Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 6s (2024)

Improved Methodology for Breast Cancer Prediction through Integration of Hard Voting Ensemble Classifier on WDBC Data Set

Archana Singh[1], Kuldeep Singh Kaswan[2]

[1,2]Department of Computer Science & Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India.

Article History:

Received: *30-05-2024*

Revised: *20-07-2024*

Accepted: *02-08-2024*

Abstract:

Introduction: Disease that is prevalent and highly fatal is the breast cancer disease and it affects many people in the world. To effectively prevent the fatality rate caused by breast cancer, tools need to be developed that are capable of early diagnosis and efficient treatment. Researchers and medical experts across the world have pointed out several diagnostic techniques for this sickness; however, higher enhancement of such present methods is still needed to enhance a perfect and effective diagnosis of this disease.

Objective: It's an objective of this research to establish quick and precise forecasts of breast cancer, which is estimated to rank second as a leading killer of women globally.

Methodology: In this paper, we provide a methodology based on hard voting ensemble classifier that combines three machine learning algorithms: logistic regression, support vector machine and decision tree to diagnose the kind of breast cancer, whether benign or malignant. The proposed model's performance is evaluated in this study using the Wisconsin Diagnostic Breast Cancer dataset (WDBC), with random oversampling (ROS) being used to balance the dataset and Standard Scaler being used for feature scaling.

Results: The suggested approach achieved an accuracy of 0.9825, a precision of 0.9859, a recall of 0.9859, F1 score of 0.9859and AUC of .9813. Using a 10-fold cross validation it obtained a mean accuracy of .9738.

Conclusions: The suggested approach yielded superior results after the individual classifier and many acknowledged existing works are directly compared with the results.

Keywords: breast cancer, machine learning, prediction, WDBC, hard voting classifier

1. Introduction

The most prevalent disease in women to be diagnosed with and the leading cause of cancer-related deaths is breast cancer (BC). In 2020, 2.3 million women received a breast cancer diagnosis, and 685,000 women died from this disease, according to the WHO. The "World Health Organization" predicts that 963,000 women will lose their lives to breast cancer globally in 2021.Around the globe females get breast cancer at any age after adolescence while the incidence rises with age. The most significant risk factor to breast cancer is female gender. Men are influenced by breast cancer in a range of 0.5–1% and the care of breast cancer in males is based on the same concepts as in women [1]. Breast cancer have several stages each characterized by the extent of the disease and the spread of cancer cells. Table 1 underscores the various stages of breast cancer with description. According to WHO

information the breast cancer deaths averted by 2.5 million from predicted deaths in 20 years (by 2040). This global initiative is taken by world health organization to reduce breast cancer deaths rate [1]. To achieve this initiative early identification and accurate diagnosis of BC is most essential to improve patient outcomes and reduce mortality rate.

Table 1 Stages and classification of breast Cancer

Stages	Description	Classification
Stage 0	Breast milk duct-confined non-invasive cancer cells are known as ductal carcinoma in situ (DCIS).	Malignant
Stage I	Early-stage cancer where the tumour size is small and has not spread outside the breast tissue.	Malignant
Stage II	The cancer has grown larger and may have spread to lymph nodes in the vicinity, but it has not yet reached distant organs.	Malignant
Stage III	Locally advanced cancer that has spread to nearby lymph nodes and tissues but not too distant organs.	Malignant
Stage IV	Advanced cancer that has metastasized to distant organs such as the bones, lungs, liver and brain.	Malignant
Benign	Abnormal cells or growths that are non-cancerous and not spreading to the body's other regions.	Benign

Traditional diagnostic techniques are susceptible to human error. Machine learning (ML) approaches have been used into breast cancer prediction with hopeful results in recent years providing accurate and efficient tools for diagnosis and risk assessment. By leveraging advanced algorithms and computational techniques ML algorithms can analyse vast amounts of patient data including imaging scans, genetic markers, and clinical records, to identify patterns and relationships that may not be apparent to human observers. These algorithms have the potential to improve early detection rates leading to more timely interventions and better patient outcomes. Ongoing research in ML continues to refine and enhance predictive models enabling the integration of emerging technologies such as deep learning and ensemble methods to further enhance accuracy and reliability.

In this paper a hard voting-based classification model is purposed to predict the breast cancer. Standard metrics can be used to assess the performance of the three chosen classification methods, logistic regression (LR), support vector machine (SVM), and decision tree (DT), in terms of the class label decisions they make. Wisconsin Breast Cancer Dataset (WDBC) used in this work which is the publicly available on “UCI machine learning repository”.

2. Literature Review

Machine learning techniques developed to study breast cancer have attracted numerous investigational and clinical areas, it follows that a critical assessment is required. This section first gives a short overview of the earlier studies that are relevant to this experiment in order to set up the study. Jakhar et al., (2023) [2] proposed a hybrid stack-based ensemble learning framework named SELF for identifying BC and achieving an accuracy of 98.80%. Talatian et al. (2021) [3] presented an intelligent ensemble classification method that achieved 98.74% accuracy by utilizing both evolutionary algorithms and multi-layer perceptron-based neural networks.

The Naseem et al. (2022) [4] conducted an automatic detection framework for classifying and prognosis of breast cancer. They used an ensemble of classifiers with an accuracy of 98.83%. Srinivas et al., (2022) [5] evaluated a new classifier system for use in the detection of breast cancer which outperformed the other known classifiers by a high degree of accurate classifications with a rate of 98.50%. Jabbar et al. (2021) [6] utilized ml ensemble models for classification of breast cancer dataset and they achieved an accuracy of 97.42%. Alhayali et al., (2020) [7] served a relevant way based an ensemble Hofeding tree and Naïve Bayes classifiers and an accuracy of 95.99 % was achieved.

Batool & Byun (2024) [8] set out for application of adaptive voting ensemble learning algorithm to identify breast cancer with the highest accuracy of 97.60%. Chaurasia et al. (2018) [9] focused on data mining tools for diagnosing breast cancer benign and malignant with the accuracy value of 96%. Hashim & Yassin (2023) [10] used soft voting classifier based on ML models have been provided for breast cancer prediction and which is able to give up to 99.30% accuracy.

Sharma et al. (2024) [11] proposed an ensemble framework for BC prediction with 97.66% accuracy value. Anastraj & Chakravarthy (2019) [12] predict breast cancer using backpropagation with deep neural networks and getting 94% accuracy. While Uddin et al. (2023) [13] utilized ML for BC diagnosis achieving accuracy of 98.77%.

Table 2 Breast cancer classification summarization

S. No.	Author & Year	Algorithm Used	Datasets	Evaluation Metrices
1	Jakhar et al. (2023) [2]	Stack based ensemble	WDBC (699 instances)	Accuracy= .9880 Precision= .9909 Recall= 0.9909 F1 score= 0.9909
2	Jakhar et al. (2023) [2]	Stack based ensemble	Breakhis	Accuracy= 0.9435 Precision= 0.9245 Recall= 0.9596 F1 score= 0.9417
3	Talatian et al. (2021) [3]	MLP neural network (MLP-NN) and evolutionary algorithm	WBCD (699 instances)	Accuracy= .9874
4	Naseem et al. (2022) [4]	ANN, SVM, LR, DT, and k-NN	WDBC (569 instances)	Accuracy= .9883
5	Srinivas et al. (2022) [5]	LR, RF and SGD	WDBC (569 instances)	Accuracy= 0.985 Precision= 0.970 Recall= 0.990 F1 score= 0.980
6	Jabbar et al. (2021) [6]	Bayesian network and radial basis function	WBCD (699 instances)	Accuracy=.9742 Precision= .9672
7	Alhayali et al. (2020) [7]	Hofeding tree and Naïve Bayes	WBCD (699 instances)	Accuracy= .9599
8	Batool & Beyon (2024) [8]	Voting ensemble classifier	WDBC (569 instances)	Accuracy = .976 Precision= .964 Recall = 1.00, F1 score = .981
9	Chaurasia et al. (2018) [9]	RBF	WBCD (699 instances)	Accuracy = .96 Precision =.9623
10	Chaurasia et al. (2018) [9]	J48	WBCD (699 instances)	Accuracy = .9341 Precision = .9037
11	Hashim & Yassin (2023) [10]	Soft Voting Classifier	WDBC (569 instances)	Accuracy =.993 Precision =1.00 Recall =.9846

				F1 score =.992 AUC = 0.992.
12	Sharma et al. (2024) [11]	Stacked based ensemble classifier	WDBC (569 instances)	Accuracy= .9766
13	Anastraj et al. (2019) [12]	Back propagation network, ANN, CNN, SVM	Wisconsin breast Cancer (699) dataset	Accuracy=.94
14	Uddin et al. (2023) [13]	Voting Classifier [LR+SVM]	WDBC (569 instances)	Accuracy=.9877 Precision=.9883 Recall=.9854 F1 score= .9868

These studies validate the effectiveness of ML algorithms in accurately predicting breast cancer outcomes by taking various measures of evaluation. These findings highlight the potential of ML in assisting clinicians with early diagnosis and personalized treatment. Previous research shown in Table 2 employing a variety of classifiers has produced encouraging results. But prior research has shown that employing an ensemble of classifiers enhances the outcomes. To overcome this gap, we purpose a model in this study that improves the outcomes of breast cancer diagnosis by utilizing many classifiers or an ensemble of classifiers.

3. Purposed Methodology

The dataset and classification models used to improve classifier compression are described in this section. In order to develop an appropriate model for the high-precision prediction of breast cancer the primary stages of the suggested architecture are depicted in Fig.1. Main areas are covered by our suggested methodology: data preprocessing, balancing the dataset, scaling the features and cross validation. The dataset is balanced using the random over sampling (ROS) approach and the features were scaled using a standard scaler method. Hard voting classifier is used to make the prediction and evaluate the model performance based on the training data findings.

Procedure of purposed methodology

1. Load dataset.
2. Check for any such empty values or NaN’s in the dataset.
3. If NA or missing value is from the data, try to replace it with the appropriate value.
4. Map the target variable values: let's say 'M' is 0. If 'B' is 1, replace with it.
5. Distribute or apportion X features and the Y target variable from the data file.
6. Split the dataset into two pieces: training and testing sets (X_train, X_test, Y_train, Y_test).
7. Perform Random Over Sampling to equalize the unbalanced distribution of classes.
8. Run the features through the StandardScaler function.
9. Train the models: Logistic Regression, SVM, decision trees.
10. Create hard Voting Classifier implementation with loaded models into the script.
11. Train the hard Voting Classifier using the resampled and scaled training data.
12. Make predictions and evaluate each classifier on test set.
13. Perform 10-fold data cross-validation on Voting Classifier and Compute the mean cross-validation accuracy value.

Fig. 2 highlights the flowchart of the method being discussed in this paper to show the process flow beginning from data collection and steps down to the analysis stage. A flowchart is useful in showing flow and relationships in a process.

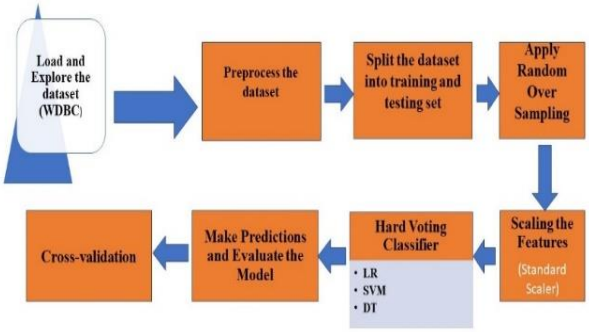

Fig.1. Purposed architecture for Breast cancer prediction

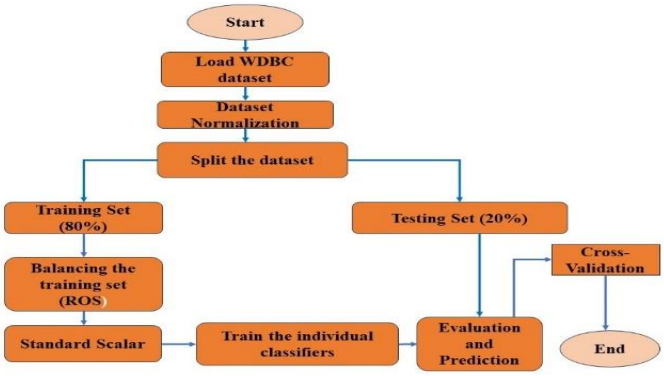

Fig.2. Flowchart of purposed methodology

In our approach, we employ voting-based ensemble learning also known as the super learning in order to develop the model which is achieved by modelling variety of classification models. Algorithm 1 describes the purposed framework steps, inputs in Table 3 are described through symbols in the algorithm.

Algorithm 1: Purposed ensemble framework for breast cancer prediction

D' = N(D)	/*normalisation of dataset*/
M=0, B=1	/* Map the target variables*/
Let G = {g1 , g2 , g3 .. . gn }	/*dataset*/
C = {C1, C2, C3, ... Cn}	/*the set of ML ensemble classifiers*/
X= the 80% dataset for training, X ⊆ G	/*80% of the dataset */
Y= the 20% dataset for testing, Y ⊆ G	/*20% of the dataset */
X'= Bs(X)	/* Balancing and scaling the training set*/
V =Voting classifier	
P = n (G)	/*where P is no. of attributes of dataset*/
Begin	
M(i) = C(i)	/* training the model on X'*/
Next i	/*loop where i is a variable*/
Mo = Mo ∪ V	/*union of model and voting classifier*/

End
Result ←Mo classifies Y
Predict← Cross-Validation

Table 3 The symbols are presented in Algorithm 1

S. No.	Symbols	Description
1	D	Dataset
2	D’	Normalized dataset
3	M	Malignant
4	B	Benign
5	N	Normalization
6	G	Attributes of datasets
7	C	Classifiers
8	X	Training set
9	Y	Testing set
10	V	Voting classifier
11	X’	Balanced and scaled set
12	Bs	Balancing & scaling
13	P	Number of attributes
14	i	Variable
15	Mo	Model

3.1. Dataset Description and Preprocessing

Most popular benchmark dataset for breast cancer classification tasks is the “Wisconsin Breast Cancer Diagnostic (WDBC)” dataset obtained from the UCI ML Repository is utilized in this work. The features in this dataset which indicate the properties of cell nuclei were calculated from digital pictures of fine needle aspirates (FNA) of breast masses. It is composed of 569 examples each of which has 30 attributes (such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) that were extracted from photographs of cell nuclei [14]. The diagnostic which indicates the kind of breast mass and is classified as either malignant (M) or benign (B) is the target variable in this dataset. The dataset is useful for creating and assessing ml models for breast cancer diagnosis because it offers a thorough description of the features of the disease. Fig.3 shows the class distribution of breast cancer of WDBC with 569 instances.

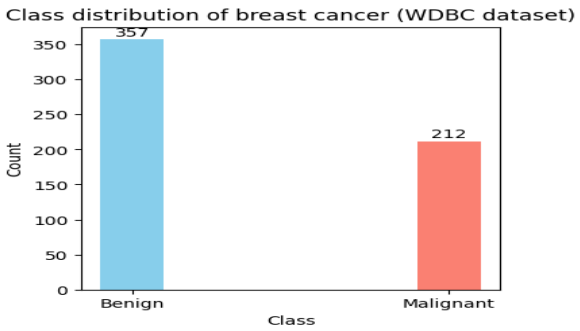

Fig.3. Class distribution of breast cancer of WDBC

3.1.1. Splitting the dataset

We obtained data from the UCI ML Repository as well and through the insightful study, we discarded outliers. The next step was to divide the data into 20% testing section and 80% section for training. Such ideas play an important role in every stage of our metalogical framework. The dataset is initially split into two categories: x represents all features not including the target, and y represents the target. Next, we'll use the Train _Test _ Split Procedure to divide the dataset into training and testing sets. Training data are utilized to educate the model while testing data are employed to evaluate the model's functionality following training. The partition of Training and Testing set in our work is depicted in Fig.4.

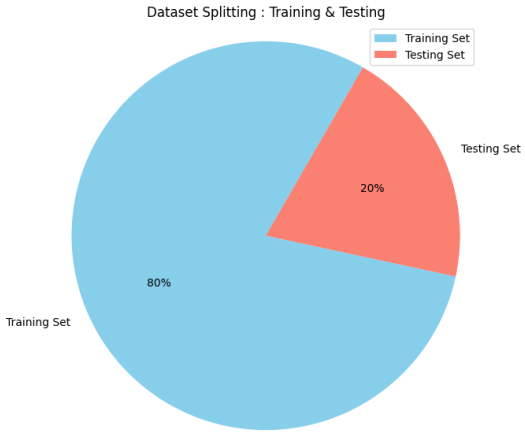

Fig.4. The partition of dataset into training and testing set

This stage is crucial for getting our dataset ready for the following stages. The models are able to identify breast cancer once they have been trained on the training set.

3.1.2. Balancing the dataset

The class balancing, which takes place in many cases where there are imbalanced categories (such as in medical diagnosis) requires means that will help straighten out the uneven distribution among the classes. Therefore, in this situation, Random Over Sampling (ROS) method is applied, which makes multiply copies of minority class samples to compensate the disparity in class distribution.

Random Over Sampler (ROS): The Random Over Sampler (ROS) technique is a method used to address class imbalance in datasets where one class significantly out numbers the other. In the context of machine learning this inequality can lead to biased models that favour the majority class ensuing in poor performance for the minority class. ROS works by randomly duplicating instances from the minority class until both classes are balanced. This oversampling technique increases the representation of the minority class and providing more examples for the model to learn from during training.

The motivation behind using ROS is to ensure that the model learns from a more balanced dataset which can improve its ability to correctly classify instances from both classes. This approach helps prevent the model from being biased towards the majority class and improves its overall performance and generalization ability. Fig.5. shows the Class Distribution before and after RandomOverSampler on training set.

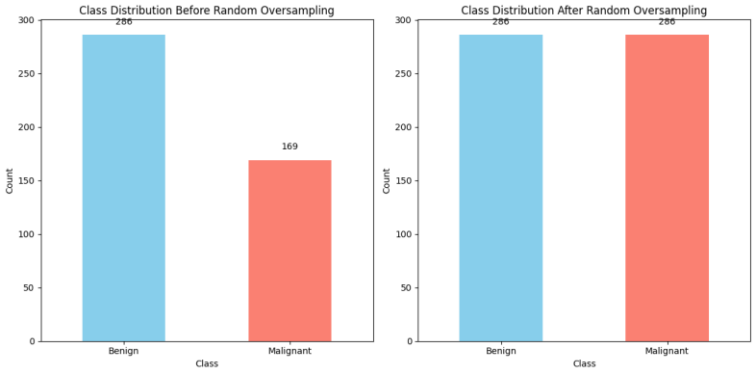

Fig.5. Class Distribution before and after RandomOverSampler (ROS) on Training Dataset

The WDBC dataset is divide into training (455 instances) and testing (114 instances) sets. After balancing with ROS, the training set exhibited a class distribution of 286 benign (B) and 169 malignant (M) instances into 572 instances. Table 4 shows whole description of the class distribution of WDBC dataset used in this work.

Table 4 Dataset distribution summary

	WDBC	**After Splitting**		**Training set after ROS**
		Training set	Testing set	
Total Instances	569	455	114	572
Class Distribution	B=357, M=212	B=286, M=169	-	B=286, M=286

3.1.3. Feature Scaling

Standard Scaler is a feature scaling technique used to normalize the range of independent variables or features in a dataset. Because it guarantees that features are on a same scale and keeps some features from predominating over others during model training, this procedure is essential to machine learning. In order to use Standard Scaler each feature's mean and standard deviation are first determined. The feature values are then transformed to have a mean of 0 and a standard deviation of 1. The aim of the ease of comparison of traits using the Standard Scaler is to ensure that the model treats all features equally and during the training session. Thereby, it ensures better rate of convergence of optimization algorithms and it addresses and prevents numerical errors that initializes when features have vastly different scales. Feature Scale with the Standard Scaler implemented is an indispensable preprocessing step in machine learning which homogenously scales features to have a range within similar ranges, leading to more stable and reliable models. To standardize the dataset by removing the mean and scaling them to unit variance, there is a machine learning technique known as StandardScaler [10]. The formula for standard scaling (z-score normalization) applied by StandardScaler is:

$$x' = \frac{x - \mu}{\sigma} \qquad (1)$$

Where x′ is the feature's standardized value, x is the feature's original value, μ is the feature values' mean, and σ is their standard deviation.

3.2. Classification Models

3.2.1. Logistic regression

The algorithm for classification LR is an approach to supervised learning. This method predicts the probability of a dichotomous target (dependent) variable. After the computation it yields probabilistic values between 0 and 1. These probabilistic values are the basis for the categorization that the algorithm uses. In cases when the technique makes use of a sigmoid function (0=output>=1), LR and linear regression are quite comparable. The only thing that differs is the hypothesis function; however, a regression function (-∞<=output>=∞) is used. Consequently, they have a variety of uses the former is employed in regression, while the latter is used in classification. It helps understand dependent and independent variables in data analysis [15].

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}} \qquad (2)$$

where a is the biased or intercept component, b is the coefficient, and P is the expected output for a given input value (X).

3.2.2. SVM

Strong supervised learning algorithms like SVM are employed for both regression and classification problems. Finding the hyperplane that optimally divides the data into classes while optimizing the margin between them is the main goal of support vector machines (SVM). SVM determines the best hyperplane that maximizes the margin that is, the separation between the nearest data points (support vectors) and the hyperplane from each class when the data is linearly separable [16]. The decision boundary (hyperplane) of SVM may be expressed mathematically as follows:

$$w^T x + b = 0 \qquad (3)$$

where x is the input feature vector, b is the bias term, and w^T stands for the transpose of w. W is the weight vector perpendicular to the hyperplane.

3.2.3. Decision Tree

DT classifier is a ML method for classification. With the help of this approach different classes may be classified using a tree structure called a decision tree (DT) in which each leaf node symbolizes the class that the decision represents, internal (decision) nodes reflect attributes of a dataset that is used to make any choice, and branches indicate decision rules. In essence a DT poses a question and divides the tree into subtrees according to the answer (yes/no). They offer a strong framework for weighing your alternatives. Entropy is a suitable term to represent the amount of data required to characterize a sample. Thus, the entropy is maximal when the sample is evenly split; otherwise, it is zero if the sample is homogenous meaning that all of the elements are similar [15].

At each node m of the tree, a decision function $f_m(x)$ is applied to determine the splitting criterion based on a feature x_j and a threshold t, such that:

$$f_m(x) = \begin{cases} 1, & if\ x_j \le t \\ 0, & otherwise \end{cases} \qquad (4)$$

where j and t are chosen to maximize the information gain or minimize impurity.

3.2.4. Voting Classifier

A voting classifier is a technique for group learning that generates a final prediction by aggregating the predictions of several different classifiers. Working on the basis of "majority voting," the projected class is determined by tallying the votes cast for each classifier with each prediction helping to shape the final result. Voting classifiers come in two varieties: soft voting and harsh voting. In hard voting the class that receives the most votes is the winner and in soft voting the class that receives the greatest average probability is determined by averaging the class probabilities predicted by each classifier [8]. The prediction of a voting classifier may be expressed mathematically as follows:

$$\hat{y} = argmax_i \sum_{j=1}^{M} P_{ij} \qquad (5)$$

Where: $\hat{y}$ is the predicted class, M is the number of classifiers, Pij is the probability predicted by the i^{th} classifier for class j.

3.3. Performance Measures

Five cross-validation matrices are evaluated in this study: accuracy, recall, F1 score, and precision and AUC. The confusion matrix's values, which are TP that is the prediction and the actual data are both yes can be used to compute these matrices, TN both the actual data and the prediction are negative, FP and FN refer to the following: yes, for the forecast and no for the actual data. It is possible to compute precision, recall, F1 score, and accuracy using the following formulas [20]. A generalized confusion matrix that aids in defining the various performance metrics is shown in Fig.6.

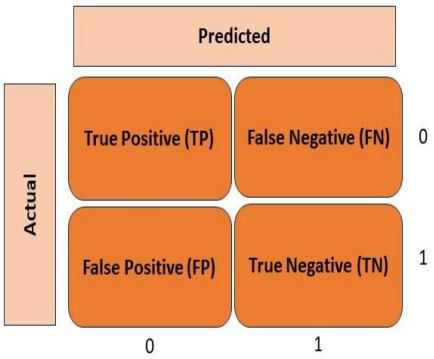

Fig.6. A general view of confusion matrix

The following equations apply to the performance indicators include F1-Score, AUC-ROC, recall, accuracy, and precision.

Accuracy: One of the most often used metrics for evaluating a classifier's performance is accuracy [11]. It is defined as follows and represented as a percentage of correctly identified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \qquad (6)$$

Precision: It is defined as the ratio of real positive occurrences to those that a potential classifier predicts would be positive [21].

$$Precision = \frac{TP}{TP + FP} \qquad (7)$$

Sensitivity/ Recall: It is also known as true positive rate (TPR) refers to the capacity of a classifier to accurately forecast a positive outcome when a disease is present [21]. It is defined as

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

F1-Score: The weighted average of Precision and Recall yields the F1 Measure [21]. The calculation is performed using equation 9.

$$F - score = \frac{2*precision*Recall}{Precision+Recall} \quad (9)$$

AUC-ROC: "Area under the curve-receiver operating characteristics (AUC-ROC)", provides the performance metrics across all classification criteria. It shows how well a classifier can distinguish between different classes. The probability curve is represented by ROC while the degree or measure of separability is represented by AUC [2].

$$AUC - ROC = \frac{1}{2} * \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (10)$$

4. Experimental Results and Discussion

The confusion matrices, accuracy, F1 score, precision, recall, AUC-ROC performance data that we obtained by applying the suggested methods are displayed and discussed in this section. We use balanced datasets to compare the output of the models contained in our proposed model, the hard Voting Classifier with the output of the other models. The outcomes of our suggested model are further assessed using 10-fold cross-validation on the balanced dataset. Python is used for the purpose of implementing the proposed methodology as this language supports highly efficient data analysis and ML libraries. This made the research data to be processed in an efficient way as well as analysing the research findings in an accurate manner.

We split the dataset (WDBC 569 instances) into two parts: testing (20% with 114 instances) and training (80% with 455 instances). After balancing the training set using ROS it is converted into 572 cases (M = 286, B = 286). After that, we train and test the performance of the hard voting classifier as well as the models that we use in this study, LR, DT, and SVM. Fig.7. shows the confusion matrices for all individual classifier and hard voting classifier.

Table 5 represents the comparison of the results of baseline classifier and voting classifier. The comparison chart of evaluation metrics of baseline machine learning models and hard voting classifier shown in Fig.8. Classification algorithms for breast cancer dataset in Table 6 encapsulates a detailed framework. Moreover, the proposed Voting Classifier, unlike other techniques, composed by LR, SVM, and DT, has achieved high performance, obtaining an accuracy of 98.25% in the WDBC dataset.Fig.9. and Fig. 10 indicates, comparison chart of accuracy and precision of the currently developed models has been pulled up against the intended one.

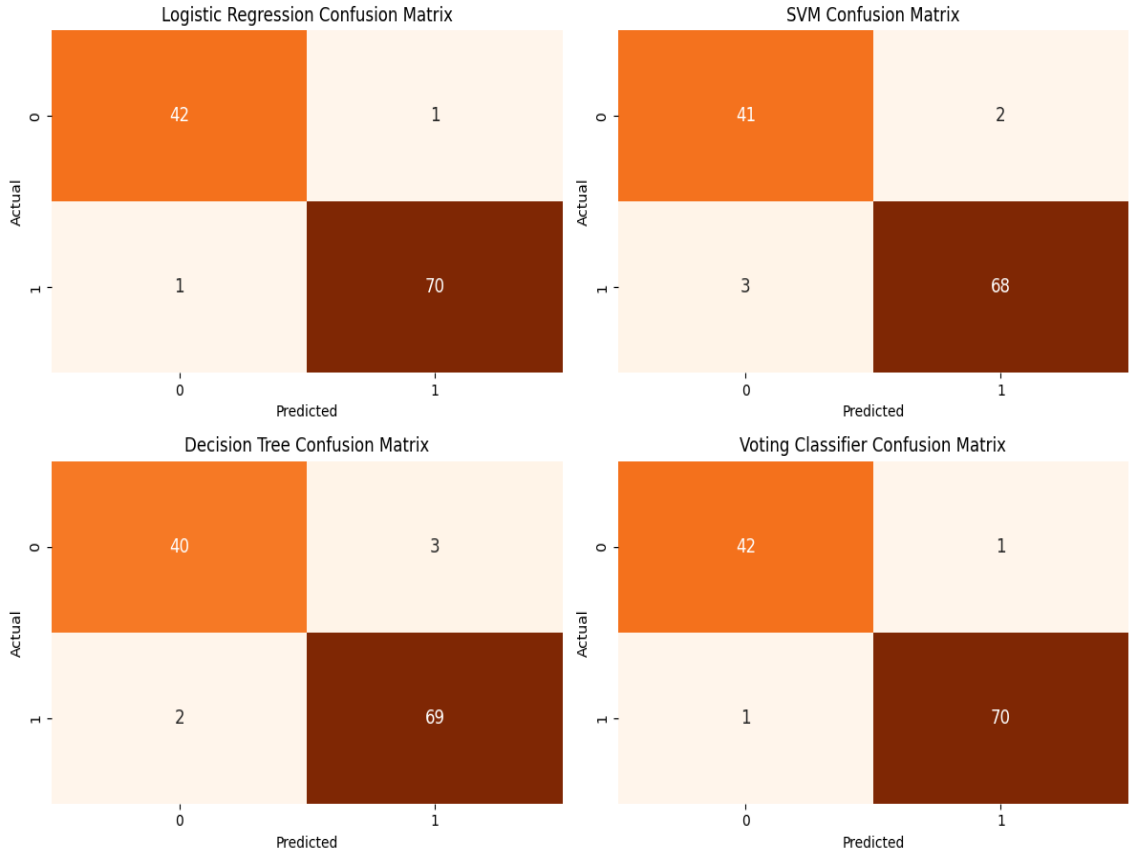

Fig.7. Analysing the performance of different classifiers: logistic regression, support vector machine, decision tree and voting classifier through confusion matrix

Table 5 Comparison of evaluation metrics of baseline machine learning models and hard voting classifier.

Algorithms	Accuracy	Precision	Recall	F1-Score	AUC-ROC
LR	0.9824	0.9859	0.9859	0.9859	0.9813
SVM	0.9561	0.9714	0.9577	0.9645	0.9556
DT	0.9561	0.9583	0.9718	0.9650	0.9510
Hard voting classifier	**0.9825**	**0.9859**	**0.9859**	**0.9859**	**0.9813**

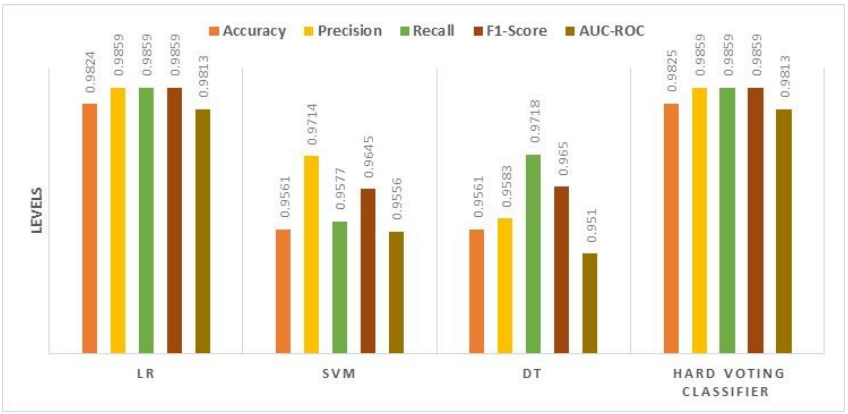

Fig.8. Comparison chart of evaluation metrics of baseline machine learning models and hard voting classifier

Table 6 Comparison of evaluation metrics of existing models and purposed model.

Year & Reference	Algorithm	Dataset	Accuracy	Precision	Recall	F1-Score	AUC-ROC
(2018) [9]	RBF	WBCD (699)	.96	.9623	-	-	-
(2019)[12]	Back propagation network, ANN, CNN, SVM	WBCD (699)	.94	-	-	-	-
(2020) [7]	Hofeding tree and Naïve Bayes	WBCD (699)	.9599	-	-	-	-
(2021) [6]	Bayesian network and radial basis function	WBCD (699)	.9742	.9672	-	-	-
(2021)[18]	DT	WDBC (569)	.9253				
(2023)[19]	XGBoost	WDBC (569)	.974	0.960	1.00	0.980	
(2024) [11]	Stacked based ensemble classifier	WDBC (569)	.9766	-	-	-	-
2024 [17]	stacking with logistic regression ensemble model	WDBC (569)	.9737	-	-	-	-
Purposed Model	**Voting Classifier (LR+SVM+DT)**	**WDBC (569)**	**.9825**	**0.9859**	**0.9859**	**0.9859**	**0.9813**

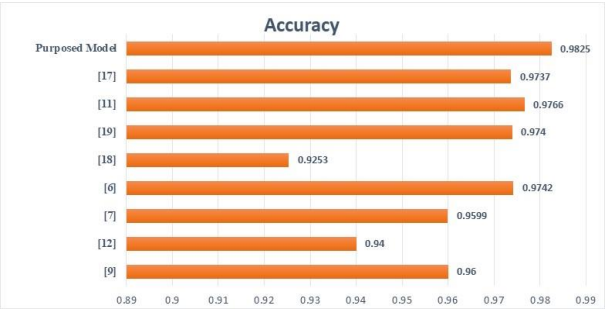

Fig.9. Comparison chart of accuracy of existing works with purposed model

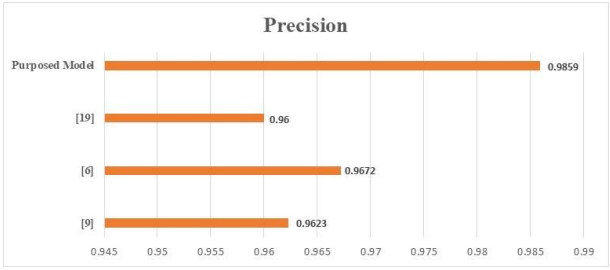

Fig.10. Comparison chart of precision of existing works with purposed model

Finally, we assess our model for estimating the out-of-sample error using the 10-fold cross-validation approach. The 10-fold cross-validation improves the evaluation of our model's performance by preventing overfitting and creating a more generalized model. This method is used with both the balanced data set and the suggested "hard voting classifier". The testing results revealed a mean accuracy of .9738.

5. Conclusion and future works

In this paper, an ensemble classification method has been employed on the WDBC datasets in order to early and accurate breast cancer prediction that is based on the voting strategies (LR, DT, and SVM). The suggested model performed better than other cutting-edge models, with accuracy of .9825, precision of .9859, recall of .9859, F1 score of .9859, and AUC of 0.9813. Upon doing a 10-fold cross-validation comparison, the accuracy of the suggested model was found to be .9738 surpassing that of previous published models. Due to small sample of population its findings may not apply to larger

groups of people. Future research should focus on using clinical datasets. To make the classification more accurate this study could also look into adding different optimization strategies to the suggested approach.

Conflict of Interest : The authors declare that there is no conflict of interest.

Acknowledgement : The authors have no support for this work including funding, grants, or other financial support to be able to write this manuscript.

References

[1] World Health Organization, Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed 2024.
[2] Jakhar, A.K., Gupta, A., Singh, M., (2023). SELF: a stacked-based ensemble learning framework for breast cancer classification. Evol. Intell. 1–16. https://doi.org/10.1007/s12065-023-00824-4.
[3] Talatian Azad, S., Ahmadi, G., Rezaeipanah, A. (2021). An intelligent ensemble classification method based on multi-layer perceptron neural network and evolutionary algorithms for breast cancer diagnosis. J. Exp. Theor. Artif. Intell. 1–21.
[4] Naseem, U., Rashid, J., Ali, L., Kim, J., Haq, Q.E.U., Awan, M.J., Imran, M. (2022). An automatic detection of breast cancer diagnosis and prognosis based on machine learning using ensemble of classifiers. IEEE Access 10, 78242–78252.
[5] Srinivas, T., Karigiri Madhusudhan, A.K., Arockia Dhanraj, J., Chandra Sekaran, R., Mostafaeipour, N., Mostafaeipour, A. (2022). Novel based ensemble machine learning classifiers for detecting breast cancer. Math. Probl. Eng. https://doi.org/10.1155/2022/9619102.
[6] Jabbar, M.A., 2021. Breast cancer data classification using ensemble machine learning. Eng. Appl. Sci. Res. 48 (1), 65–72.
[7] Alhayali, R.A.I., Ahmed, M.A., Mohialden, Y.M., Ali, A.H.(2020). Efficient method for breast cancer classification based on ensemble Hoeffding tree and Naïve Bayes. Indones. J. Electr. Eng. Comput. Sci. 18 (2), 1074–1080.
[8] Batool, Byun, Y.C. (2024). Towards improving breast cancer classification using an adaptive voting ensemble learning algorithm. IEEE Access.
[9] Chaurasia, V., Pal, S., Tiwari, B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. J. Algorithms Comput. Technol. 12 (2), 119–126. https://doi.org/10.1177/1748301818756225.
[10] Hashim, M.S., Yassin, A.A. (2023). Breast cancer prediction using soft voting classifier based on machine learning models. IAENG Int. J. Comput. Sci. 50 (2).
[11] Sharma, D., Goyal, R., Mohana, R. (2024). An ensemble learning-based framework for breast cancer prediction. Decis. Anal. J. 10, 100372.
[12] Anastraj, K., Chakravarthy, T. (2019). Analysis of breast cancer using back propagation with deep neural network. Int. J. Comput. Sci. Eng. 7 (4), 844–847.
[13] Uddin, K.M.M., Biswas, N., Rikta, S.T., Dey, S.K. (2023). Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. Comput. Methods Programs Biomed. Update 3, 100098.
[14] Wolberg, W., Mangasarian, O., Street, N., Street, W. (1995). Breast Cancer Wisconsin (Diagnostic), UCI Machine Learning Repository.
[15] Hossin, M.M., Shamrat, F.J.M., Bhuiyan, M.R., Hira, R.A., Khan, T., Molla, S. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. Bull. Electr. Eng. Inform. 12 (4), 2446–2456.
[16] Kadhim, R.R., Kamil, M.Y. (2023). Comparison of machine learning models for breast cancer diagnosis. IAES Int. J. Artif. Intell. 12 (1), 415.
[17] Laghmati, S., Hamida, S., Hicham, K., et al. (2024). An improved breast cancer disease prediction system using ML and PCA. Multimed. Tools Appl. 83, 33785–33821. https://doi.org/10.1007/s11042-023-16874-w.
[18] Assegie, T.A., Tulasi, R.L., Kumar, N.K., (2021). Breast cancer prediction model with decision tree and adaptive boosting. IAES Int. J. Artif. Intell. 10 (1), 184.
[19] Chen, H., Wang, N., Du, X., Mei, K., Zhou, Y. Cai, G.(2023). Classification prediction of breast cancer based on machine learning. Comput. Intell. Neurosci.
[20] Rasool, C., Bunterngchit, L., Tiejian, L., Islam, M.R., Qu, Q., Jiang, Q. (2022). Improved machine learning-based predictive models for breast cancer diagnosis. Int. J. Environ. Res. Public Health 19 (6), 3211.
[21] Houfani, D., Slatnia, S., Kazar, O., Remadna, I., Saouli, H., Ortiz, G., Merizig, A. (2023). An improved model for breast cancer diagnosis by combining PCA and logistic regression techniques. Int. J. Comput. Digit. Syst.