Computer Methods and Programs in Biomedicine Update 3 (2023) 100098

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage: www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

Machine learning-based diagnosis of breast cancer utilizing feature optimization technique

Khandaker Mohammad Mohi Uddin [a,*], Nitish Biswas [a], Sarreha Tasmin Rikta [a], Samrat Kumar Dey [b]

[a] Department of Computer Science and Engineering, Dhaka International University, Dhaka, 1205, Bangladesh
[b] School of Science and Technology, Bangladesh Open University, Gazipur, 1705, Bangladesh

ARTICLE INFO

Keywords:
Breast cancer
Feature optimization
Diagnosis
WBCD
Voting classifier
Web application

ABSTRACT

Breast cancer disease is recognized as one of the leading causes of death in women worldwide after lung cancer. Breast cancer refers to a malignant neoplasm that develops from breast cells. Developed and less developed countries both are suffering from this extensive cancer. This cancer can be recuperated if it is detected at an early stage. Many researchers have proposed several machine learning techniques to predict breast cancer with the highest accuracy in the past years. In this research work, the Wisconsin Breast Cancer Dataset (WBCD) has been used as a training set from the UCI machine learning repository to compare the performance of the various machine learning techniques. Different kinds of machine learning classifiers such as support vector machine (SVM), Random Forest (RF), K-nearest neighbors(K-NN), Decision tree (DT), Naïve Bayes (NB), Logistic Regression (LR), AdaBoost (AB), Gradient Boosting (GB), Multi-layer perceptron (MLP), Nearest Cluster Classifier (NCC), and voting classifier (VC) have been used for comparing and analyzing breast cancer into benign and malignant tumors. Various matrices such as error rate, Accuracy, Precision, F1-score, and recall have been implemented to measure the model's performance. Each Algorithm's accuracy has been ascertained for finding the best suitable one. Based on the analysis, the result shows that the Voting classifier has the highest accuracy, which is 98.77%, with the lowest error rate. Finally, a web page is developed using a flask micro-framework integrating the best model using react.

1. Introduction

Cancer is now one of the most horrific diseases in the world. People worldwide have terrible cancer such as lung, cervical, thyroid, liver, and breast cancer; among these cancers, breast cancer is hazardous. Breast Cancer mainly occurs in women. The World Health Organization (WHO) says, approximately 0.5-1% of breast cancers occur in men. It has been a known disease for a long time because this cancer was noted by Ancient Egyptians more than 3,000 years ago [1,2]. An uncontrolled growth that originates in the breast is called Breast cancer. The unchecked growth of the cell is called tumors. These tumors are mainly of two types- 'Benign,' non-cancerous, and 'Malignant' which are cancerous [3]. Benign tumors grow slowly, and they have distinct borders, never invading surrounding tissue or other body parts. On the other hand, malignant tumors proliferate, do not have any different edges, and quickly spread to the other parts of the body through a process called metastasis [4].There are some factors behind this breast cancer. Fig. 1 represents factors that affect breast cancer disease, and Table 1 illustrates the stages of development of breast cancer. This disease easily transits women's chest, glands, and milk ducts. Breast cancer is now the utmost common cancer all over the world, especially in developed countries, but now it is rapidly increasing in middle-income and low-income countries, according to WHO (World Health Organization), with more than 2.3 million women were diagnosed with breast cancer in 2020 and 685,000 died among them [5]. The number of dead women has increased from 685,000 to 963,000 in 2021, as WHO predicted. From the Breast Cancer Statistics and Resources (BCRF), it is known that a woman is diagnosed with breast cancer every 14 seconds somewhere in the world [6]. The mortality rate will increase in the future, and the world Health Organization portends that this could reach 2.9 million globally [7].Different medical tests are suggested to capture breast cancer, such as breast ultrasound, and breast exams by checking the doctor, Removing a sample of breast cells for testing

* Corresponding author.
E-mail addresses: jilanicsejnu@gmail.com (K.M.M. Uddin), nitishbiswas.cse@gmail.com (N. Biswas), tasmin.sarreha@gmail.com (S.T. Rikta), samrat.sst@bou.ac.bd (S.K. Dey).

https://doi.org/10.1016/j.cmpbup.2023.100098

Available online 19 February 2023
2666-9900/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

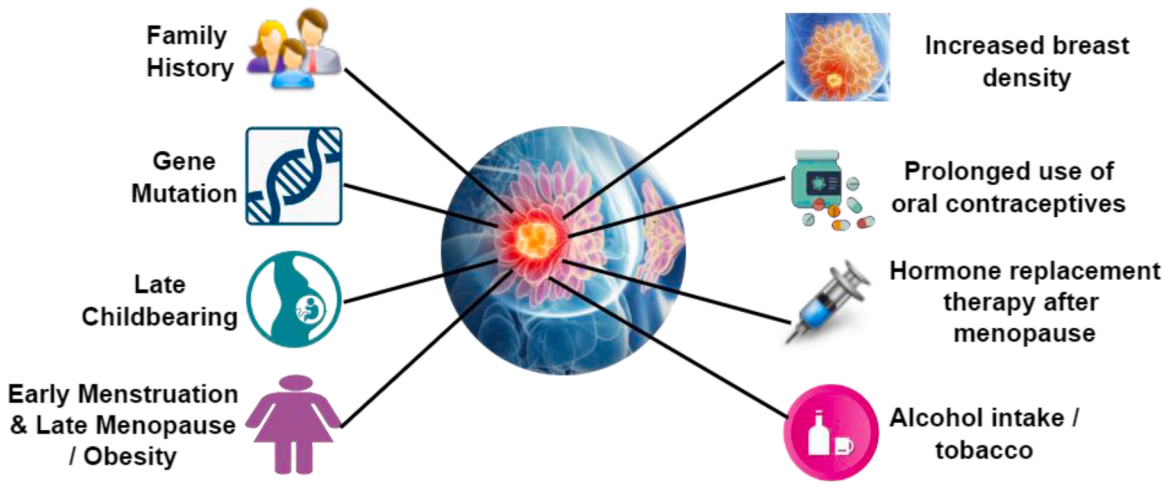


Fig. 1. Factors affecting Breast cancer disease.

Table 1
The stages of development of Breast Cancer.

Stage	Description	Size	Treatment stage
I	Non-invasive cancer	Small size	Breast-Conserving Surgery (BCS), Simple mastectomy
II	Invasive	2 cm or less	Surgery, Radiation Therapy
III	Invasive	2 to 5 cm	Chemotherapy, Hormone therapy, target drugs, Immunotherapy
IV	Invasive	Greater than 5 cm	Chemotherapy, Hormone therapy, Immunotherapy
V	Invasive	It can be any size	Chemotherapy, Hormone therapy, Immunotherapy

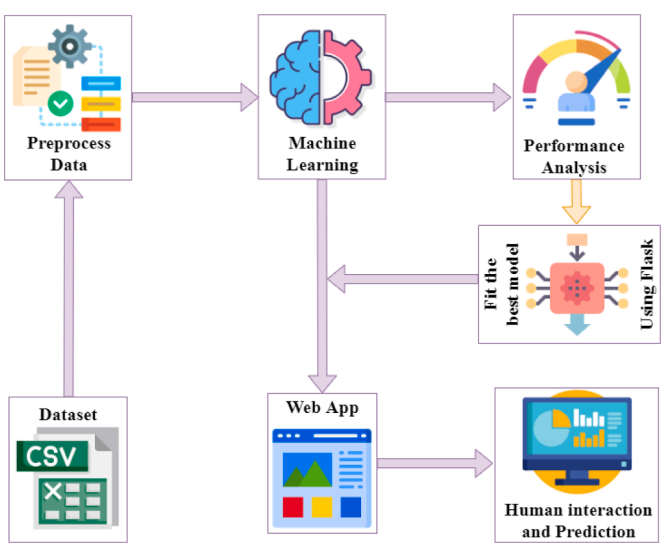


Fig. 2. Overview of the proposed work

(biopsy), Mammogram, and Breast magnetic resonance imaging (MRI). The biopsy is a definitive way to capture breast cancer among these testing techniques. After testing all these medical techniques, radiologists examine them and discuss them with doctors, and analyze whether the cells are cancerous or not. But the problem is that expert doctors are rare, most of the time, these tumors do not appear with any symptoms, and this process takes much time. In addition, all these medical testing are costly. However, patients can avoid unnecessary treatments if detected in the initial stage [8,9].Therefore, researchers proposed a Computer-Aided Diagnosis (CAD) that can detect tumors and show accurate results quickly without the help of any experts or radiologists [10]. Machine learning is a part of Artificial Intelligence that is widely used to predict diseases through experience and data. It applies different methods to improve performance, such as optimization, statistical, and probabilistic [11]. It is found in a survey that specialists and most experienced doctors can detect breast cancer with 79% accuracy, whereas the machine learning technique's accuracy is 91% which can prevent diseases very accurately than traditional techniques [12].

Therefore, machine learning can play a significant role in reducing the mortality rate. Various machine learning techniques are applied to capture breast cancer, such as Support Vector Machine, Naïve Bayes, K-nearest neighbors, and AdaBoost. However, eleven machine learning classifiers were used in this research work to discover the best suitable model for predicting breast cancer.

Hyperparameter tuning was implemented with the grid search in this study. This will produce an optimized classifier to detect Breast cancer with the highest accuracy. The overview of the proposed work is shown in Fig. 2. This study focuses on distinct objectives:

- To analyze an experiment in WBCD dataset from the UCI machine learning repository and make a relation with which feature has a concrete relationship to the class feature.
- To perform state-of-the-artMachine Learning classifiers on the dataset and compare their results.
- To evaluate the best model with a good prediction and promising accuracy.

These are the ways that we contribute to this proposed research:

1 This work outperforms previous research on this particular topic conducted by other researchers, with an accuracy of 98.77%.
2 Eleven classifiers and different machine learning techniques, including feature optimization techniques like PCA and feature scaling, hyperparameter tuning, and cross-validation, are employed in this research work to reach the best result.
3 A web page is developed based on this research that can calculate the result accurately using real-time inputs.

The rest of the part of this paper is divided into four sections. Section 2 briefly describes the related works. Materials and methods used in this work to evaluate breast cancer have been illustrated in Section 3; Section 4 discusses the result and analysis of the proposed model. Finally, the entire work will be concluded in Section 5.

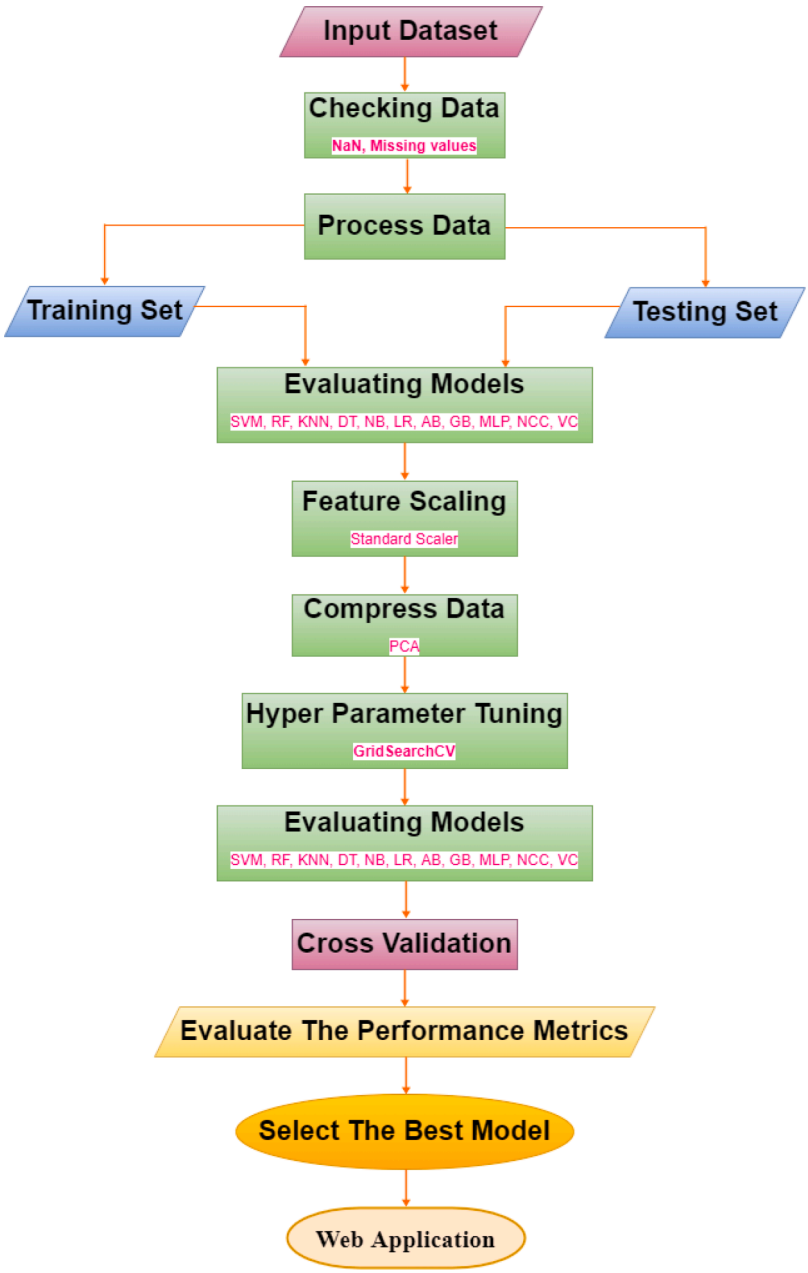


Fig. 3. Working procedure of proposed methodology.

2. Related works

In this section, the related work is surveyed on the classification of breast cancer data.Many researchers have already proposed several techniques in the past years and applied various machine learning algorithms to different breast cancer data repositories. The performance of each model differs from Algorithm to Algorithm and the dataset used by researchers.

Jabbar et al. [12] proposed a model to diagnose breast cancer using Bayesian Network (BN) and Radial Basis Function (RBF) algorithms. The authors used Wisconsin Breast Cancer Dataset (WBCD) with 699 instances and 10 attributes. In this work, BN and RBFs are involved in modeling ensemble classifiers. Ensemble classifier is a process that merges multiple other models to improve machine learning results. By using this model, they obtain 97.42% with the highest accuracy.

Gopal et al. [13] proposed Random Forest (RF), Logistic Regression (LR), and Multi-layer perceptron (MLP) models with 10 cross-fold methods for breast cancer prediction. In this work, the WBCD dataset is used from the UCI repository with 569 instances with 32 attributes, among which only 11 facts have been eliminated and selected using PCA (Principal Component Analysis), a linear feature selection method. In this proposed model, the MLP classifier shows a higher accuracy with 98% than Random Forest, Logistic Regression.

Ara et al. [14] applied a machine learning classifier in the WBCD dataset with 569 instances and 32 features collected from the UCI repository. They implement a feature selection method in the dataset for achieving the highest accuracy. In this research, they proposed Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), and K-Nearest Neighbors (KNN) classifier for prognosis breast cancer. This work shows that the SVM achieves maximum accuracy (96.5%) with fewer error rates.

Worldwide different authors [15–17] have performed various machine learning classifiers on the Wisconsin Breast Cancer Dataset (WBCD) of the UCI machine learning repository. Varshini et al. [15]

Algorithm 1
Working procedure of breast cancer prediction

Input: UCIMachine LearningRepository Breast Cancer Dataset
Output: Predicted value Malignant or Benign
1 **Begin**
1 data ← load dataset
1 if data.value is equal NaN or empty
1 replace NaN or missing_value
1 pre-processing:
1 if data.target is equal M
1 replace M with 0
1 else
1 replace B with 1
1 x ← data.drop[target]
1 y ← data.target
1 x_1, x_2, y_1, y_2 ← split_data of x and y
1 model ← train_model using $x_1$and y_1
1 predict ← testing_model using x_2 and y_2
1 s_x ← scaling data of x
1 s_x ← compress s_x data using PCA
1 apply hyperparameter tuning for each classifier
1 classifier ← train_model using s_x
1 model ← apply_voting_classifier using classifier
1 predict ← cross_validation with model
1 computes performance evaluation metrics
1 **End**

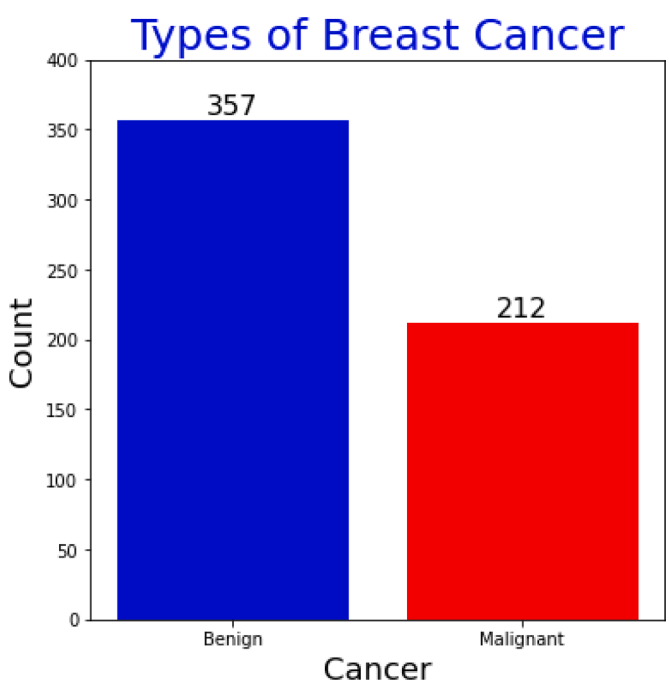


Fig. 4. Class distribution of breast cancer.

observed that the Random forest classifier has the highest accuracy with 96.5%.On the other hand, Chaurasia et al. [16] used PCA for decreasing the dimensions and also used SFS (sequential feature Selector) method to extract the classifier accuracy. RF, KNN, SVC, and SVC achieved the equivalent accuracy of 97.66%.Subramania et al. [17] proposed LR, K-NN, GNB, DT, and RF classifiers where they use feature scaling during the data pre-processing to standardize the feature in the dataset into a finite range and use PCA Linear discriminant analysis (LDA), which are used to reduce dimensionality. Finally, they apply cross-validation to utilize data better. The highest obtained accuracy is achieved at 96.59% from Random Forest.

Aesegie et al. [18] developed a model to diagnose breast cancer and adaptive boosting and decision tree are applied by the authors to the Kaggle breast cancer data repository, consisting of 569 observations and 31 features. The authors use Pearson's correlation to explain the relationship between two data samples. In addition, they employ data visualization in the breast cancer data repository to find out the relationship between the features and class. After that, the proposed model is performed; and by confusion matrix, the performance of a

Table 2
The importance of predictive variables in Breast Cancer.

Features	Priority ratio (%)
perimeter_worst	0.473412
area_worst	0.464993
radius_worst	0.453620
concave points_mean	0.440203
concave points_worst	0.437734
perimeter_mean	0.402974
concavity_mean	0.374344
radius_mean	0.367268
area_mean	0.361632
area_se	0.340142
concavity_worst	0.316236
perimeter_se	0.274725
radius_se	0.249887
compactness_worst	0.226390
compactness_mean	0.212241
concave points_se	0.126875
texture_worst	0.122004
concavity_se	0.117078
smoothness_worst	0.102887
texture_mean	0.097148
symmetry_worst	0.091050
smoothness_mean	0.074627
compactness_se	0.073327
fractal_dimension_worst	0.071018
symmetry_mean	0.055354
fractal_dimension_se	0.038078
symmetry_se	0.014087
smoothness_se	0.013692
fractal_dimension_mean	0.007138
texture_se	0.000000

Table 3
Statistical analysis of the dataset.

Features	Mean	Standard Division	Max	Min
radius_mean	14.127	3.524	6.981	28.11
texture_mean	19.289	4.3	9.71	39.28
perimeter_mean	91.969	24.29	43.79	188.50
area_mean	654.889	351.91	143.5	2501.00
smoothness_mean	0.096	0.014	0.052	0.16
compactness_mean	0.104	0.0528	0.019	0.34
concavity_mean	0.088	0.079	0	0.42
concave points_mean	0.048	0.038	0	0.20
symmetry_mean	0.181	0.027	0.106	0.30
fractal_dimension_mean	0.062	0.007	0.049	0.09
radius_se	0.405	0.277	0.11	2.87
texture_se	1.2168	0.55	0.36	4.88
perimeter_se	2.866	2.021	0.757	21.98
area_se	40.337	45.49	6.802	542.20
smoothness_se	0.007	0.003	0.0017	0.03
compactness_se	0.025	0.017	0.002	0.13
concavity_se	0.031	0.030	0	0.39
concave points_se	0.011	0.006	0	0.05
symmetry_se	0.020	0.008	0.0078	0.07
fractal_dimension_se	0.003	0.002	0.00089	0.029
radius_worst	16.269	4.833	7.93	36.04
texture_worst	25.677	6.146	12.02	49.54
perimeter_worst	107.26	33.602	50.41	251.20
area_worst	880.58	569.356	185.2	4254.00
smoothness_worst	0.132	0.022	0.071	0.22
compactness_worst	0.254	0.157	0.027	1.058
concavity_worst	0.272	0.208	0	1.252
concave points_worst	0.114	0.065	0	0.29
symmetry_worst	0.290	0.0618	0.156	0.66
fractal_dimension_worst	0.083	0.018	0.055	0.207

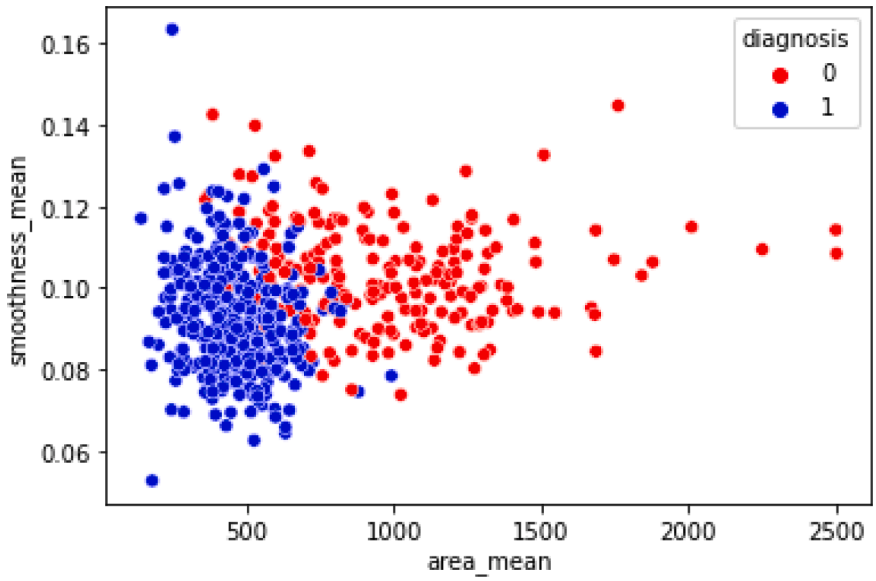


Fig. 5. Scatter plot of Features [smoothness_mean, area_mean]

classification model is evaluated on a set of test data where valid values are known. In this work, they observe adaptive boosting outperformed a Decision tree.

Acquisition et al. [19] proposed K-NN,SVM, and NB for predicting breast cancer. They used the breast cancer Wisconsin dataset from the UCI repository. Different cost, gamma, and various Kernel functions are applied to gain accuracy in SVM. Besides that, K-fold cross-validation with 10-folds is employed to prevent the overfitting of data. SVM and K-NN, both classifiers, give the best result with 96.85% with fewer error rates.

Mangukiya et al. [20] proposed various classifiers such as SVM, DT, NB, and AdaBoost, where they achieved 98.24% accuracy using XGboost with the lowest error rate to diagnose breast cancer. Aesegie et al. [21] applied an optimized K-NN classifier to detect breast cancer. A grid search is employed in this model for hyperparameter tuning, which results in the most accurate prediction. This research's highest accuracy is 94.35%, which is attained by using an optimized K-NN model.

Commission et al. [22] investigated the WBCD, containing 699 instances and 10 features. They proposed various machine learning models such as SVM, ANN, and NB, where they use a 10-fold cross-validation method, and from this research, the best model is considered SVM with maximum accuracy of 96.72%.

After observing the literature survey, it is observed that different classifiers and methods showed different results. Breast cancer is now more dangerous than any other disease, and it needs to predict it more accurately. As a result, the motivation is obtained to work with breast cancer to achieve a more accurate result.Inthis research work, we have achieved comparatively higher results than all other research works described in this session.

3. Materials and methods

This section describes our proposed approach and techniques employed to diagnose breast cancer. The performance of any machine learning research work differs from ML model to ML model, different attributes to different attributes. Appropriate feature selection methods obtain the best performance. For feature optimization, standard scaler technique and Principal Component Analysis (PCA) are applied in this research work. As a result, a reliable approach must be implemented to determine breast cancer more accurately.

StandardScaler performs the task of standardization. Usually, a dataset contains variables that are different in scale. For e.g. in this dataset contain radius_mean column with values on scale 6.981-28.11 and area_mean column with values on scale 143.5-2501.00. As these two columns are different in scale, they are standardized to have a common scale while building a machine learning model. PCA is performed after StandardScaler operation. The working procedure of our proposed methodology is shown in Fig. 3.

This proposed method explored feature selection techniques with standard scalar where each feature's values in the data have zero mean and unit variance, compress data using PCA, Hyperparameter tuning with Grid search can be used to evaluate the best model, and cross-validation is also employed here to detect overfitting. These features are fed into the proposed classifiers to determinebreast cancer with the highest accuracy. Using all machine learning classifiers in this experiment showed a promising result to predict breast cancer. Algorithm 1 of this research work is shown below for better understanding.

3.1. Dataset

Nowadays, machine learning classifiers are widely used in medical diagnosis. The Wisconsin Breast Cancer Dataset is publicly used in the experiments. The name of the developer of this dataset is Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, USA [14]. This dataset consists of 569 instances and 30 features that have been collected from the UCI repository, and this dataset has no missing value. This datasetcomprises 357 benign or non-cancerous and 212 malignant or cancerous breast cancer, as shown in Fig. 4.

The ratio of training and testing sets of this total dataset has been randomly split into 3:1.This dataset has a column named id, which never affects the result. So, in this work, this column is eliminated to avoid redundancy. Table 2 represents the predictive variables in breast cancer. Each feature has a direct relationship with the result. And in this table priority ratio of the segments means how much they are co-related with the outcome.

3.2. Data Pre-processing

Data pre-processing is operated to attain the best result in the dataset. The dataset may have noise, missing values or information, and unbalanced data, diminishing the outcome accuracy. As a result, we must clean these unwanted things from the dataset before performing the machine learning model. Diagnosis columns M and B represent malignant and benign in this data set. These string data are converted

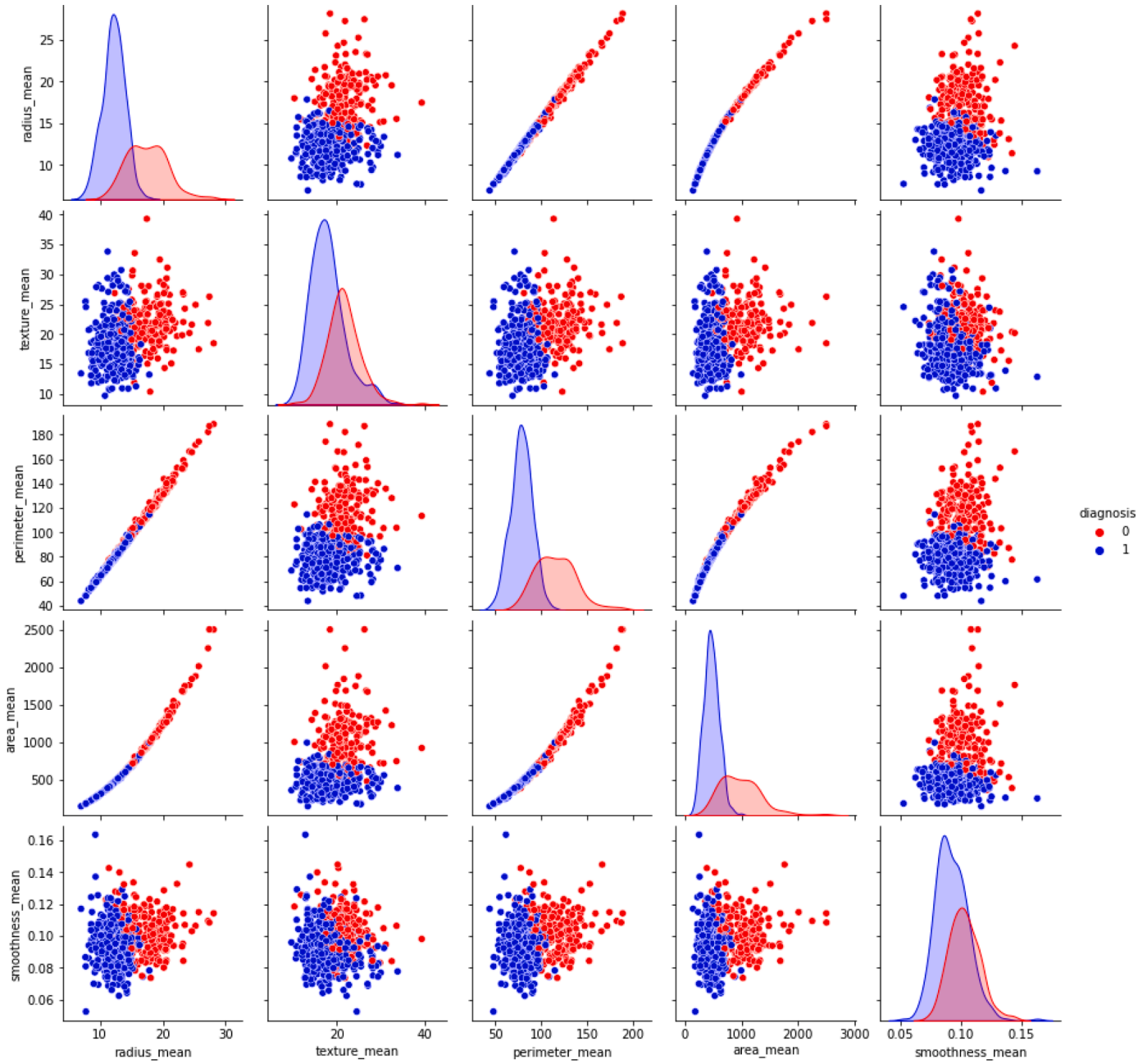


Fig. 6. Pair plot of Features [radius_mean, texure_mean, perimeter_mean, area_mean,smoothness_mean]

into numerical data 0, and 1, where 0 refers to malignant and 1 refers to benign. After that, Max, Min Normalization is performed to ensure all features have the same scale and data standardization ensures all features have the same influence on the distance metric. The dataset contains the following distinct features after pre-processing the data. Statistical analysis of the dataset is shown in Table 3.

3.3. Feature optimization using PCA

The used dataset in this study has high dimensionality that contains 30 features. These large number of features hinder reaching the best result, creating overfitting. As a result, Principal Component Analysis (PCA) is used in this dataset that transforms the 30 features into 16 features for increasing the performance of the result. PCAis an unsupervised machine learning algorithm, used to reduce the dimensionality that converts a large number of features in a dataset into a small number of features known as the Principal component. The main advantages of PCA are to reduce overfitting, remove correlated features, and elevate machine learning algorithm performance. It works based on the following method. Before applying PCA, data is normalized using Eq. (1) [23]. Then, build the covariance matrix using Eq. (2).

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

$$cov(A, B) = \frac{1}{n-1} \sum_{i=1}^{n} (Ai - \overline{a})(Bi - \overline{b}) \tag{2}$$

After that find out the eigenvectors and eigenvalues. And finally, eigenvalues are sorted in descending order and the number of principal components is selected.

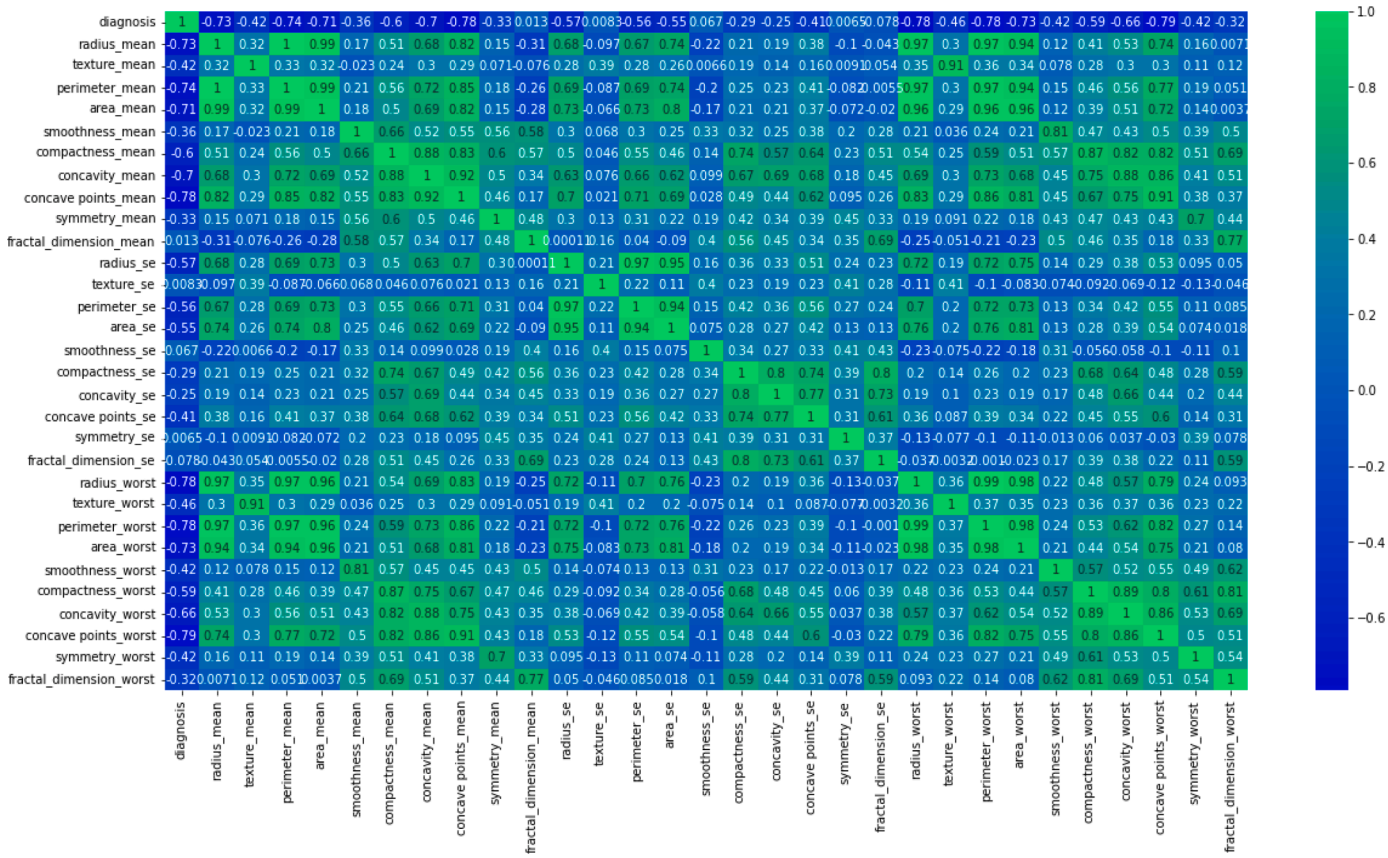


Fig. 7. Correlation with each feature.

Table 4
Environment setup of the proposed model system.

Resource	Details
CPU	Intel® Core™ i-3-1005G1 CPU @ 1.20GHz
RAM	12GB
GPU	Intel® UHD Graphics
Software	Anaconda
Language	Python

3.4. Data visualization

When information and data are represented graphically, data visualization makes data easier for the human brain.We have used a scatterplot to visualize the two variables for a set of data, whether they have a relationship or not.The pair plot represents the pairwise relationship on a dataset. The following Figs. 5 and 6 represent the scatterplot and pair plot, respectively.

The correlation is used to find potential relationships between different features and shows the strength of these relationships, where green color represents positive and blue color represents negative. Some variables' relationship like radius-mean is strongly related to perimeter_mean, area_mean, radius_worst, perimeter_wrost, and area_wrost; texutre_mean has also strong relations with texture_wrost; perimeter_mean has also good relations with radius_mean, area_mean,radius_wrost, area_wrost, smoothness_wrost. These all feature's relationships are more than 90%. Fig. 7 depicts the correlation matrix of the breast cancer dataset.

3.5. Machine learning classifiers

Classification plays a vital role in machine learning techniques used worldwide to diagnose any disease. This section will briefly discuss the working procedures of SVM, RF, K-NN, DT, NB, LR, AB, GB, MLP, NCC, and VC classifiers, and among all classifiers, we investigate the best-suited classification model comparatively.

3.5.1. Support Vector Machine

SVM is the supervised machine learning algorithm used for classification, regression, and outlier detection. In this Algorithm, each data item is plotted in n-dimensional space. Then we have to employ classification to find out the best decision boundary among multiple lines or decision boundaries. The obtained best border is called the hyperplane of SVM. The main motto of SVM is to find an optimal hyperplane that separates the training data into two classes.

SVM can be stated mathematically as follows: the training data for the two classes is first stacked into an n a p × q matrix X. The number of variables is denoted by q, while the number of observations is denoted by p. x_i is used to represent the ith row of X. A diagonal p × p matrix Y with -1 and +1 specifies whether each x_i is in class +1 or 1. The major problem in SVM is to use a hyperplane to divide a collection of training vectors into two different groups [24].

$$D = \{(x^1, y^1), \ldots, (x^l, y^l)\}, \; x \in \mathbb{R}, \; y \in \{-1, 1\} \tag{3}$$

The hyperplane is used to have divided the set of vectors optimally if it did so without error and the distance between the vectors closest to the hyperplane was the shortest.

3.5.2. Random Forest

This approach generates a vast number of decision trees that are linked together. In this method, decision trees serve as pillars. A "random forest" is a collection of decision trees created during the pre-processing stage [25]. The best feature is chosen from a random selection of features after numerous trees have been constructed. The random forest

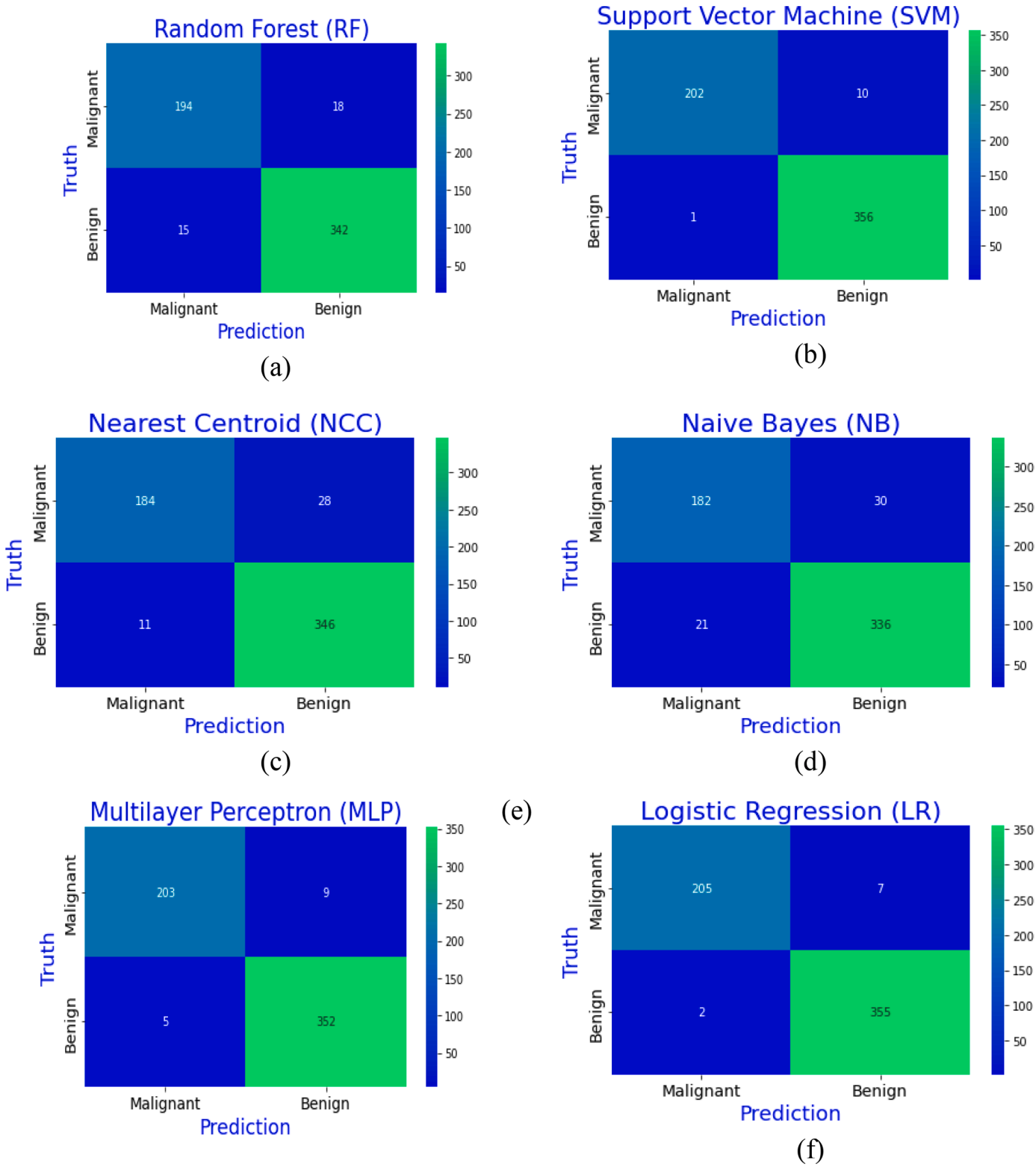


Fig. 8. Confusion matrix for(a) Random Forest (b) Support Vector Machine (c) Nearest Centroid (d) Naïve Bayes (e) Multilayer perceptron (f) Logistic Regression (g) Decision Tree (h) AdaBoost (i) K-Nearest Neighbor (i) Gradient Boosting (k) Voting Classifier

classifier's mathematical formula is given in Eq. (4).

$$P_{ij} = m_i T_j - m_{left(j)} T_{left(j)} - m_{right(j)} T_{right(j)} \quad (4)$$

Where, *Pi sub* (j) = node j's significance
m sub (j) = samples arriving to node j
T sub (j)= the value of node j's impurity
left (j) =node j splits a child node from the left
right (j) =node j splits a child node from the right

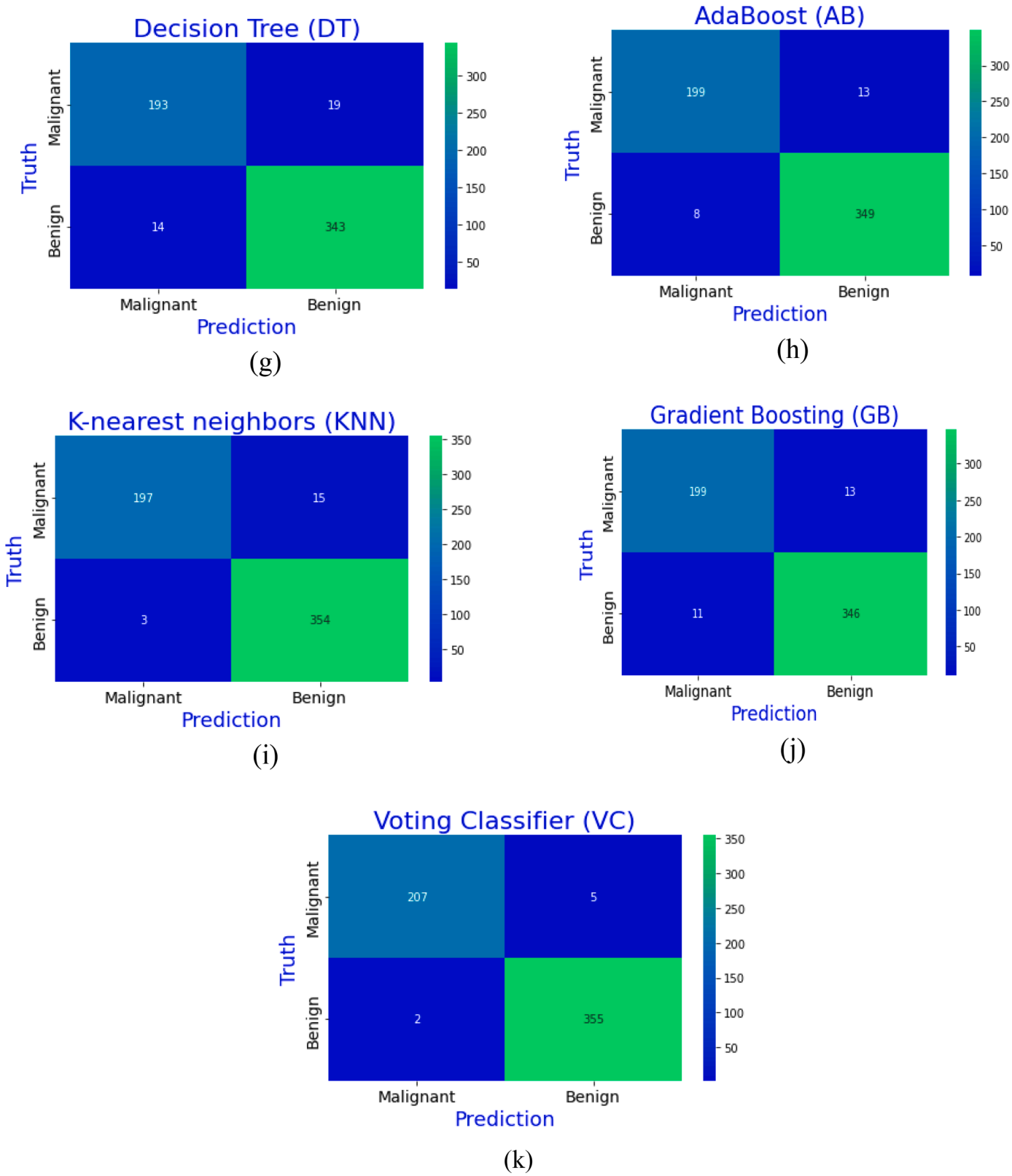


Fig. 8. (*continued*).

3.5.3. *K-Nearest Neighbor*

K-Nearest Neighbor is a straightforward machine learning algorithm, easy to implement, more effective for extensive training data, and mainly used for classification and regression based on supervised technique. It is a non -parametric classifier that uses feature similarity. New data points values are predicted by feature similarity. The latest data points are assigned a discount depending on the matching points, which are closer in the training dataset [25].This classifier evaluates the distance using different formula. A popular distance formula is given below-

Table 5
Evaluation of classification methods before feature optimization.

Methods	Precision	Recall	F_Measure	Accuracy	Error
SVM	94.12 %	88.68 %	90.49 %	91.61 %	0.083 %
RF	96.88 %	94.34 %	95.39 %	95.80 %	0.042 %
KNN	94.34 %	93.62 %	93.40 %	94.41 %	0.055 %
DT	92.92 %	93.84 %	93.33 %	93.71 %	0.062 %
NB	94.23 %	92.29 %	93.11 %	93.71 %	0.062 %
LR	95.50 %	95.50 %	95.50 %	95.80 %	0.042 %
AB	97.39 %	96.61 %	96.98 %	97.20 %	0.028 %
GB	98.39 %	97.17 %	97.72 %	97.90 %	0.021 %
MLP	95.29 %	92.84 %	93.85 %	94.41 %	0.055 %
NCC	92.86 %	85.85 %	87.91 %	89.51 %	0.104 %
VC (LR+SVM)	96.10 %	96.45 %	96.27 %	96.50 %	0.035 %

Table 6
Evaluation of classification methods after feature optimization.

Methods	Precision	Recall	F_Measure	Accuracy	Error
SVM	98.28 %	97.61 %	97.92 %	98.07 %	0.019 %
RF	93.91 %	93.65 %	93.78 %	94.20 %	0.058 %
KNN	97.22 %	96.04 %	96.58 %	96.84 %	0.061 %
DT	93.99 %	93.56 %	93.77 %	94.20 %	0.058 %
NB	90.73 %	89.98 %	90.33 %	91.04 %	0.089 %
LR	98.55 %	98.07 %	98.30 %	98.42 %	0.015 %
AB	96.27 %	95.81 %	96.03 %	96.31 %	0.036 %
GB	95.57 %	95.39 %	95.48 %	95.78 %	0.042 %
MLP	97.55 %	97.18 %	97.36 %	97.54 %	0.024 %
NCC	93.44 %	91.86 %	92.54 %	93.15 %	0.068 %
VC (LR+SVM)	**98.83 %**	**98.54 %**	**98.68 %**	**98.77 %**	**0.012 %**

$$Euclidean = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{5}$$

3.5.4. Decision Tree

The input characteristics are indicated by the nodes of a decision tree (DT). The height information collected for that particular feature is then used to project the outcome. The characteristics that were not utilized in the preceding steps make up the subtree. Internal nodes represent input components, branching represent outputs, and leaf nodes represent classes in DT [26]. It is made up of several layers of nodes, with the root at the top and the leaves at the bottom. Eq. (6) calculates the entropy of each class, whereas Eq. (7) calculates the entropy using two characteristics.

$$H(S) = \sum_{i=1}^{c} -P_i log_2 P_i \tag{6}$$

$$H(S, D) = \sum_{c \in D} P(c).E(c) \tag{7}$$

3.5.5. Naïve Bayes

It is a supervised machine learning algorithm based on the Bayes theorem, assuming each input variable independently [27]. This classifier is very easy to implement on many complex problems. Eq. (8) is used to map this classifier, Where P(M|N) is the posterior probability, P (N|M)is the likelihood. P(M) is called prior probability or the initial probability. P(N) is the prior probability of the input instance.

$$P(M|N) = \frac{P(N|M) \times P(M)}{P(N)} \tag{8}$$

3.5.6. Logistic Regression

Because the purpose of both is to estimate the values of the parameter's coefficients, logistic regression and linear regression are quite similar. Instead of utilizing linear regression, the outcome prediction is altered using a nonlinear function known as the logistic function. For estimating logistic function coefficients, gradient descent can be utilized [28]. Logical regression equation is shown in Eq. (9). Here, the expected output is y, the bias is b0, and the single input value x has a coefficient of b_1.

$$y = \frac{e^{(b0+b1 \times x)}}{1 + e^{(b0+b1 \times x)}} \tag{9}$$

3.5.7. Adaptive Boosting

Adaptive boosting is also called AdaBoost, used as an ensemble method that transforms weak predictors into strong predictors [29,30]. The Eq. (10) represents AdaBoost classification approach. Where m is the corresponding weight and f_m denominate the mth weak classifier.

$$F(x) = \text{sign}\left(\sum_{m=1}^{m} \theta_m f_m(x)\right) \tag{10}$$

3.5.8. Gradient Boosting

Gradient boosting is based on the concept of seeing the boosting process as a set of cost functions that must be optimized. As a consequence, the loss function reduction gradient descent paths may be turned into decision trees and used to enhance the classifier sequentially [31].

Given a training dataset containing each data point $((x, y) \sim D$, the purpose of gradient boosting is to learn a function $\widehat{F}$ from all plausible hypotheses H while minimizing the expectation of loss over the distribution D. Eq. (11) aids in the calculation of $\widehat{F}$ [32].

$$\widehat{F} = \underset{F \in \mathscr{H}}{\text{argmin}} \underset{(x,Y) \sim D}{E} L(y, F(x)) \tag{11}$$

3.5.9. Multi-layer perceptron

A multilayer perceptron (MLP) is a type of neural network that generates a separate vector for each pair of inputs. There are three levels in the MLP structure: an output layer, an input layer, and a hidden layer. Eq. (12) represents MLP's input layer.

$$a^x = x \tag{12}$$

In Eq. (12), x is the input and a^1 is the first layer of the network, and each layer's input is the weighted output of the preceding layer [33], which corresponds to Eq. (13).

$$\left(a^{(l+1)} = \Psi w^{(l)} a^{(l)} + b^{(l)}\right. \tag{13}$$

If l is a specific layer, $w^{(l)}$signifies layer weight,$b^{(l)}$denotes the bias symbol in layer l, and Ψ denotes a nonlinear function used in this network. It's possible that this function is a sigmoid function, a hyperbolic tangent, or something else.

3.5.10. Nearest Centroid

The Nearest Centroid classifier involves observations of the class of training samples' labels using the centroids near the statement [34].The training procedure is performed by labeling $\{(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_1)\}$.

Which class labels $y_i \Sigma$ Y and then computes the per class centroid using the Eq. (14).

$$\overrightarrow{\mu}_l = \frac{1}{|C_l|} \sum_{i \in C_l} \overrightarrow{x}_i \tag{14}$$

Where c_l representsthe set of indices of samples belonging to class $l \in Y$.Finally, the class is assigned to observation for predicting by the following function given in Eq. (15).

$$\overrightarrow{x} \text{ is } \widehat{y} = arg\ min_{l \in} Y \left|\left| \overrightarrow{\mu}_l - \overrightarrow{x} \right|\right| \tag{15}$$

3.5.11. Voting Classifier

A Voting Classifier is an ensemble method of numerous models and

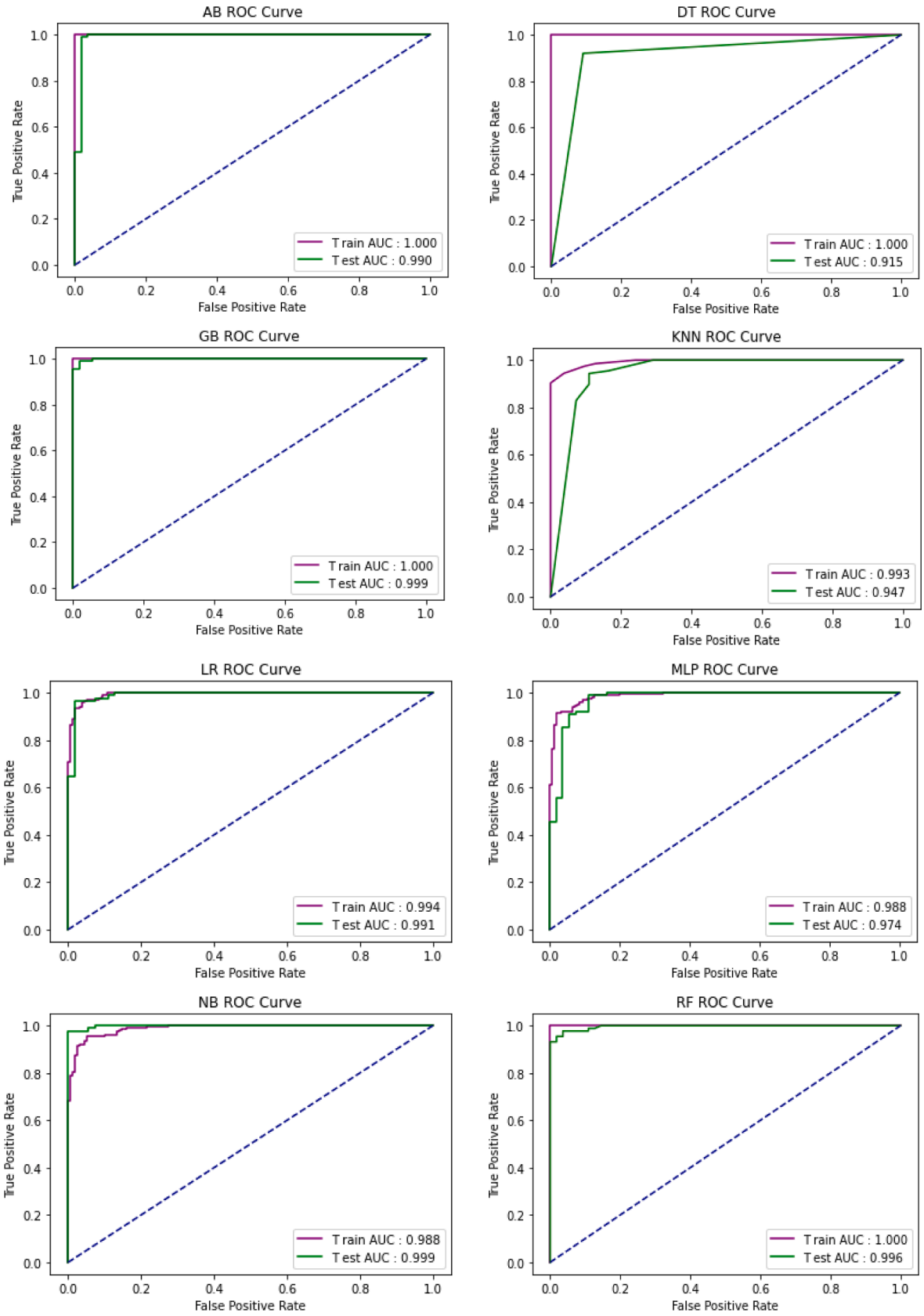


Fig. 9. ROC curve (a) Before optimization (b) After optimization

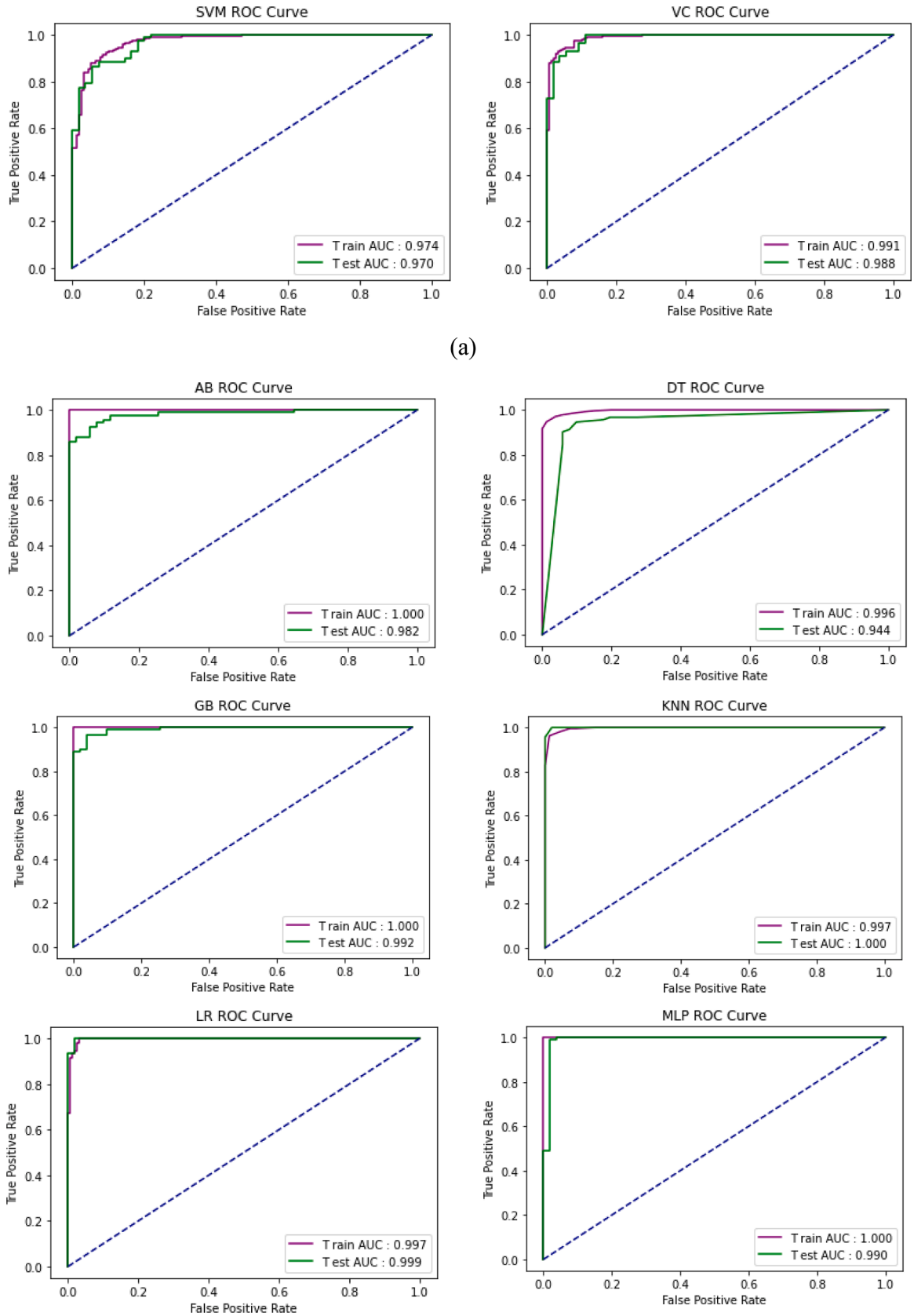
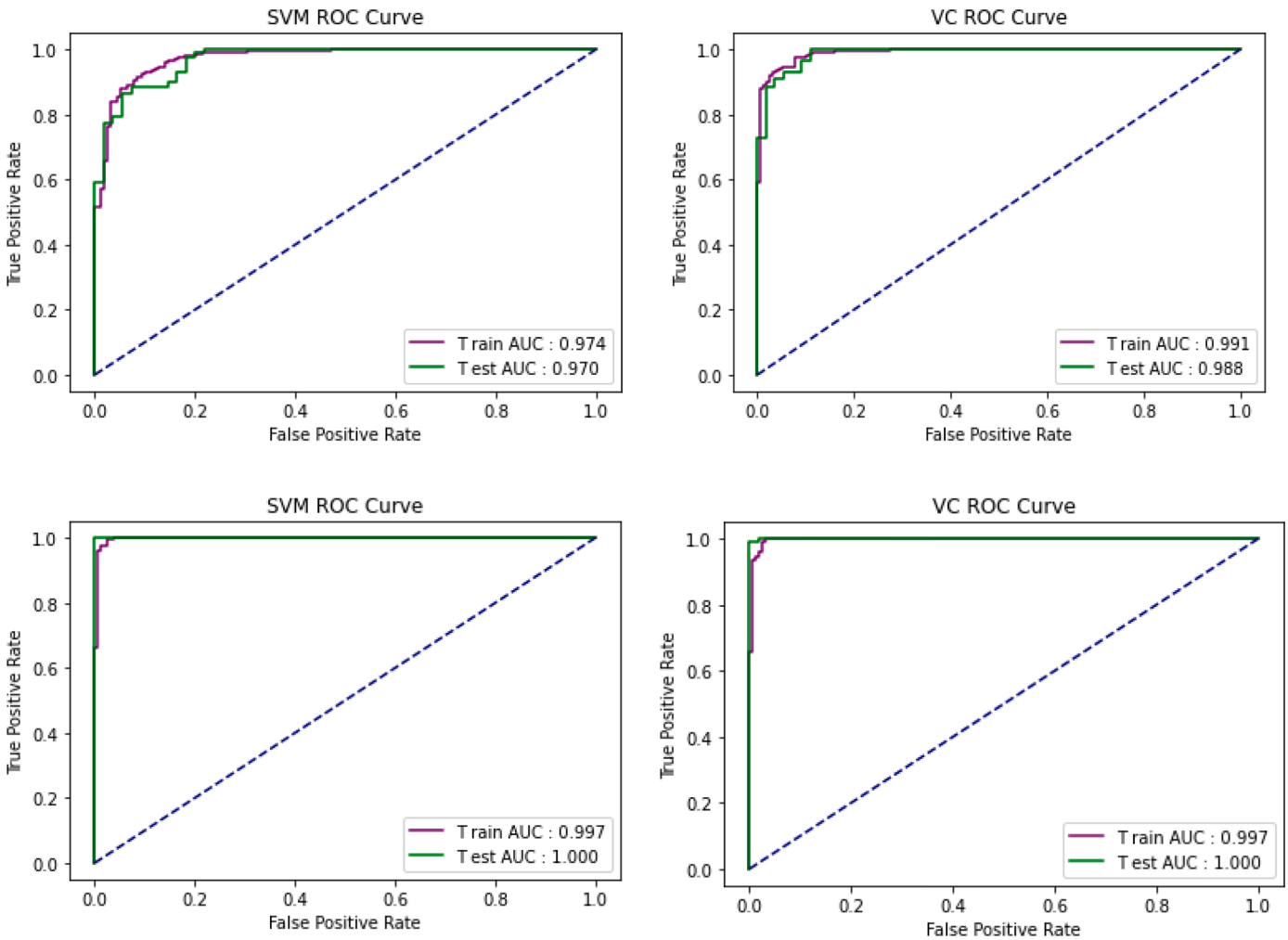


Fig. 9. (*continued*).



(b)

Fig. 9. (*continued*).

Table 7
Comparison of our work with the most related works.

Author	Dataset	Method	Accuracy
Jabbar et al. 2020 [12]	WBCD, 699 instances, 11 features	BN + RBF	97.42%
Gopal et al. 2021 [13]	UCI, 569 instances, 32 features	MLP	98.0%
Ara et al. 2021 [14]	UCI, 569 instances, 32 features	SVM	96.5%
Varshini et al. 2021 [15]	UCI, 569 instances, 32 features	RF	96.5%
Chaurasia et al. 2021 [16]	UCI, 569 instances, 32 features	RF	97.66%
Subramanian et al. 2022 [17]	UCI, 569 instances, 32 features	RF	96.59%
Assegie et al. 2021 [18]	Kaggle, 569 instances, 31 features	DT	92.53%
Acquisition et al. 2019 [19]	UCI, 699 instances, 10 features	KNN	96.85%
Mangukiya et al. 2022 [20]	UCI, 569 instances, 32 features	XGboost	98.24%
Assegie et al. 2020 [21]	UCI, 569 instances, 32 features	KNN	94.35%
Comission et al., 2019 [22]	WBCD, 699 instances, 10 features	SVM	96.72 %
Proposed	**WBCD, 569 instances, 30 features**	**VC (LR+SVM)**	**98.77%**

predicts the class with the highest probability of the chosen category. There are mainly two types of classifiers. They are Hard and soft voting. In this study, hard voting [35,36] is used. Hard voting is calculated when the predicted outputs class has the majority of votes. The formula of voting classifier is shown in Eq. (16).

$$\sim a = mode\left(N_c\left(y_t{}^1\right), N_c\left(y_t{}^2\right), \ldots, N_c\left(y_t{}^n\right)\right) \tag{16}$$

Where $N_c(y_t)$ is the highest number of votes.

3.6. Environment Setup

We have used different types of environments for completing this research work. Table 4 represents the environment setup of this proposed work.

4. Result and discussion

This section describes the result of the developing system. The experimental result, confusion matrix for each classifier, and accuracy will be briefly described here.

4.1. Confusion Matrix and Performance Analysis

A confusion matrix is a table that measures the predictive

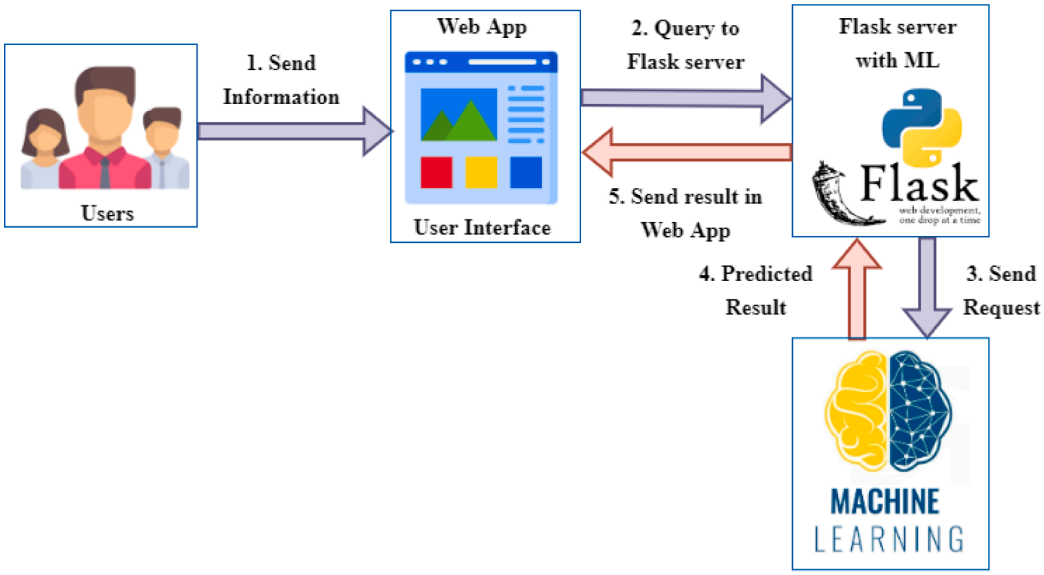


Fig. 10. Working flow of the web application

performance of a classification problem. This study used 11 classifiers to predict breast cancer, and each classifier showed mixed results. Using the confusion matrix, we compared each class with the other class and then observed how many samples were classified incorrectly. The performance of these models using the confusion matrix is shown in Fig. 8.

In addition, the performance of the available ML classifiers were evaluated using a range of assessment parameters. Eqs. (17) - (21)) illustrate the steps for calculating various Confusion Matrix assessment metrics. The accuracy (AC), precision (P), recall (R), and F1-score are obtained by computing correctly classified samples (True positive (TP) and True Negative (TN)) and incorrectly classified samples (False positive (FP) and False Negative (FN)).

$$\text{Accuracy}\ (\%) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \tag{17}$$

$$\text{Precision}\ (\%) = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{18}$$

$$\text{Recall}\ (\%) = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{19}$$

$$\text{F} - \text{Measure}\ (\%) = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \times 100 \tag{20}$$

$$\text{Error}\ (\%) = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \tag{21}$$

Using these performance metrics, the score of all proposed classifiershave been measured before feature optimization and after feature optimization, shown in Tables 5 and 6, respectively. Table 5 ensures that the Voting classifier gives the highest accuracy with 96.50% with 0.035% error without any optimization. However, after optimization, it reaches 96.50% to 98.77% with a 0.012% error. After observing these two tables this research ensure that optimization techniques enhances some accuracy in terms of detection. The AUC-ROC curve, which is depicted in Fig. 9, enables us to see how well our machine learning classifier is functioning. Basically, a binary classifier's sensitivity and specificity are examined and used in ROC analysis. ROC is a probability curve, and AUC is a measure of separability's degree [37]. It indicates the degree to which the model can distinguish between classes.

The model will be more effective at differentiating between patients with disease and those who do not have it if the AUC is higher. TPR and FPR are shown on the y-axis and x-axis, respectively, to represent the ROC curve. The classifier can accurately discriminate between all Positive and Negative class points when AUC = 1 [38]. The classifier would have predicted that all positives were negatives and all negatives were positives if the AUC had been 0. There is a good possibility that the classifier will be able to tell the difference between the positive class values and the negative class values when 0.5<AUC<1. This is the case because more True positives and True negatives can be detected by the classifier than False positives and False negatives. The classifier is unable to discriminate between Positive and Negative class points when the AUC value is less than 0.5. Consequently, either the classifier predicts a random class or a fixed class for each data point. Therefore, the better a classifier is in differentiating between positive and negative classifications, the greater its AUC score. This is how the ROC curve for the classifier mentioned above operates.

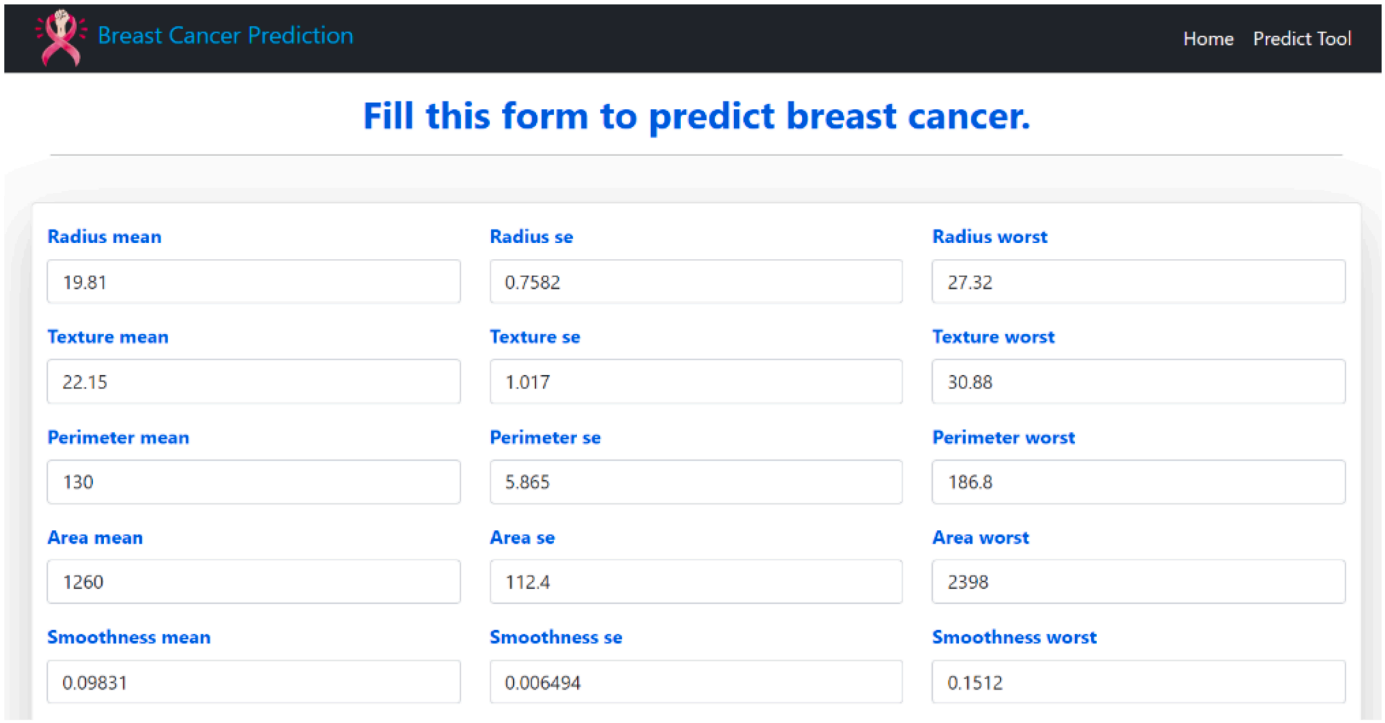
However, this study also investigates the K-fold cross-validation after hyperparameter tuning. Using cross-validation, we estimate the performance of a model on unseen data. In K-fold, K means how many splits we want to do in the given set. This study has applied the most famous value 10 for K to evaluate these models. It is also observed that, after cross-validation, it is estimated the maximum promising result by using the performance metrics. Thus, the best model is selected among all these classifiers.

4.2. Comparison with others Works

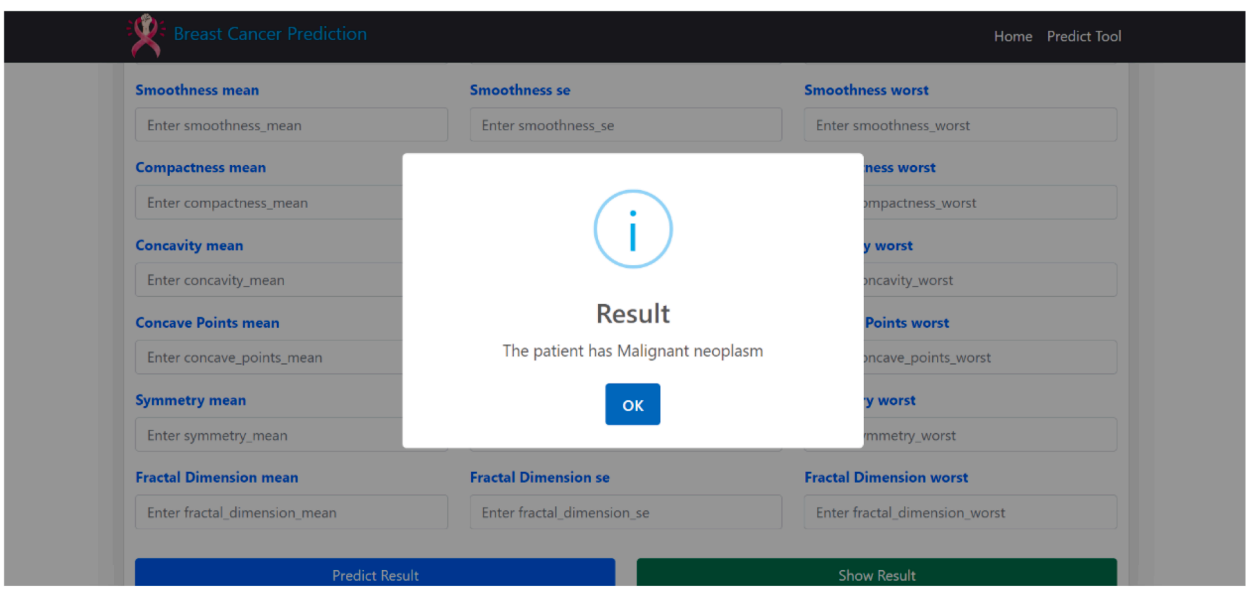
However, during this study, many researchers have worked on breast cancer in recent years and proposed different techniques and methodologies that showthe diverse and perplexing results. As a result, we have been motivated to work with this disease to overcome limitations. Table 7 shows the comparative analysis that confirms the accuracy of our proposed methodology is higher than any other research work that has been mentioned in this table.For showing the comparison, this study only considered the similar dataset of WBCD and the available machine learning approaches authors employed in their research. Though some of the studies produced slightly higher accuracy than our results, still in terms of feature optimization and the use of relevant features, our proposed model VC(LR+SMV) achieved significant results in terms of others.

4.3. Implementation of Web application

This section describes the web application for predicting the breast



(a)

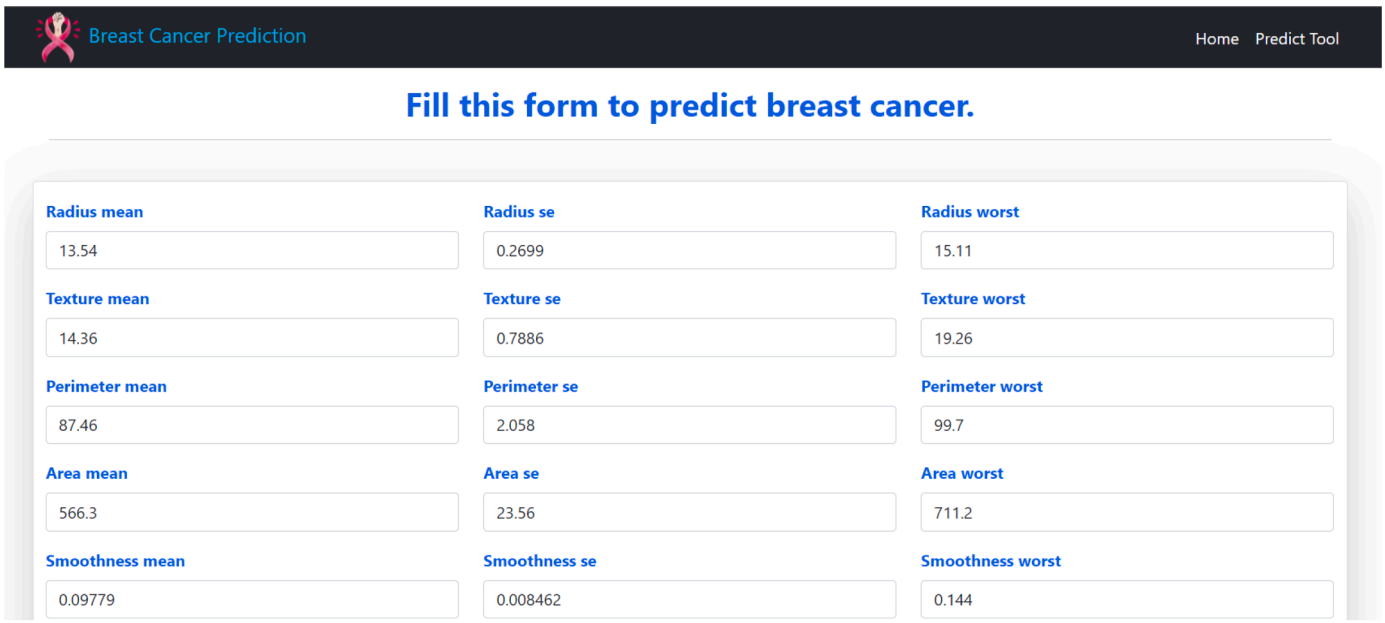


(b)

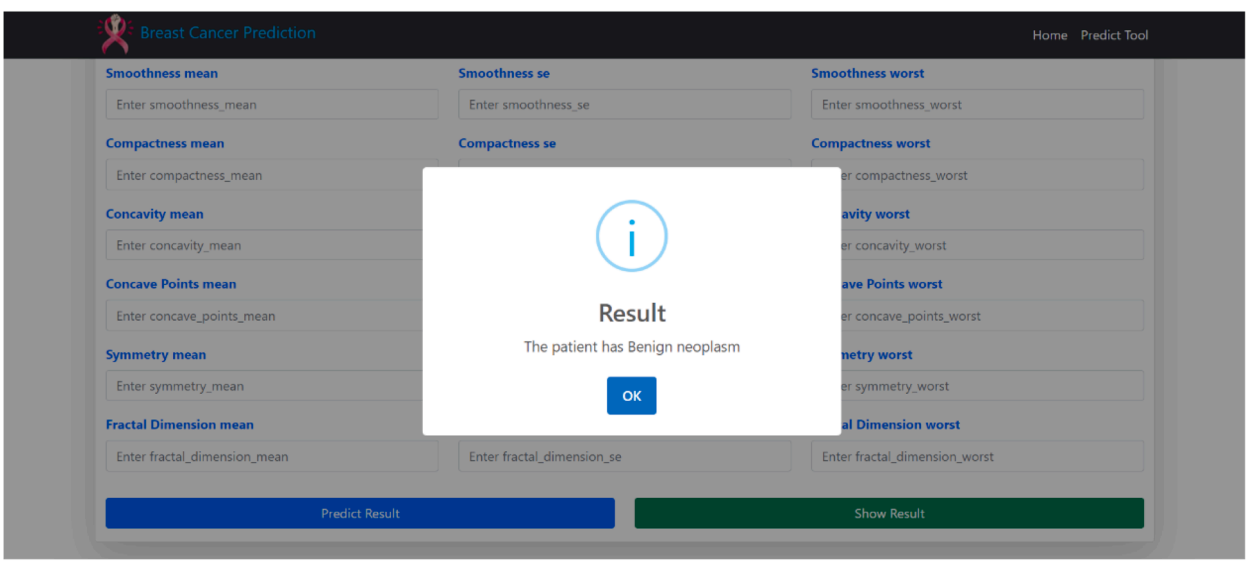
Fig. 11. Prediction result of Malignant (a) input form of a user and (b) predict the result.

cancer. The react is used for developing this web app. The models are trained on the jupyter notebook and then a pklfile is built. This pkl file with machine learning model is deployed in Heroku cloud storage which generates an api using flask application. Working flow of the web application is shown in Fig. 10.

The flask is developed using python code which makes a gorge between the web page and machine learning model. Anyone can predict this disease submitting with the necessary input perameters in the input form. The given inputs are sent to the flask and flask sends this request to the machine learning models to predict the result. After predicting the stroke, the result is sent to the flask and flask resends this result to the web. Thus, a user can show the result based on given inputs. For malignant neoplasms, the input form is shown in Fig. 11(a), and the predicted result is shown in Fig. 11(b). For benign neoplasms, the input form is shown in Fig. 12(a), and the predicted result is shown in Fig. 12 (b).



(a)



(b)

Fig. 12. Prediction result of Benign (a) input form of a user and (b) predict the result.

5. Conclusion

Machine learning is now most prevalent in the medical field, where machine learning models are readily employed in numerous large datasets. Nowadays, the most frequent cancer worldwide is breast cancer, which takes away valuable lives in the early stage of life. But early detection of breast cancer can reduce the mortality rate and save valuable lives. Therefore, a real time smart web application is built based on best machine learning models from which any user can predict the breast cancer giving some inputs and can save the expensive cost and time. The proposed method of this paper presents different machine learning models for the prediction of breast cancer. Finally from the proposed model, it is obtained that the voting classifier gives the highest accuracy with 96.50% with 0.035% error without any optimization. However, after optimization, it reaches 96.50% to 98.77% with a 0.012% error. To define the model's accuracy, transparency, fairness, and results, this study may be further expanded for the classification of breast cancer utilizing an explainable machine learning approach. Also, this research can be extended further utilizing a self-developed dataset.

Declaration

Author Contribution

KMMU, NB, and SKD had the idea for and designed the study, had full access to all the data in the study, and took responsibility for the data and accuracy of the model generation. All the authors contributed to the writing of the article. STR, and NB contributed to the critical revision of the report. NB, and KMMU developed all the results and data presentation methods. All authors contributed to data acquisition and analysis and reviewed and approved the final version.

Data Availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Funding

None

Ethical Approval

Not required

Consent to participate

Not required

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] W Yue, Z Wang, H Chen, A Payne, X. Liu, Machine learning with application in breast cancer diagnosis and prognosis, Designs 2 (2018) 1–17.
[2] LA Torre, F Bray, RL Siegel, Lortet-TieulentJ FerlayJ, A Jemal, A global cancer statistics 2012, CA Cancer J. Clin. 65 (2) (2015) 87–108.
[3] V. Chaurasia, S. Pal, Applications of machine learning techniques to predict diagnostic breast cancer, SN Comput. Sci. 1 (5) (2020) 1–11.
[4] J. Sivapriya, A. Kumar, S. Siddarth Sai, S. Sriram, Breast cancer prediction using machine learning, Int. J. Recent Technol. Eng. (IJRTE) 8 (2019).
[5] CE DeSantis, F Bray, J Ferlay, J Lortet-Tieulent, BO Anderson, A Jemal, International variation in female breast cancer incidence and mortality rates, Cancer Epidemiol. Biomark. Prev. 24 (10) (2015) 1495–1506.
[6] O Ginsburg, CH Yip, A Brooks, A Cabanes, M Caleffi, JA Dunstan Yataco, et al., Breast cancer early detection: A phased approach to implementation, Cancer 126 (Suppl 10) (2020) 2379–2393. http://www.ncbi.nlm.nih.gov/pubmed/32348566.
[7] M Mutebi, BO Anderson, C Duggan, C Adebamowo, G Agarwal, Z Ali, et al., Breast cancer treatment: A phased approach to implementation, Cancer 126 (Suppl 10) (2020) 2365–2378. http://www.ncbi.nlm.nih.gov/pubmed/32348571.
[8] N. Biswas, K.M.M. Uddin, S.T. Rikta, S.K. Dey, A comparative analysis of machine learning classifiers for stroke prediction: a predictive analytics approach, Healthc. Analyt. 2 (2022), 100116.
[9] S.K. Dey, M.M. Rahman, A. Howlader, U.R. Siddiqi, K.M.M. Uddin, R. Borhan, E. U. Rahman, Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: a machine learning approach, PLoS One 17 (7) (2022), e0270933.
[10] K. Doi, Computer-aided diagnosis in medical imaging: historical review, current status and future potential, Comput. Med. Imaging Graph. 31 (4-5) (2007) 198–211.
[11] R. Aggrawal, S. Pal, Sequential feature selection and machine learning algorithmbased patient's death events prediction and diagnosis in heart disease, SN Comput. Sci. 1 (6) (2020) 1–16.
[12] M.A. Jabbar, Breast cancer data classification using ensemble machine learning, Eng. Appl. Sci. Res. 48 (1) (2021) 65–72, https://doi.org/10.14456/easr.2021.8.
[13] V.N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, M. Rajesh, Feature selection and classification in breast cancer prediction using IoT and machine learning, Meas. J. Int. Meas. Confed. 178 (October 2020) (2021), 109442, https://doi.org/10.1016/j.measurement.2021.109442.
[14] S. Ara, A. Das, A. Dey, Malignant and Benign breast cancer classification using machine learning algorithms, in: 2021 Int. Conf. Artif. Intell. ICAI 2021, no. June, 2021, pp. 97–101, https://doi.org/10.1109/ICAI52203.2021.9445249.
[15] K. Varshini, R.K. Sethuramamoorthy, V. Kumar, S.A. Shree, S. Deivarani, Breast cancer prediction using machine learning techniques, Int. J. Adv. Sci. Technol. 29 (6 Special Issue) (2020) 2026–2032, https://doi.org/10.26483/ijarcs.v10i5.6464.
[16] V. Chaurasia, M. Pandey, S. Pal, Prediction of presence of breast cancer disease in the patient using machine learning algorithms and SFS, IOP Conf. Ser. Mater. Sci. Eng. 1099 (1) (2021), 012003, https://doi.org/10.1088/1757-899x/1099/1/012003.
[17] D. Subramanian, "Breast Cancer Prediction using Some Machine Learning Models by Dimensionality Reduction of Various Features Breast Cancer Prediction using Some Machine Learning Models by Dimensionality Reduction of Various Features," no. February, 2022.
[18] T.A. Assegie, R.L. Tulasi, N.K. Kumar, Breast cancer prediction model with decision tree and adaptive boosting, IAES Int. J. Artif. Intell. 10 (1) (2021) 184–190, https://doi.org/10.11591/ijai.v10.i1.pp184-190.
[19] A. D. Acquisition, "Classification of Breast Cancer Data Using Machine Learning Algorithms," pp. 462–465, 2019.
[20] M. Mangukiya, A. Vaghani, and M. Savani, "Breast Cancer Detection with Machine Learning February 2022," no. February, 2022, doi: 10.22214/ijraset.2022.40204.
[21] T.A. Assegie, An optimized K-nearest neighbor based breast cancer detection, J. Robot. Control 2 (3) (2021) 115–118, https://doi.org/10.18196/jrc.2363.
[22] E. Commission, M. T. Islam, and E. Commission, "Performance Comparison of Three Classifiers for the Classification of Breast Cancer Dataset," no. December, 2019, doi: 10.1109/EICT48899.2019.9068816.
[23] S. Cateni, et al., Variable selection and feature extraction through artificial intelligence techniques, Multivar. Analy. Manag., Eng. Sci. (2012) 103–118, chapter 6.
[24] S.K. Dey, K.M.M. Uddin, H.M.H. Babu, M.M. Rahman, A. Howlader, K.A. Uddin, Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease, Intell. Syst. Applic. 16 (2022), 200144.
[25] A.A. Imran, M.N. Amin, F.T. Johora, Classification of chronic kidney disease using logistic regression, feedforward neural network and wide amp; deep learning, in: 2018 International Conference on Innovation in Engineering and Technology (ICIET), 2018, pp. 1–6.
[26] Y. Zhang, Support vector machine classification algorithm and its application, in: C. Liu, L. Wang, A. Yang (Eds.), Information Computing and Applications, Springer, Berlin, Heidelberg, 2012, pp. 179–186.
[27] T.H.H. Aldhyani, A.S. Alshebami, M.Y. Alzahrani, Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms, J. Healthc. Eng. 2020 (2020), e4984967.
[28] S. Dreiseitl, & L. Ohno-Machado, Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inform. 35 (5-6) (2002) 352–359.
[29] D.D. Margineantu, T.G. Dietterich, Pruning adaptive boosting, ICML 97 (1997) 211–218.
[30] D.C. Feng, Z.T. Liu, X.D. Wang, Y. Chen, J.Q. Chang, D.F. Wei, Z.M. Jiang, Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach, Constr. Build. Mater. 230 (2020), 117000.
[31] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140, https://doi.org/10.1007/BF00058655.
[32] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. 29 (2001) 1189–1232.
[33] L. Zhao, S. Lee, S.-P. Jeong, Decision tree application to classification problems with boosting algorithm, Electronics 10 (2021) 1903.
[34] M. Thulasidas, Nearest centroid: a bridge between statistics and machine learning, in: 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), IEEE, 2020, pp. 9–16.
[35] A Mahabub, MI Mahmud, MF Hossain, A robust system for message filtering using an ensemble machine learning supervised approach, ICIC Expr. Lett. Part B Appl. 10 (2019) 805–811, https://doi.org/10.24507/icicelb.10.09.805.
[36] A Mahabub, AZSB Habib, A voting approach of modulation classification for wireless network, in: Proceedings of the 6th international conference on networking, systems and security, ACM, 2019, pp. 133–138.
[37] P. A. R. O. C. Analysis, I. C. Sammut, and G. I. Webb, "(Eds.), Encyclopedia of Machine Learning and Data Mining (pp . 1-8)." pp. 1–8, 2016.
[38] Sarang Narkhede, Understanding AUC - ROC Curve, Towar. Data Sci. (2019) 6–11 [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.