0.04	0.05	1.00

Fig. 1 Pearson correlation for benign class



	x1	x2	x3	x4	x5	x6	x7	x8	x9
x1	1.00	0.10	0.11	-0.15	0.01	-0.04	-0.02	-0.01	0.12
x2	0.10	1.00	0.72	0.32	0.46	-0.04	0.39	0.30	0.24
x3	0.11	0.72	1.00	0.26	0.38	0.05	0.34	0.31	0.21
x4	-0.15	0.32	0.26	1.00	0.21	0.19	0.33	0.18	0.21
x5	0.01	0.46	0.38	0.21	1.00	-0.03	0.21	0.23	0.34
x6	-0.04	-0.04	0.05	0.19	-0.03	1.00	0.14	-0.08	-0.04
x7	-0.02	0.39	0.34	0.33	0.21	0.14	1.00	0.25	0.06
x8	-0.01	0.30	0.31	0.18	0.23	-0.08	0.25	1.00	0.22
x9	0.12	0.24	0.21	0.21	0.34	-0.04	0.06	0.22	1.00

Fig. 2 Pearson correlation for malignant class

two sections. These are the training and testing phase. K fold cross-validation depicts that a single fold is utilized for testing and $k-1$ folds are being used training circularly. Cross-validation is used for the avoidance of overfitting. In our study, a ten-fold cross-validation technique is used to partition data in which nine-fold are used for training and the remaining one-fold for testing in each iteration.

Theoretical Considerations

In the machine learning strategies, the learning procedure can be parted into two principal classifications such as supervised and unsupervised learning. In supervised learning, an arrangement of information cases is utilized to prepare the machine and is marked to give the right outcome. But in case of unsupervised learning, there are no pre-decided informational indexes, no idea of the usual result, which implies that the objective is harder to accomplish. Regression and classification are the most common methods that go under supervised learning. In case of regression, the target variable is continuous, and for classification, the target variable that is used for prediction is discrete.

Support Vector Machine

Support vector machine [30] is a speculation of a natural classifier called maximal edge classifier. Maximal edge classifier accompanies the meaning of hyperplane which expresses in an n-dimensional space. The hyperplane is of $(n-1)$ dimensions with level subspace that need not go through the root. It is difficult to draw a hyperplane in a higher dimension, so $(n-1)$ dimensional level subspace is still used. An SVM classifier can be constructed easily if there exists a separating hyperplane. The dataset categories cannot be divided using hyperplane, so feature space has to be enlarged using Gaussian radial basis function (RBF) or sigmoid function, cubic, quadratic or even higher order polynomial function. The hyperplane that is used in p-dimensions is as follows:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \quad (1)$$

where X_1, X_2,…, and X_p are the data points in the sample space of p-dimension and β_0, β_1, β_2,…, and β_p are the hypothetical values.

K-Nearest Neighbors

K-nearest neighbor algorithm is utilized for grouping and used in pattern recognition. It is widely used in predictive analysis. On the arrival of new data, the K-NN algorithm identifies existing data points that are nearest to it. Any attributes that can differ on a large scale may have sufficient influence on the interval between data points [30]. The feature vectors, as well as class labels, are stored in the training phase. K-NNs assume that the data samples are represented in a metric space. In the classification phase, first, the quantity is characterized by neighbors of K that is the most regular among the K training sample. At that point, the calculation will discover K adjacent neighbors of the new data sample. As all the data points are in metric space, a significant concern is how the distance will be calculated.

If the number of neighbors is denoted by N in K-NNs, then N samples are considered using the following distance metric value:

$$\text{Minkowski Distance: Dist}(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (2)$$

where if $p = 1$, then it is Manhattan distance, if $p = 2$, then it is Euclidean distance, and if $p = \infty$, then it is Chebyshev distance.

Among many choices, Euclidean distance is globally used. Among these K neighbors, the calculation will then check the quantity of information that focuses on every class, and afterward, it will relegate the new information point to the classification which frames the more significant part.

Random Forests

Random forest classifier is a powerful supervised classification tool. The RF classification is an ensemble method that can be studied as a form of the nearest neighbor predictor. Ensemble learning is the method by which statistical methods like classifiers or experts are strategically developed and incorporated to solve a specific problem of computational intelligence. RF generates a forest of classification trees from a given dataset, rather than a single classification tree. Each of these trees produces a classification for a given set of attributes [30, 31].

The workflow of random forest is given below.

(i) From the training set, picked K data points randomly.
(ii) From these K data points, generate the decision trees.
(iii) From generated trees, choose the number of N-tree and repeat steps (i) and (ii).
(iv) Form the N-tree that predicts the category to which the data points relate for a new data point, and assign the new data point via the category with the highest probability.

Artificial Neural Networks

Artificial neural network algorithm is slightly inspired by biological neuron and work by following the workflow of biological neurons dendrite, soma, and axon. The internal structure of every ANN is an artificial neuron and a simple mathematical function [32, 33]. The basic architecture of an artificial neural network is a set of interconnected neurons located in three different layers named input, hidden, and output layers. This type of network generally learns to perform tasks by considering a sufficient number of examples. The neural network can be applied both for classification and regression problems. There exist two types of ANNs which are perceptron, the simplest form of ANNs used for binary classification, and multilayer ANNs, a more sophisticated form of perceptron used to solve complex classification and regression problems.

Following is the representation for forward propagation and prediction of a single neuron:

$$\text{Output} = b_i + \sum_{j=1}^{n_x} w_{ij} x_i \quad (3)$$

where w_{ij} weight from input to output layer, b_i bias value, and x_i input value.

An activation function is applied to the output value after the calculation of it. Common activation functions used in artificial neural network are as follows:

$$\text{Sigmoid: Activation}(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

$$\text{Tanh: Activation}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \quad (5)$$

$$\text{Rectified Linear Unit: Activation}(x) = \begin{cases} 0, & \text{for } x \le 0 \\ x, & \text{for } x > 0 \end{cases} \quad (6)$$

$$\text{Leaky Rectified Linear Unit: Activation}(x) = \begin{cases} 0.01x, & \text{for } x < 0 \\ x, & \text{for } x \ge 0 \end{cases} \quad (7)$$

$$\text{Softmax: Activation}(x) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}, \text{ where } i = 1, 2, \ldots, j \quad (8)$$

We have applied a feed-forward network (known as multilayer perceptron) for classification for this study. In the architecture of the ANN, the number of neurons in the input layer is equal to the number of the attribute in the dataset. Another part of the network is the hidden layer where the number of hidden layers is regarded as one layer.

After forward propagation, the loss value is calculated from predicted value and actual value by following:

$$\text{Cross Entropy: Loss (Output, True Value)} = -(T\log(O) + (1-T)\log(1-O)) \quad (9)$$

$$\text{Residual Error: Loss (Output, True Value)} = O - T \quad (10)$$

$$\text{Squared Error: Loss (Output, True Value)} = (O-T)^2 \quad (11)$$

The calculation of loss is followed by a weight update in the back-propagation step. The representation for weight update is the following:

$$\Delta w_i = \eta(T-O)x_i \quad (12)$$

$$w_i = w_i + \Delta w_i \quad (13)$$

where η learning rate.

During our work, the input layer consists of 9 neurons which connect to 13 other neurons of the first hidden layer. Then there exist 13–9 mapped connections between first hidden layers to the second hidden layer. As the problem was a binary classification problem, there exists only one neuron in the output layer. The model was tuned for seventy epochs with a batch size of five. The system architecture of artificial neural networks algorithm is illustrated in Fig. 3.

Logistic Regression

Logistic Regression is an analytical modeling technique where the likelihood of a level is associated with a set of explicative variables. It is used for analyzing a dataset in which there are one or more independent variables that decide a result. The result is measured with a binary variable (in which there are only two possible results). It is applied to

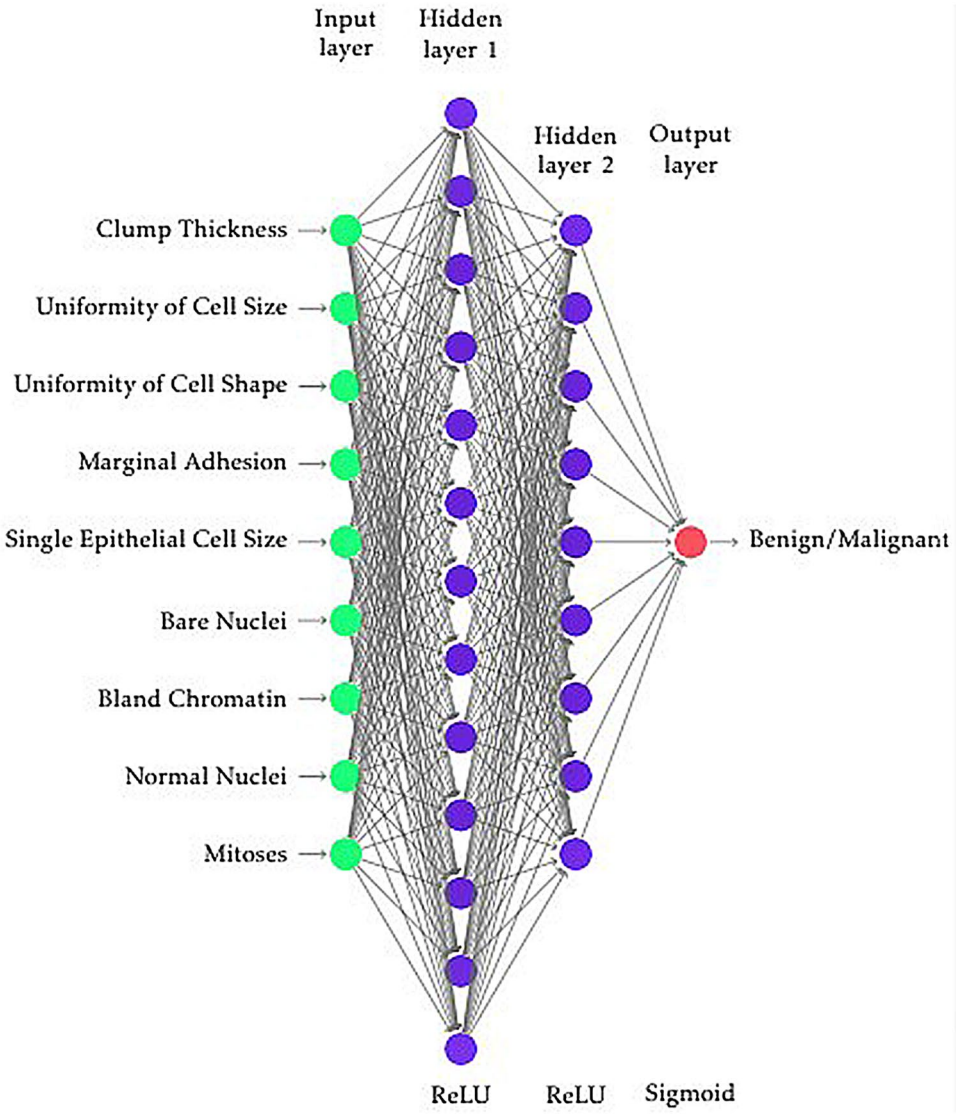


Fig. 3 The system architecture of artificial neural networks

predict a binary result (True/False, 1/0, Yes/No) given a set of independent variables. The following equations are the representation of the LR model:

$$x = c_o + \sum_{i=1}^{n} c_i x_i \tag{14}$$

$$P(x) = \frac{e^x}{1 + e^x} \tag{15}$$

where x is a quantity of the participation of the illustrative variables x_i ($i = 1,\ldots,n$), c_i is the regression coefficient that is achieved by the highest probability in association with its usual errors. Δc_i and $P(x)$ are the certain acknowledgments of variables that describe the likelihood of an excitement. In this investigation, the threshold was considered to be equal to or greater than 0.5, i.e., $P(x) \geq 0.5$,which results in a record being classified as an excitement [34]. In LR, likelihood P of a certain event can be calculated from the Bernoulli test and can be correlated with the sampling event [35–38].

Performance Measure Parameters

The performance of machine learning techniques is measured with respect to a few performance measure parameters. A confusion matrix including TP, FP, TN, and FN for actual data and predict data is formed to evaluate the parameters. The implication of the terms is given below:

TP	=	True Positive
TN	=	True Negative
FP	=	False Positive
FN	=	False Negative

In our study, the following parameters are used extensively to evaluate some terms by their corresponding formula to measure the performance of our study. There are a lot of parameters like these which describe some relationships that can help to measure the performance of a system. The comparative study's performance is evaluated by the following formulas:

Accuracy (*Acc*) The ratio of correctly classified samples to total samples:

$$\text{Accuracy (Acc)} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{16}$$

Sensitivity (*Sen*) Sensitivity is also regarded as recall. The rate of the perceived positive case with the total positive cases:

$$\text{Sensitivity (Sen)} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{17}$$

Specificity (*Spec*) Specificity means the relationship of observed negative examples with all negative examples, says the rate of predicted presence including entire examples by the presence of breast cancer:

$$\text{Specificity (Spec)} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \tag{18}$$

Precision (*Prec*) Precision is named the division of the examples which are actually positive among all the examples that we predicted positive:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{19}$$

Negative predictive value (*NPV*) NPV is the proportion of negatively classified cases that remained truly negative:

$$\text{Negative predictive value (NPV)} = \frac{\text{TN}}{(\text{TN} + \text{FN})} \tag{20}$$

False-positive rate (*FPR*) False-positive rate is measured as the quantity of false-positive predictions partitioned by the total amount of negatives. The valid false-positive rate is 0.0 through the most exceedingly highest is 1.0:

$$\text{False - positive rate (FPR)} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \tag{21}$$

False-negative rate (*FNR*) The rate of the event of negative test brings about the individuals who have the quality or sickness for which they are examined:

$$\text{False - negative rate (FNR)} = \frac{\text{FN}}{(\text{FN} + \text{TP})} \tag{22}$$

F1 score F1 score is defined as the harmonic mean between precision and sensitivity:

$$\text{F1 score} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \tag{23}$$

Matthews correlation coefficient (*MCC*) For binary classification, MCC is used. Here the range is + 1 to − 1. When the value is + 1, the best performance is shown and when the value is – 1, the worst performance is shown. It is represented as:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{24}$$

Table 2 SVM-confusion matrix for ten-fold cross-validation

	Benign	Malignant
Benign	TP = 44 (62.86%)	FP = 1 (1.43%)
Malignant	FN = 0 (0.00%)	TN = 23 (32.85%)

Table 3 KNN-confusion matrix for ten-fold cross-validation

	Benign	Malignant
Benign	TP = 5 (64.29%)	TP = 45 (64.29%)
Malignant	FN = 1 (1.43%)	TN = 23 (32.85%)

Table 4 RF-confusion matrix for ten-fold cross-validation

	Benign	Malignant
Benign	TP = 4 (62.86%)	TP = 44 (62.86%)
Malignant	FN = 2 (2.86)	TN = 23 (32.85%)

Table 5 ANN-confusion matrix for ten-fold cross-validation

	Benign	Malignant
Benign	TP = 45 (64.29%)	TP = 45 (64.29%)
Malignant	FN = 0 (0.00%)	TN = 24 (34.28%)

Table 6 LR-confusion matrix for ten-fold cross-validation

	Benign	Malignant
Benign	TP = 45 (64.29%)	TP = 45 (64.29%)
Malignant	FN = 2 (2.86%)	TN = 22 (31.42%)

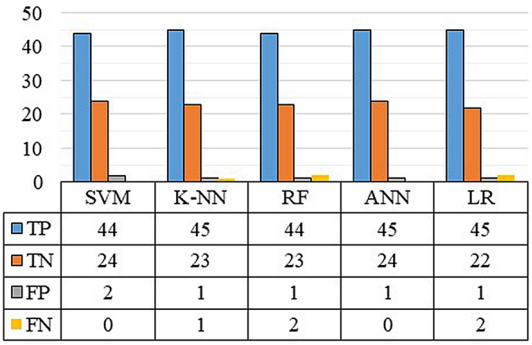


Fig. 4 Confusion matrix for the prediction of breast cancer using five machine learning techniques

Implementation and Result Analysis

Experimental Setup

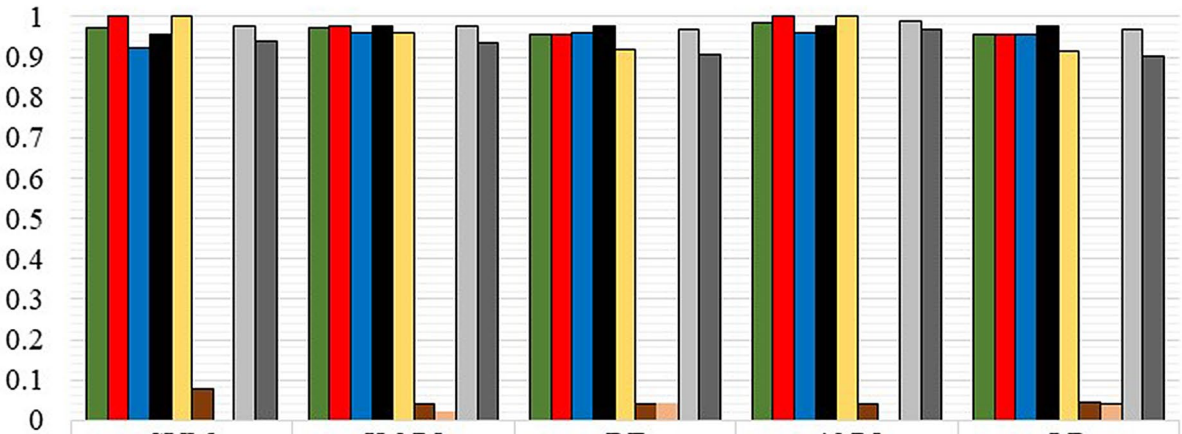
To predict whether a cell is benign or malignant, we have used five machine learning techniques such as SVM, K-NNs, RFs, ANNs, and LR individually. We used an Intel Core i7 powered computer with 32 GB RAM for processing purposes. Scikit-learn, an open-source machine learning library in Python programming language is used. Jupyter Notebook is an open-source web application that permits to develop and share reports that include live code, visualizations, equations, and narrated text which is utilized to fulfill our goal.

Results and Discussion

We have applied a ten-fold cross-validation strategy, i.e., the data set was part of ten portions. The ten-fold cross-validation technique is used to endorse the deliberate model. In this technique, nine-fold is used for training and the remaining one for testing. The confusion matrix is calculated for each technique. From the dataset of 699 instances, we used 629 instances which are 90% of the total data to train for all five techniques. We used 70 instances to test both our trained models. The confusion matrix of used machine learning strategies is illustrated in Tables 2, 3, 4, 5, and 6 which provides the prediction outcome of SVM, K-NNs, RFs, ANNs, and LR, respectively.

The combined confusion matrix is illustrated in Fig. 4 which depicts that the K-NNs, ANNs, and LR model predicts the largest number of the true positives (45 out of 70 test samples) among the five techniques. In addition, SVM and ANN models predict the largest number of true negatives and the lowest number of false negatives (24 among 70 test samples) and (0 among 70 test samples), respectively. The lowest number of false positive (1 out of 70 samples) is achieved by K-NNs, RFs, ANNs, and LR respectively.

The calculated performance measures are illustrated in Fig. 5 and Table 7. Figure 5 depicts that ANNs outperformed all other machine learning techniques so far we have studied with the highest accuracy of 98.57% whereas K-NNs and SVM achieved the second highest accuracy of 97.1%. Additionally, the highest specificity of 96% is obtained by the ANNs, and the lowest specificity of 92.3% is obtained by the SVM. ANNs, KNNs, and LR are outperformed by a precision of 97.8%. The lowest false-positive rate and false-negative rate are achieved by ANNs. SVM and ANNs have negative predictive values of 1 whereas the second next value is 0.958, which depicts that these methods are sensitive in the verification of positive tested samples. All the techniques have an F1 score of nearly 97%, which is comparatively better.



	SVM	K-NN	RF	ANN	LR
Accuracy	0.9714	0.9714	0.9571	0.9857	0.9571
Sensitivity	1	0.9782	0.9565	1	0.9574
Specificity	0.923	0.9583	0.9583	0.96	0.9565
Precision	0.9565	0.9782	0.9777	0.9782	0.9782
NPV	1	0.9583	0.92	1	0.9166
FPR	0.0769	0.0416	0.0416	0.04	0.0434
FNR	0	0.0217	0.0434	0	0.0425
F1 Score	0.9777	0.9782	0.967	0.989	0.9677
MCC	0.9396	0.9365	0.9062	0.969	0.9043

Fig. 5 Performance measurement parameters for the prediction of breast cancer using five machine learning techniques

ROC and PR-AUC

The ROC curve is a key appliance for analytic test estimation. In a ROC curve, the true-positive rate (sensitivity) is plotted against the false-positive rate (1 − specificity) at various threshold settings. ROC curve expresses a relation between true-positive rate vs. false-positive rate. The ROC curve for the breast cancer prediction using five machine learning techniques is illustrated in Fig. 6.

The precision–recall (PR) curve denotes a relation between precision vs. recall. Precision is a measure of how many of the individuals are predicted by the classifier as positive in case of total positive. The recall is a measure of the likelihood that estimates 1 given all the examples whose correct class label is 1. The PR-AUC for the breast cancer prediction using five machine learning techniques is illustrated in Fig. 7.

Table 7 Performances of breast cancer prediction system

	SVM	K-NN	RF	ANN	LR
Accuracy (%)	97.14	97.14	95.71	98.57	95.71
Sensitivity (%)	100	97.82	95.65	100	95.74
Specificity (%)	92.3	95.83	95.83	96	95.65
Precision (%)	95.65	0.9782	0.9777	0.9782	0.9782
NPV (%)	100	95.83	92	100	91.66
FPR (%)	7.69	4.16	4.16	4	4.34
FNR (%)	0	2.17	4.34	0	4.25
F1 score	0.9777	0.9782	0.967	0.989	0.9677
MCC	0.9396	0.9365	0.9062	0.969	0.9043

Comparative Study

A comparison study is illustrated in Table 8 for breast cancer prediction. The accuracy achieved by the Kernel-based orthogonal transform [39] is 98.53%. The authors in [20] measured the performance of SDT, BDT, and DTF for the prediction of breast cancer. The accuracy obtained by the techniques is 95.75%, 97.07%, and 97.51% in SDT, BDT, and DTF, respectively. Local linear wavelet neural network (LLWNN) [21] obtained an accuracy of 97.2%. The classification accuracy achieved by RBFNN-KPSO is 97.85%,

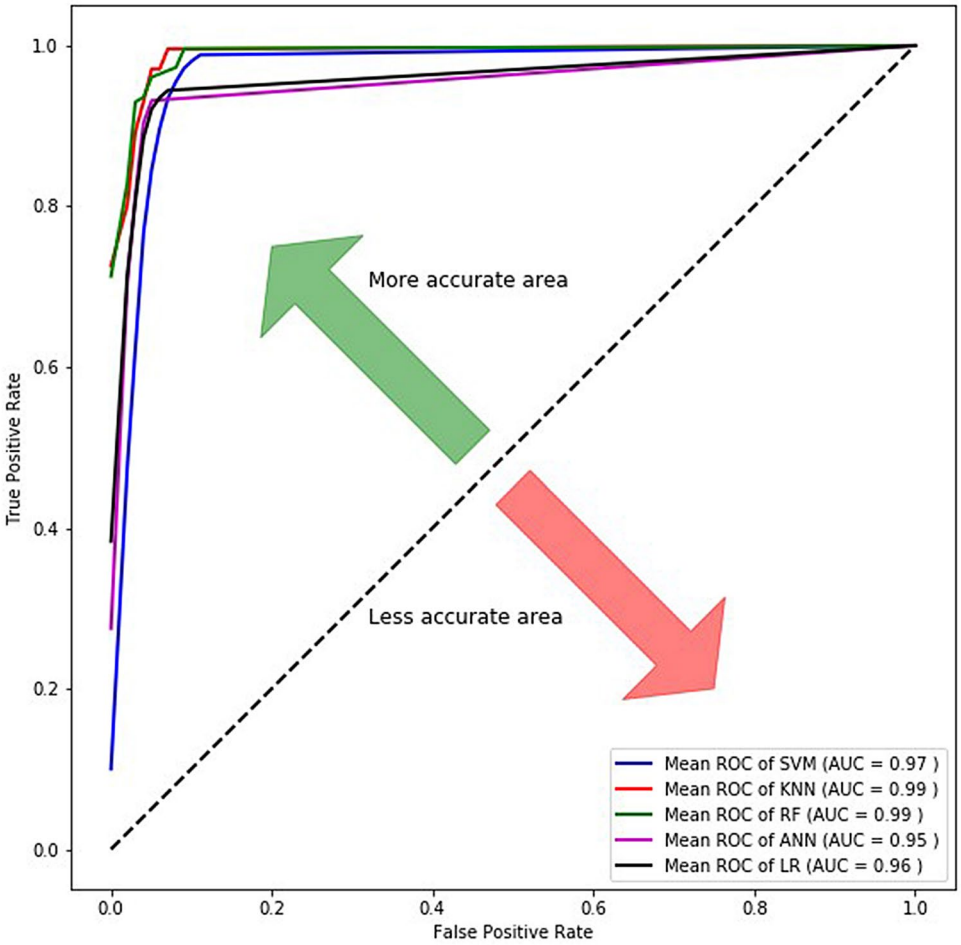


Fig. 6 ROC curve for the prediction of breast cancer using five machine learning techniques

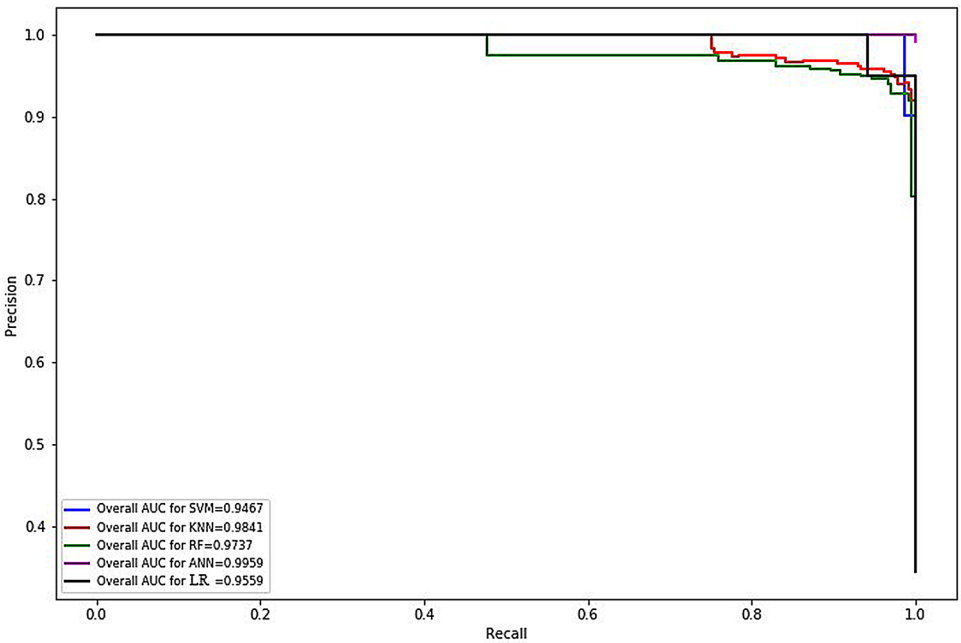


Fig. 7 PR-AUC for the prediction of breast cancer for five machine learning techniques

Table 8 The comparison of our study with the state of the art

Authors	Year	Method	Accuracy
Xu et al. [39]	2012	Kernel-based orthogonal transform	98.53
Azar et al. [20]	2013	SDT	95.75
Azar et al. [20]	2013	BDT	97.07
Azar et al. [20]	2013	DTF	95.51
Senapati et al. [21]	2013	LLWNN	97.20
Senapati et al. [22]	2014	RBFNN-KPSO	97.85
Senapati et al. [22]	2014	RBFNN extended	96.42
Azar et al. [24]	2014	LPSVM	97.14
Azar et al. [24]	2014	LSVM	95.42
Azar et al. [24]	2014	SSVM	96.57
Azar et al. [24]	2014	PSVM	96
Azar et al. [24]	2014	NSVM	96.57
Azar et al. [24]	2014	St-SVM	94.86
Latchoumi and Parthiban [40]	2017	WPSO-SSVM	98.42
Kumar et al. [41]	2017	SVM-Naive Bayes-J48	97.13
Sakri et al. [15]	2018	Naive Bayes	81.3
Sakri et al. [15]	2018	RepTree	80
Sakri et al. [15]	2018	k-NNs	75
Banu and Subramanian [18]	2018	Bayes belief network	91.7
Banu and Subramanian [18]	2018	Boosted augmented Naive Bayes	91.7
Banu and Subramanian [18]	2018	Tree augmented Naive Bayes	94.11
Chaurasia et al. [19]	2018	Naive Bayes	97.36
Chaurasia et al. [19]	2018	RBF network	96.77
Chaurasia et al. [19]	2018	J48	93.41
Our study	–	RF	95.71
		LR	95.71
		SVM	97.14
		K-NN	97.14
		ANN	98.57

and the RBFNN-extended Kalman filter is 96.4235% in [22]. The accuracy obtained by LPSVM, LSVM, SSVM, PSVM, NSVM, St-SVM are 97.1429%, 95.4286%, 96.5714%, 96%, 96.5714% and 94.86%, respectively, in [24].

A weighted-particle swarm optimization (WPSO) with a smooth support vector machine (SSVM) for prediction accuracy achieved 98.42% [40]. The proposed system in [41] combined SVM, Naive Bayes, and J48 using the voting classifier method to achieve accuracy of 97.13% which is better than each of individual classifiers. The system developed in [15] acquired 70%, 76.3%, and 66.3% accuracy for NB, RepTree, and K-NNs, respectively. With PSO implemented, they have found four features which are best for this classification task. For NB, RepTree, and K-NNs with PSO, they obtained 81.3%, 80%, and 75% accuracy value, respectively. According to the findings in [18] with the help of gradient boosting, 91.7%, 91.7%, and 94.11% accuracy have been achieved for BBN, BAN, and TAN, respectively. The obtained results in [19] illustrated that the Naive Bayes algorithm performed great with the accuracy of 97.36%, RBF network with the accuracy of 96.77%, and the J48 came out to be the third with an accuracy of 93.41%. In the overall comparison, ANN model performs comparatively better than the other techniques in our study.

Conclusion

This paper presented a comparative study of five machine learning techniques for the prediction of breast cancer, namely support vector machine, K-nearest neighbors, random forests, artificial neural networks, and logistic regression. The basic features and working principle of each of the five machine learning techniques were illustrated. The highest accuracy obtained by ANNs is 98.57% whereas the lowest accuracy derived from the RFs and LR is 95.7%. The diagnosis procedure in the medical field is very expensive as well as time-consuming. The system proposed that machine learning technique can be acted as a clinical assistant for the diagnosis of breast cancer

and will be very helpful for new doctors or physicians in case of misdiagnosis. The developed model by ANNs is more consistent than any other technique stated, and it may be able to bring changes in the field of prediction of breast cancer. From the study, we can conclude that machine learning techniques are able to detect the disease automatically with high accuracy.

Acknowledgements This research was partially supported by Universiti Malaysia Pahang (UMP) through UMP Flagship Grant (RDU192206).

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiol Soc N Am. 2018;286(3):800–9.
2. Breast Cancer: Statistics, Approved by the Cancer.Net Editorial Board, 04/2017. [Online]. Available: http://www.cancer.net/cancer-types/breast-cancer/statistics. Accessed 26 Aug 2018.
3. Mori M, Akashi-Tanaka S, Suzuki S, Daniels MI, Watanabe C, Hirose M, Nakamura S. Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts. Springer. 2016;24(1):104–10.
4. Kurihara H, Shimizu C, Miyakita Y, Yoshida M, Hamada A, Kanayama Y, Tamura K. Molecular imaging using PET for breast cancer. Springer. 2015;23(1):24–32.
5. Azar AT, El-Said SA. Probabilistic neural network for breast cancer classification. Neural Comput Appl. 2013;23(6):1737–51.
6. Nagashima T, Suzuki M, Yagata H, Hashimoto H, Shishikura T, Imanaka N, Miyazaki M. Dynamic-enhanced MRI predicts metastatic potential of invasive ductal breast cancer. Springer. 2002;9(3):226–30.
7. Park CS, Kim SH, Jung NY, Choi JJ, Kang BJ, Jung HS. Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions. Springer. 2013;22(2):153–60.
8. Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. IETE J Res. 2020;. https://doi.org/10.1080/03772063.2020.1713916.
9. Muhammad LJ, Islam MM, Usman SS, Ayon SI. Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery. SN Comput Sci. 2020;1(4):206.
10. Islam MM, Iqbal H, Haque MR, Hasan MK. Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 226–229.
11. Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK. Performance evaluation of random forests and artificial neural networks for the classification of liver disorder. In: Proc. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, 2018, pp. 1–5.
12. Ayon SI, Islam MM. Diabetes prediction: a deep learning approach. Int J Inf Eng Electron Bus (IJIEEB). 2019;11(2):21–7.
13. Islam MZ, Islam MM, Asraf A. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, 2020. pp. 1–20.
14. Hasan MK, Islam MM, Hashem MMA. Mathematical model development to detect breast cancer using multigene genetic programming. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), pp. 574–579, 2016.
15. Sakri SB, Rashid NBA, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access. 2018;6:29637–47.
16. Juneja K, Rana C. An improved weighted decision tree approach for breast cancer prediction. In: International Journal of Information Technology, 2018.
17. Yue W, et al. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018;2(2):13.
18. Banu AB, Subramanian PT. Comparison of Bayes classifiers for breast cancer classification. Asian Pac J Cancer Prev (APJCP). 2018;19(10):2917–20.
19. Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. J Algorithms Comput Technol. 2018;12(2):119–26.
20. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput Appl. 2012;23(7–8):2387–403.
21. Senapati MR, Mohanty AK, Dash S, Dash PK. Local linear wavelet neural network for breast cancer recognition. Neural Comput Appl. 2013;22(1):125–31.
22. Senapati MR, Panda G, Dash PK. Hybrid approach using KPSO and RLS for RBFNN design for breast cancer detection. Neural Comput Appl. 2014;24(3–4):745–53.
23. Hasan MK, Islam MM, Hashem MMA (2016) Mathematical model development to detect breast cancer using multigene genetic programming. In: Proc. 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, 2016, pp. 574–579.
24. Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl. 2013;24(5):1163–77.
25. Ferreira P, Dutra I, Salvini R, Burnside E. Interpretable models to predict Breast Cancer. In: Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, 2016, pp. 1507–1511.
26. Jhajharia S, Verma S, Kumar R. A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. In: Proc. International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1–7.
27. Islam MM, Rahaman A, Islam MR. Development of smart healthcare monitoring system in IoT environment. SN Comput Sci. 2020;1(3):185.
28. Rahaman A, Islam M, Islam M, Sadi M, Nooruddin S. Developing IoT based smart health monitoring systems: a review. Rev d'Intell Artif. 2019;33(6):435–40.
29. Breast Cancer Wisconsin (Original) Data Set, [Online]. https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data. Accessed 25 Aug 2018.
30. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. 1st ed. New York: Springer; 2013.
31. Guido S, Mller AC. Introduction to machine learning with python. Sebastopol: O'Reilly Media Inc.; 2016.
32. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2016;29(10):685–93.

33. Ratner B. Statistical and machine-learning data mining: techniques for better predictive modeling and analysis of big data. Oxford: Chapman and Hall/CRC; 2017.
34. Dong L, Wesseloo J, Potvin Y, Li X. Discrimination of mine seismic events and blasts using the fisher classifier, naive bayesian classifier and logistic regression. Rock Mech Rock Eng. 2015;49(1):183–211.
35. Hosmer DW Jr, Lemeshow S. Applied logistic regression. New York: Wiley; 2004.
36. Schumacher M, Roner R, Vach W. Neural networks and logistic regression: part I. Comput Stat Data Anal. 1996;21(6):661–82.
37. Vach W, Roner R, Schumacher M. Neural networks and logistic regression: part II. Comput Stat Data Anal. 1996;21(6):683–701.
38. Hajmeer M, Basheer I. Comparison of logistic regression and neural network-based classifiers for bacterial growth. Food Microbiol. 2003;20(1):43–55.
39. Xu Y, Zhu Q, Wang J. Breast cancer diagnosis based on a kernel orthogonal transform. Neural Comput Appl. 2012;21(8):1865–70.
40. Latchoumi TP, Parthiban L. Abnormality detection using weighed particle swarm optimization and smooth support vector machine. Biomed Res. 2017;28:4749–51.
41. Kumar UK, Nikhil MBS, Sumangali K. Prediction of breast cancer using voting classifier technique. In: Proc. IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, 2017, pp. 108–114.