Ensemble Machine Learning for Enhanced Breast Cancer Prediction: A Comparative Study

Md. Mijanur Rahman[1]*, Khandoker Humayoun Kobir[2], Sanjana Akther[3], Md. Abul Hasnat Kallol[4]
Assistant Professor, Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh[1]
Student, Department of Computer Science and Engineering, Southeast University, Dhaka, Bangladesh[2, 3, 4]

***Abstract*—Breast cancer poses a significant threat to women's health, affecting one in every eight women globally and often leading to fatal outcomes due to delayed detection in advanced stages. Recent advancements in machine learning have opened doors to early detection possibilities. This study explores various machine learning algorithms, including K- Nearest Neighbor (KNN), Support Vector Machine (SVM), Multi- Layer Perceptron (MLP), Decision Tree (DT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), Ada Boost (AB), Gradient Boosting (GB), and XGboost (XGB). The employed algorithms, along with nested ensembles of Bagging, Boosting, Stacking, and Voting, predicted whether a cell is benign or malignant using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Utilizing the Chi-square feature selection technique, this study identified 21 essential features to enhance prediction accuracy. Results of this study indicate that MLP LR achieved the highest accuracy of 98.25%, closely followed by SVM with 97.08% accuracy. Notably, the Voting classifier yielded the highest accuracy of 99.42% among the ensemble methods. These findings suggest that the research model holds promise for accurate breast cancer prediction, thus contributing to increased awareness and early intervention.**

Keywords*—Breast cancer; detection; machine learning; bagging; boosting; stacking; voting; chi square; ensemble; hybrid ensemble; bioinformatics

I. Introduction

This Breast cancer is one of the alarming signs of female health, as many patients are added to the breast cancer queue every year globally. Mortality rate and late detection problems confirm that early detection is a must. According to the WHO report of July 2023, about two and nearly half million women were diagnosed, and 685,000 died in 2020 globally. On the other hand, in the last five years, only 7.8 million women survived after being diagnosed with breast cancer, ensuring it is the world's most prevalent cancer [1]. Similarly, as per the Daily Star report, more than 12 thousand women and seven hundred are diagnosed with breast cancer every year in Bangladesh, and about 6,844 of them don't make it [2]. Despite significant advancements in medical science, early detection remains the primary obstacle in effectively treating breast cancer. Timely detection is crucial, as failure to do so can result in fatal outcomes for patients [3].

Cancer occurs when healthy tissues undergo uncontrolled growth, forming masses or clusters of cells called tumors. Tumor cells can be of two types: Cancerous (called Malignant) and non-cancerous (named Benign) [4]. Malignant is harmful because this cell can grow and spread to other body parts, whereas Benign does not spread but grows. Cancer is detectable through physical examination, biopsy, or mammograms. These ways are effective but time-consuming, costly, and painful.

The primary challenge in breast cancer diagnosis lies in distinguishing between cancerous (malignant) and non-cancerous (benign) cells. Machine learning algorithms have emerged as a solution to this problem, leveraging previous patient data to develop various models. Over the past few decades, these algorithms, particularly Artificial Neural Networks and Support Vector Machines (SVM), have consistently demonstrated high accuracy and effectiveness. Their reliability makes them valuable tools for achieving better diagnostic outcomes [3].

Ensemble machine learning algorithms have long been employed to improve detection accuracy. In a recent study by R. Murtirawat et al. [3], an update on Ensemble Learning techniques involving five machine learning algorithms (LR, KNN, LDA (Linear Discriminant Analysis), SVM, RF) was presented. The study achieved an impressive accuracy of 99.30% using a 75% training and 25% testing dataset. Therefore, this paper represents a significant milestone by achieving even higher accuracy with a 70:30 training-testing data ratio.

In a study conducted by E. Strelcenia et al. [5], the accuracy of several machine learning methods, including LR, DT, RF, KNN, MLP, and XGB Classifier, was evaluated. Their findings revealed accuracies of 96%, 98%, 97%, 89%, 92%, and 94%, respectively, based on their dataset. Machine learning techniques consistently demonstrate notable accuracy percentages across various applications. In the context of breast cancer predictions, these algorithms proved to be particularly effective, yielding higher accuracy rates. For instance, both MLP and LR achieved an accuracy of 98%. Additionally, SVM, KNN, and XGboost demonstrated performances with 97% accuracy.

The aim of this research is to use ten different computer programs to guess if someone has breast cancer, using a dataset of 21 features. The goal is to accurately differentiate between benign and malignant cases using a range of algorithms, including KNN, SVM, DT, RF, MLP, NB, LR, ADB, GB, and XGB, along with other ensemble techniques such as Bagging, Voting, and Stacking. Ensemble techniques play a crucial role in breast cancer prediction and diagnosis, offering enhanced accuracy, particularly in the early stages. The proposed study utilizes proper statistical feature selection techniques to

effectively detect breast cancer by distinguishing between benign and malignant cases. Furthermore, the research model holds promise for advancing developments in breast cancer research and improving patient care and treatment.

This study is structured into several sections. Firstly, it serves as a literature review, providing relevant information and discussion. Following this, the methodology section outlines the techniques and algorithms used in the model. Subsequently, the results section presents outcomes in terms of matrices and parameters. Discussions ensue, where the findings are analyzed and compared with existing works. Finally, the conclusion summarizes the study's key insights.

II. Literature Review

Ensemble methods have been used in breast cancer detection for quite some time now, mainly because they've been proven to boost accuracy. With the continuous advancements in medical science and machine learning algorithms, the impact on breast cancer research is becoming increasingly evident, attracting more researchers to the field each day. Many have focused their efforts on the WBCD dataset due to its extensive statistical data, allowing for more thorough experimentation.

In a recent study conducted by R. Shafique et al. [6], the significance of feature selection techniques was highlighted by comparing the performance of PCA, chi-square, and SVD on specific datasets. Models constructed using RF, SVM, GBM, LR, MLP, and KNN algorithms demonstrated enhanced accuracy with all three techniques. Notably, KNN achieved the highest accuracy of 95% on the WDBC dataset across the three feature selections. Following this, the study addressed dataset imbalance by employing upsampling, which involves adding extra samples to test the model's performance. This approach aimed to mitigate potential inaccuracies resulting from neglecting the minor class, ultimately improving accuracy. The following study displayed that Chi2 offered more impactful results than PCA comparatively in different models on statistical datasets, especially WDBC.

A study by M. Kumar et al. [7] introduced the OSEL (Optimized Stacked Ensemble Learning) model, which combines various algorithms such as KNN, RF, LR, SVM, DT, ADBM1, GB, SGB (Stochastic Gradient Boosting), and Cat Boost. This model achieved impressive metrics, including 99.4% accuracy, 99% precision, 98% recall, and 99% F-measure. As an effective heterogeneous ensemble method, Stacking demonstrated superior performance compared to other Boosting classifiers, resulting in higher accuracy, precision, recall, and F1-measure. This makes the combined model particularly relevant in the current research landscape. Another notable ensemble model for diagnosis was established by U. Naseem et al. [8], utilizing a combination of four classification methods (SVM, LR, NB, DT) as base learners and artificial neural networks (ANN) as the meta-learner. This model achieved an accuracy of 97.6% without sampling and 98.83% with sampling. In the prognosis case, the ensemble model performed best with SVM, LR, and RF as base learners and ANN as the meta-learner, achieving an accuracy of 83.15% without sampling and 88.33% with sampling. Notably, SVM consistently outperformed other classification models across both diagnosis and prognosis datasets when used individually.

The study by R. Murtirawat et al. [3] garnered attention for showcasing remarkable accuracy through the Voting ensemble technique. Their updated Ensemble Model (LR, KNN, LDA, SVM, and RF) achieved an impressive accuracy of 99.42% with a 75% training dataset and 25% testing dataset. However, this notion was challenged by another report from A. Assiri et al. [9], whose Voting ensemble boasted even higher accuracy of 99.42%, achieved solely through majority Voting. This majority-based algorithm was constructed using the top three algorithms (logistic learning, SVM with SGD, and multilayer perceptron) from the initial eight classification techniques, which then determined the final result through a voting mechanism. Interestingly, this study revealed that the majority-based ensemble model outperforms the soft voting accuracy (98.83%), showcasing its comparative effectiveness.

V. Nemade et al. [10] presented a model comprising two sections: standard ML algorithms and ensemble techniques. Achieving an accuracy of 97% with XGboost, the evaluation was based on the confusion matrix labels, including True Negative (TN), False Negative (FN), True Positive (TP), and False Positive (FP). Notably, this model utilized AUC as a metric, distinguishing itself from others that typically rely on accuracy, precision, recall, and AUC-ROC.

M. Ramakrishna et al. [11] proposed an AdaBoost ensemble model that leveraged recognized feature patterns. Notably, Adaboost-RF and Adaboost-NB took 8.52s and 18.32s, respectively, to develop the model. Impressively, Adaboost-RF achieved an accuracy of 97.95%, demonstrating commendable performance. Further evidence of the effectiveness of the AdaBoost algorithm was provided by N. Mashudi et al. [12], who implemented it on the WBCD dataset and achieved an accuracy of 98.77%. Through various cross-validation techniques such as 2-fold, 3-fold, and 5-fold, AdaBoost demonstrated consistent high accuracy, with scores of 98.41% and 98.24% for 2-fold and 3-fold cross-validation, respectively. Additionally, SVM displayed a notable accuracy of 98.60% in 5-fold cross-validation.

In a study by M. Momtahen et al. [13], a DOB-Scan probe was introduced to classify breast tissues as healthy or unhealthy. They devised a technique utilizing bagging and boosting on machine learning classifiers, achieving 100% accuracy in classifying 68 tissue-mimicking liquid phantom samples. Similarly, the effectiveness of the voting classifier was demonstrated in a paper by Q. Nguyen et al. [14]. In their research, the Ensemble-voting classifier, SVM tuning, and logistics regression achieved an accuracy of 98.83%. The study utilized PCA for feature extraction and implemented a 90:10 training-testing ratio with 10-fold cross-validation to mitigate the risk of overfitting.

III. Methodology

For early detection, it's crucial to determine whether the affected cell is cancerous (Malignant) or not (Benign). The process discussed in this research involves several phases, including data analysis, model preparation, training, and ensemble techniques. During data analysis, researchers of this

study conducted data description, collection processing, and feature selection. The prepared data is then used for model preparation, implementing various ML algorithms. This includes individual machine training, ensemble approaches, and evaluation metrics.

In Fig. 1, the primary task was to pre-process the data. After selecting the WBCD dataset and dividing it into training (70%) and testing (30%) sets, this study applied standard scaling to standardize features. Ensemble and machine learning models of this research aim to detect cancer cells (Benign or Malignant), incorporating the effectiveness of all algorithms and ensemble models with differences in their performance.

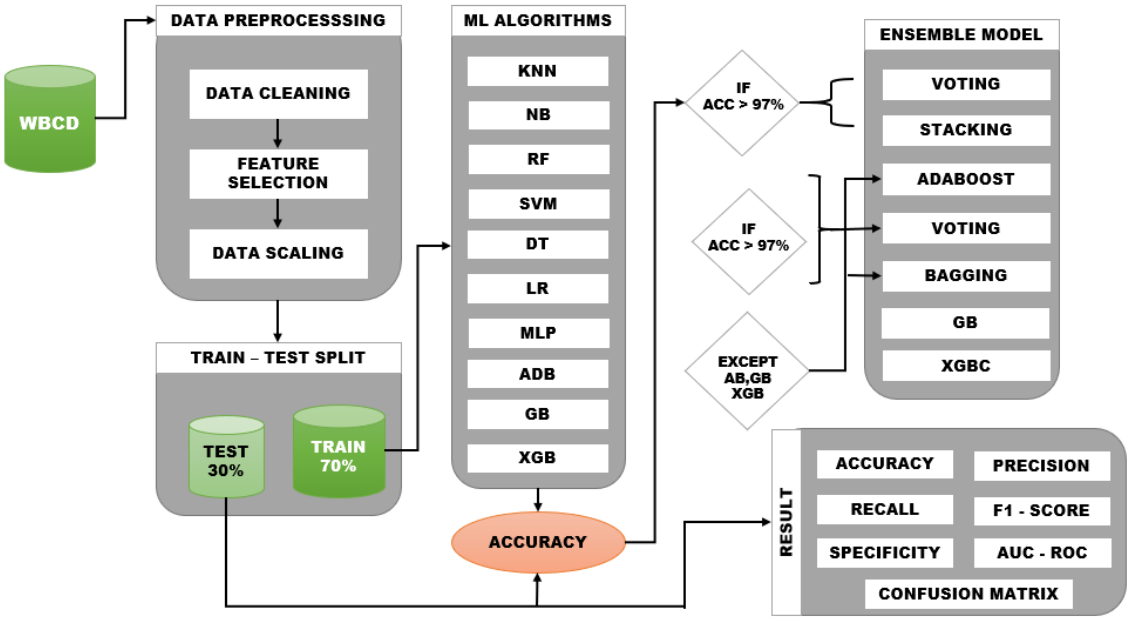

Fig. 1. Diagram of proposed methodology.

A. Dataset Description

The 'WBCD' dataset, curated by Dr. William H. Wolberg from the University of Wisconsin Hospital in Madison, comprises 569 rows and 33 columns. For each cell nucleus, ten actual valued features are computed, including radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area-1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1). Additionally, attributes such as ID Number and Diagnosis (M=malignant, B=benign) are included, with 357 instances classified as Benign and 212 as Malignant.

B. Data Collection and Pre-processing

1) Data collection: Collected this 'WBCD' dataset from Kaggle (a renowned platform for dataset collection). This dataset contains many features related to breast cancer, which helps in determining whether it's Benign or Malignant.

2) Data exploration: Providing a comprehensive explanation of the dataset is crucial for a proper understanding of the data. To achieve this, this research conducted descriptive statistics, checked for missing values, and visualized distributions using box plots, heat maps, histograms, and correlation matrices. These techniques are invaluable for gaining insight into the dataset and effectively addressing any issues that may arise.

3) Cleaning: For cleaning purposes, the researches of this study removed the 'ID' and 'unnamed' columns as they are not necessary for cancer detection or prediction. After removing them, the dataset became more significant and accurate, leading to easy working processes for proper detection.

4) Feature selection: The feature selection method of this model is Chi-square(chi2). It helped to find 21 features to detect whether the actual cell is benign or malignant. Of the 569 records, 37% were classified as Malignant, accounting for 212 records. Conversely, 63% of the cells were classified as Benign, resulting in 357 records. This approach focuses on extracting features that are both informative and straightforward for cell determination. Fig. 2 shows percentage of patients.

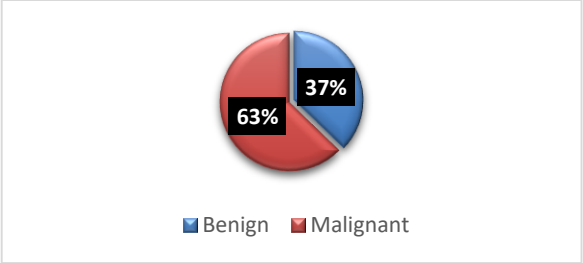

Fig. 2. Percentage of patients.

5) Train-test split: The dataset was divided into two parts: (1) Training and (2) Testing. 70% of the data is allocated for training, as this portion teaches the model and determines the actual results. The remaining 30% of the dataset was reserved for testing, providing insights into the model's performance and evaluating the training process's effectiveness.

6) Scaling: Scaling method uses data transform technique to fit within a specific scale. In this case, the research used standardization as the scaling method, which involves converting data to have a mean of zero and a standard deviation of 1 so that all features can be expressed in a comparable way.

7) Encoding: For encoding, we've used data labeling. This approach assigns a unique number to label each class inside a definite feature. It expresses records by changing them into a numerical layout, ensuring compatibility with the algorithms' requirement for numerical inputs, and even maintaining the specific feature's information.

8) Final data set: This is the final dataset resulting from an exhaustive study. With 569 samples and numerous features, the model can effectively detect whether a cell is Benign or Malignant. The main approach involved selecting 21 features using the chi-square method, which proved instrumental in identifying cancer cells. Through various techniques and methods, this research successfully achieved accurate cancer detection and more.

C. Algorithms

1) Chi-square: It is [6] a feature selection technique to select the best correlational feature from independent variables. It is a well-performed method for feature selection in statistical datasets. Chi-square performs to determine the GOF (Goodness of it), which measures the closeness of the prediction from a hypothesis [6]. The formula of chi-square is:

$$x^2 = \sum(f_0 - f_e)/f_e \quad (1)$$

Where,

f_0= observed frequency

f_e =expected frequency when no relation existed between variables.

In Fig. 3, also shows the visualization of the importance of 30 features through chi2, among which we select the top 21 features for our model prediction.

2) K-nearest neighbor: In short, KNN is a non-parametric, supervised learning algorithm that works to classify similar points near one another and make an individual group of them. To measure the new Knearest point, the researchers of this study calculated with Euclidean distance. It's a distance measurement to deal with big datasets. [7] The equation is:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

3) Logistic regression: A binary classification process to work with a linear function. Here, the sigmoid function is used as it refers to the assumption or probability [5]. The equation of this is:

$$P(Y = 1|x_i) = \sigma(x_i^T W) = \frac{1}{1+e^{-(w_0+w_1x_{1,1}+w_2x_{2,2}+\cdots+w_dx_{i,d})}} \quad (3)$$

4) Multilayer perceptron: MLP is a supervised, feed-forward back-propagation network comprising multiple layers: input, output, and hidden layers. These layers play a crucial role in extracting essential information during learning and adjusting weights accordingly [5]. MLP employs a stimulation function across all neurons, calculated using the following formula [25]:

$$f(x_i) = \mathrm{b} + \sum_{i=1}^{n} x_i\, w_i \quad (4)$$

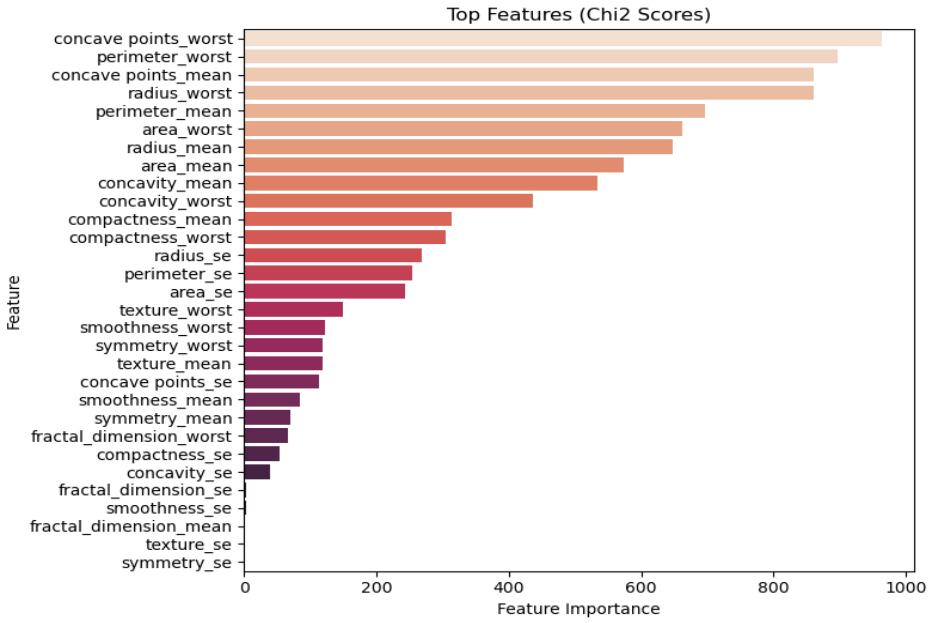

Fig. 3. Important features representation.

where,

x_i = inputs of incoming layers.

w_i = weights of hidden layers neurons.

b = initial weight.

5) Random forest: It is another type of supervised learning algorithm with a building block of machine learning, a way of making new data predictions based on previous data. RF is built with a set of decision trees of randomly chosen features, which gives the best prediction from the voting of every tree [25].

6) Adaboost: AdaBoost is an ensemble method that utilizes boosting, combining numerous "weak classifiers" to form a "strong classifier" by updating their weights iteratively. Training continues until a minimized error is achieved [11]. AdaBoost is a boosted classifier of the form:

$$F_T(\mathrm{x}) = \sum_{t=1}^{T} f_t\, x \tag{5}$$

Where each f_t is a weak learner that takes an object as input and returns a value indicating the class of the object [26].

7) Gradient boosting: GB is a numerical optimization technique that addresses classification and regression problems [26]. It operates sequentially, gradually improving weak learners by focusing on high-value data points. It can be defined as:

$$\mathrm{Y} = \mathrm{ax} + \mathrm{b} + \mathrm{e} \tag{6}$$

Where e represents the error and shows the inexplicable data [6].

8) XGboost: The XGB is a high-scalability decision tree that minimizes the loss function to get an additive expansion of the function [5]. XGB uses extensive and complex datasets to classify objects [7].

9) Bagging: Bagging is a way of reducing variance from noisy datasets by converting some random subsets into decision trees. It trains multiple instances of weak learners and finally predicts by averaging on regression and voting for classification, which tends to reduce overfitting and makes it more sustainable [27].

10)Boosting: Boosting is a sequential process of reducing errors in a predicted model. In this method, the base classifier allocates updated weight to the occurrences of misclassification, which improves the performance of the model sequentially [27].

11)Voting: Voting is a method of combining the prediction of multiple independent models. It could either be Soft voting or Majority Voting. It is useful when the base model shaves multiple predictions, like high or low, but can have the majority or average performance percentages [27].

12)Stacking: Stacking is a technique of taking outputs from multiple base models and passing them as an input of a Meta model for final prediction. It takes base predictions optimally and leads to better performance [27].

IV. Result

The proposed model of this study is a combination of machine learning algorithms, including ensemble methods. To preprocess the WDBC dataset, this study applied standardization and then used the Chi-square method for univariate feature selection, resulting in 21 effective features for classification. Initially, the dataset was divided into two parts: 70% for training and 30% for testing, using the train_test_split function from the Sklearn model selection package. The random state number 42 was used for this function. The researchers of this study developed the model using Python (version 3.11) and Anaconda (Anaconda Inc., Austin, TX, USA) as the software platform. Built-in functions such as Sklearn, Numpy, Pandas, Matplotlib, and Seaborn were utilized to conduct experiments and evaluate results. After splitting the dataset, ten different ML classification and regression algorithms were applied —KNN, SVM, MLP, NB, RF, DT, LR, XGB, AB, and GB—to train all the ML models on 70% of the training data. Subsequently, the remaining 30% of the data allowed the prediction of cancer cell types, assessing the research model's performance on unknown data. The Chi-square method significantly contributed to achieving an accuracy of over 98%, compared to PCA [6] [28], which yielded an accuracy of only 94%.

To determine the supremacy of any model, researches need to measure the performance via metrics. ROC can help to represent its performance; the higher the curve, the better performance is provided by the model [29].

1) Confusion matrix: A confusion matrix is a table describing the performance of a classification model based on test data whose original valid values are known before. Though it is simply stable, other parameters are slightly confusing (see Fig. 4).

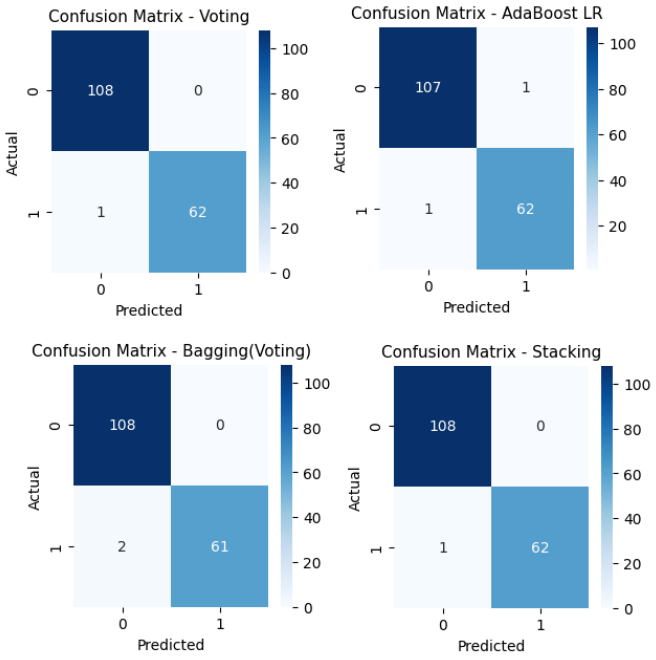

Fig. 4. Confusion matrix of ensemble techniques.

2) Accuracy: Accuracy measures the number of correctly detected Breast tumors [27]. It describes the model's correct output prediction and measures how accurately a model works so that the model can prove (see Fig. 6) itself more effective. High accuracy produces high success of models. The formula is given below:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (7)$$

3) Precision: It evaluates the accurate classification of positive samples and the true positive rate. This paper presents the valid Malignant rate, indicating the perfect positive output correctly identified by the model. This can be calculated using the formula below:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (8)$$

4) Recall: It is an output of total positive classes that confirms the correct prediction of a model. The recall is as preferable as higher. The formula for the recall is:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

5) Specificity: It measures the correct negative sample classification. It can be mentioned as the actual negative rate. The formula of it:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (10)$$

6) F1 score: When two models exhibit significant differences between precision and recall points—such as high precision and low recall, or vice versa—comparing them becomes challenging. This is where the F-score comes into play, as it aims to balance recall and precision simultaneously. The F-score reaches its maximum value when recall is equal to precision. The formula below can be used to calculate it:

$$\text{F1} - \text{Score} = \mathbf{2} \times \frac{Recall \times Precision}{(Recall + Precision)} \quad (11)$$

7) AUC-ROC: AUC mainly measures the ranking of good prediction; on the other hand, ROC is a graph showing the performance of all classification models (see Fig. 5).

Table II presents the accuracy, precision, recall, f1, f2, f3 scores, and AUC-ROC. MLP and LR achieved the highest accuracy prediction of 98.25%, followed by KNN, SVM, XGB, and RF with 97.08% accuracy. AB and GB both attained 95.91% accuracy, while DT and NB achieved 94.15% and 93.57% accuracy, respectively. The highest precision score of 0.9839 was obtained from MLP and LR. Additionally, recall, f1, f2, and f3 scores, and AUC-ROC yielded the highest values of 96.83%, 97.60%, 97.13%, 96.98%, and 97.95%, respectively. It is evident that when considering all parameters, MLP and LR performed the best as individual algorithms.

Following that, the ensemble algorithm is presented in Table III, showcasing the results of the research model after applying the Ensemble Technique. Once the top algorithms have been identified, they are utilized for the ensemble technique (shown in Table II). As SVM, LR, MLP, and XGBC worked well (shown in Table I), this study chose them for Voting and Stacking, which gave the highest accuracy of 99.42% as well as precision, recall, F1, and F2 scores of 1.0, 0.9841, 0.9920, and 0.9873, respectively. Despite having the lowest accuracy of NB and DT (shown in Table III), they performed much better after applying Bagging. So, it is evident that ensemble techniques are always more effective in performance. Fig. 7 shows performance of different EML algorithms.

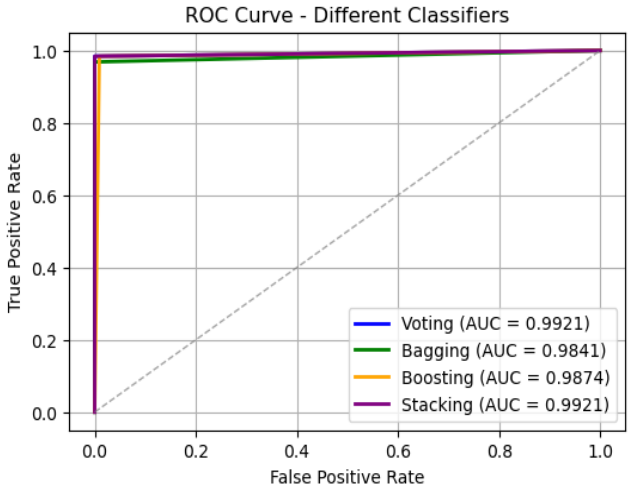

Fig. 5. ROC curve of different ensemble methods.

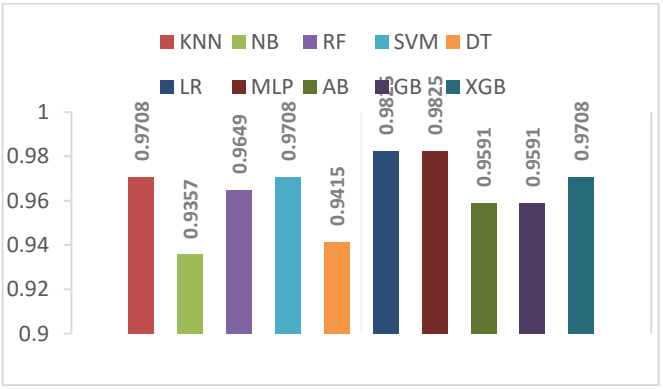

Fig. 6. Accuracy of ML algorithms.

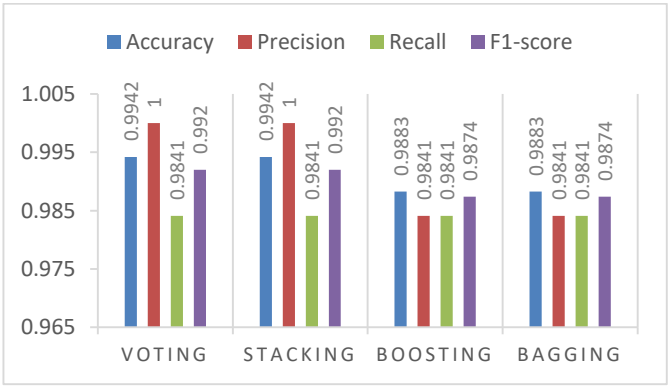

Fig. 7. Performance of different EML algorithms.

TABLE I. RELATED PAPERS ON MACHINE LEARNING AND ENSEMBLE TECHNIQUES IN BREAST CANCER DATASET (WBCD)

Author	Year	Technique	Accuracies (%)
E. Strelcenia *et al.* [5]	2023	ML Algorithm – LR, **DT**, RF, KNN, MLP, XGboost	96%, **98%**, 97%, 89%, 92%, and 94%.
V. Chaurasia *et al.* [15]	2020	Ensemble - ABC, GBC, RF, **ET**, Bagging, XGboost, Stacking (SVC, DT, LR, KNN, RF, NB)	94.73%, 93.85%, 94.72%, **95.17%,** 94.51%, 95.16, 92.65, and 98.24%.
M. Naji *et al.* [16]	2021	ML Algorithm – **SVM,** NB, C4.5, LR, RF Ensemble - **Majority Voting**	ML: **97.8%,** 92.6%, 93.1%, 96.8%, 97.1%, and 95.9%. Ensemble: **98.1%.**
M. Jabbar [17]	2021	ML Algorithm – SVM-NB, **AR-ANN**, FMM-CART, LP-SVM, et. Ensemble - Majority **Voting BN+RBF**	ML: 97.13%, **97.40%,** 97.29%, and 97.33%. Ensemble: **97.42%.**
T. Srinivas *et al.*[18]	2022	ML & Ensemble-KNN, LR, DT, RF, SVM, **SGD**, SMO, Gradient booster, AdaBoost M1, Logit Boost, Bagging	95%, 95%, 95%, 97%, 95%, **98%**, 97%, 97%, 95%, 96%, and 95%.
T. Mahesh *et al.*[19]	2022	ML Algorithm –NB, **AltTeee,** RF, RedEPT Ensemble - XGboost-NB, XGboost-AltDt, **XGboost-RF**, XGboost-RedEPT	ML:88.5%, **95.6%,** 94.5%, and 89.23% Ensemble: 81.55%, 96.5%, **98.2%,** and 82.25%
A. Assiri *et al.*[9]	2020	ML Algorithm – **LR,** SVM+SGD, MLP, DT, RF, SVM+SMO, KNN, NB. Ensemble – **Voting (Majority**, Average, Product, Minimum, Maximum)	ML: **98.25%,** 97.88%, 97.66%, 91.81%, 96.49%, 97.08%, 97.08%, and 91.81%. Ensemble: **99.42%,** 98.83%, 98.12%, 98.46%, and 99.41%.
U. Naseem *et al.*[8]	2022	Ensemble – Stacking **(SVM, LR, NB, DT) +ANN,** (SVM, LR, NB, RF) + ANN, (SVM, LR, RF, DT) +ANN, **(SVM, LR, RF, NB) + ANN**, (SVM, LR, RF) +ANN, (SVM, LR) +ANN with up sampling	Diagnosis: **98.83%**, 98.24%, 98.24%, 98.24%, 98.14%, and 96.46%. Prognosis: 84.70%, 88.13%, 84.74%, **88.33%,77.96%,** and 76.27%.
M. Elsadig *et al.*[20]	2023	ML & Ensemble- KNN, DT, **SVM**, RF, MLP, NB, STACK	92.9%, 92.3%, **97.0%**, 95.5%, 96.5%, 93.0%, 94.4%, and 96.3% for training-testing [70:30].
A. Khalid *et al.*[21]	2023	ML Algorithm – **RF**, DT, LR, KNN, LSVC, SVC	**96.49%,** 93.85%, 92.98%, 92.10%, 89.47%, and 87.71%.
T. Islam *et al.*[22]	2023	ML Algorithm –**LR**, RF, DT, GB, SVC, KNN, ABC, NB, GS, XGB Bagging- **LR**, RF, DT, GB, SVC, KNN, ABC, NB, GS, XGB Boosting - **LR**, RF, DT, GB, ABC, SVC, NB, XGB	ML: **95.6%,** 92.9%, 93.8%, 93.8%, 95.6%, 93.8%, 93.8%, 91.2%, 95.6%, and 92.1%. Bagging: **92.9%,** 92.1%, 92.1%, 92.9%, 92.1%, 92.9%, 92.1%, 90.3%, 92.9%, and 91.2%. Boosting: **95.6%,** 92.9%, 94.8%, 93.8%, 93.8%, 93.8%, 74.5%, and 58.7%.
M. Gupta *et al.*[23]	2018	ML Algorithm – **SVM,** KNN, DT, LR Ensemble – **Voting(soft)**	ML**: 93.98%,** 90.12%, 92.15%, and 89.12%. Ensemble: **97.88%.**
A. Bataineh *et al.*[24]	2019	ML Algorithm – **MLP**, KNN, CART, NB, SVM	**99.12%,** 95.61%, 93.85%, 94.73%, and 98.24%.

TABLE II. PERFORMANCE OF ML ALGORITHMS WITHOUT ENSEMBLE TECHNIQUE

ML Algorithm	Accuracy	Precision	Recall	F1 Score	F2 Score	F3 Score	Roc Auc
K-Nearest Neighbor	97.08%	0.9677	0.9524	0.9600	0.9554	0.9539	0.9669
Naïve Bayes	93.57%	0.9194	0.9048	0.9120	0.9076	0.9062	0.9292
Random Forest	96.49%	0.9672	0.9365	0.9516	0.9425	0.9395	0.9590
Support Vector Machine	97.08%	0.9531	0.9683	0.9606	0.9652	0.9667	0.9702
Decision Tree	95.32%	0.9231	0.9524	0.9375	0.9464	0.9494	0.9530
Logistic Regression	**98.25%**	**0.9839**	**0.9683**	**0.9760**	**0.9713**	**0.9698**	**0.9795**
Multilayer Perception	**98.25%**	**0.9839**	**0.9683**	**0.9760**	**0.9713**	**0.9698**	**0.9795**
AdaBoost	95.91%	0.9516	0.9365	0.9440	0.9395	0.9380	0.9544
Gradient Boosting	95.32%	0.9365	0.9365	0.9365	0.9365	0.9365	0.9497
XGboost	97.08%	0.9531	0.9683	0.9606	0.9652	0.9667	0.9702

TABLE III. PERFORMANCE OF ML ALGORITHMS WITH ENSEMBLE TECHNIQUE

Ensemble Technique	Model	Accuracy	Precision	Recall	F1 Score	F2 Score	Roc Auc
Voting (hard and soft)	Voting (SVM, MLP, LR, XGB)	**0.9942**	1.0	0.9841	0.9920	0.9873	0.9921
Stacking (meta =RF)	Stacking base (SVM, MLP, LR, XGB)	0.9942	1.0	0.9841	0.9920	0.9873	0.9921
AdaBoost	RF	0.9708	0.9833	0.9365	0.9593	0.9455	0.9636
	SVM	0.9766	1.0	0.9365	0.9672	0.9486	0.9683
	NB	0.9649	0.9385	0.9683	0.9531	0.9621	0.9621
	LR	0.9883	0.9841	0.9841	0.9841	0.9841	0.9874
	DT	0.9532	0.9104	0.9683	0.9385	0.9561	0.9563
AdaBoost Voting (hard)	Voting (AB+RF, AB+SVM, AB+LR)	0.9825	1.0	0.9524	0.9756	0.9615	0.9762

Gradient Boosting	Gradient Boosting (n_estimators=1000, learning rate=.3, subsample=.3, random state=42)	0.9766	0.9538	0.9841	0.9688	0.9779	0.9782
XGboost	XGBC (n_estimators=1000, learning rate=.1, subsample=.1, random state=42)	0.9883	1.0	0.9683	0.9839	0.9744	0.9841
Bagging	RF	0.9708	0.9833	0.9365	0.9593	0.9455	0.9636
	DT	0.9649	0.9524	0.9524	0.9524	0.9524	0.9623
	SVM	0.9825	0.9839	0.9683	0.9760	0.9713	0.9795
	MLP	0.9883	1.0	0.9683	0.9839	0.9744	0.9841
	LR	0.9825	0.9839	0.9683	0.9760	0.9713	0.9795
	KNN	0.9649	0.9672	0.9365	0.9516	0.9425	0.959
	NB	0.9708	0.9677	0.9524	0.9600	0.9554	0.9669
Bagging Voting (hard)	Voting (BC+RF, BC+SVM, BC+MLP, BC+LR, BC+NB)	0.9883	1.0	0.9683	0.9839	0.9744	0.9841

V. DISCUSSION

Table IV below provides a comprehensive comparison between existing models and the model created within this study, all applied to the same 'WBCD' dataset. Despite the numerous proposals on breast cancer, these represent some recent works by various researchers. The research model outperforms many of these in multiple aspects. This comparison is organized based on accuracy, precision, recall, specificity, F1 score, AUC-ROC, and train-test performance of the existing models, allowing for an overall performance contrast.

The effectiveness of the stacking algorithm has been demonstrated by author A. Abdar et al. [30], achieving over 98% performance in accuracy, precision, and recall. Another model, SELF by A. Jakhar et al. [26], attained 98.80% accuracy using this technique, along with high precision, recall, F1-score, and AUC-ROC scores exceeding 99% on an 80:20 training-testing ratio. On the other hand, the recent model OSEL by author M. Kumar et al. [7] garnered attention with the highest accuracy of 99.45%. However, the precision, recall, and F1-score scores were 99%, 98%, and 94%, respectively, leading to somewhat less satisfaction. In terms of the research model, stacking achieved an accuracy of 99.42%, with 100% precision, 98.41% recall, 100% specificity, 99.2% F1-score, and AUC-ROC scores on a 70% training dataset and 30% testing dataset, showcasing the highest overall performance among the models

A voting classifier is another technique aimed at enhancing performance, as described by Q. Nguyen et al. [14], achieving an accuracy of 98.83% and nearly 99% in other scores. However, this performance was surpassed by another model by M. Murtirawat et al. [3], utilizing the same classifier and achieving 99.30% accuracy, along with 100% precision, 97.8% recall, and 98.87% F1-score on a 75:25 train-test ratio. A recent paper by A. Assiri et al. [9] provided a remarkable accuracy of 99.42%, including more than 99% precision, recall, and F1-score with a complex voting ensemble technique, which was the highest reported at that time. Nevertheless, the approach of this study managed to achieve an accuracy of 99.42% along with other parameter scores such as precision, recall, specificity, F1-score, and AUC-ROC of 1, .98, 1, 99.2, and 99.2, respectively, using both hard and soft voting techniques, setting a new benchmark.

In a paper by author N. Mashudi et al. [12], AB achieved an accuracy of 98.77%, along with 99.42% precision and 97.66% specificity. In comparison to other ensemble techniques like bagging and boosting, this study surpassed recent papers with 98.83% accuracy and high scores in other parameters.

TABLE IV. COMPARISON WITH EXISTING WORK

Work	Author	Model	Accuracy	Precision	Recall	Specificity	F1-Score	Auc Roc	Train Test
EXISTING WORKS	M. Kumar *et al.*[7]	OSEL	99.45%	0.99	0.98	-	0.94	-	-
	A. Assiri *et al.*[9]	Voting (hard)	99.42%	0.9940	0.994	-	0.994	-	70:30
	M. Murtirawat *et al.*[3]	Voting	99.30%	1.0	0.978	-	0.9887	-	75:25
	Q. Nguyen *et al.*[14]	Voting	98.83%	0.99	0.99	-	0.99	0.9844	70:30
	A. Jakhar *et al.*[26]	SELF Stacking	98.80%	0.9909	0.9909		0.9909	0.9906	80:20
	N. Mashudi *et al.*[12]	AdaBoost	98.77%	0.9944	-	0.9766	-	-	
	A. Abdar *et al.*[30]	Stacking	98.07%	0.9810	0.9810	-	0.9810	0.9760	K=10
OUR WORKS	Voting (hard and soft)	Voting (SVM, MLP, LR, XGBC)	99.42%	1.0	0.9841	1.0	0.9920	0.9921	70:30
	Stacking(meta=rf)	Stacking base (SVM, MLP, LR, XGBC)	99.42%	1.0	0.9841	1.0	0.9920	0.9921	70:30
	Boosting	AB(LR)	98.83%	0.9841	0.9841	0.9841	0.9841	0.9874	70:30
	Bagging	Voting (BC_RF, BC_SVM, BC_MLP, BC_LR)	98.83%	1.0	0.9683	1.0	0.9839	0.9841	70:30

VI. CONCLUSION

Breast cancer stands as a formidable cause of mortality among women, underscoring the critical need for early detection. The challenge lies not only in uncovering the presence of cancer but also in doing so at its nascent stage, thereby curbing the mortality rate. The amalgamation of medical science with ML classifiers has emerged as a powerful tool in tackling this challenge. Over time, it has become evident that enhancing a model's predictive performance significantly aids in this realm. Ensemble techniques, taking a step further, amalgamate multiple classification methods, thereby exhibiting superior performance. This study traverses this path, showcasing the efficacy of breast cancer prediction with an accuracy of 99.42%, along with precision, recall, F1, F2 score, and AUC-ROC scores of 99%, 99%, 99%, and 99%, respectively. Positioned as one of the premier models, it outshines existing ones in both early detection capability and performance prowess. Through rigorous training and testing, the model's efficiency on the WBCD dataset is attested, adeptly discerning between Benign and Malignant cases.

Looking ahead, this model holds promise for further enhancement by integrating new optimization techniques. Researchers exploring additional ensemble techniques stand poised to achieve even more noteworthy results. Ultimately, the proposed ensemble learning system promises to become an indispensable tool for cancer specialists, facilitating the early recognition of breast cancer.

REFERENCES

[1] "Breast cancer." Accessed: Dec. 17, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] "Breast cancer takes 6,844 lives in Bangladesh every year: report | The Daily Star." Accessed: Dec. 17, 2023. [Online]. Available: https://www.thedailystar.net/city/news/breast-cancer-takes-6844-lives-bangladesh-every-year-report-1812172

[3] R. Murtirawat, S. Panchal, V. K. Singh, and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," in Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, 2020, pp. 534–540. doi: 10.1109/ICESC48915.2020.9155783.

[4] "Breast Cancer: Introduction | Cancer.Net." Accessed: Dec. 17, 2023. [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/introduction .

[5] E. Strelcenia and S. Prakoonwit, "Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study," BioMedInformatics, vol. 3, no. 3, pp. 616–631, Sep. 2023, doi: 10.3390/biomedinformatics3030042.

[6] R. Shafique et al., "Breast Cancer Prediction Using Fine Needle Aspiration Features and Upsampling with Supervised Machine Learning," Cancers (Basel)., vol. 15, no. 3, Feb. 2023, doi: 10.3390/cancers15030681.

[7] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized Stacking Ensemble Learning Model for Breast Cancer Detection and Classification Using Machine Learning," Sustain., vol. 14, no. 21, Nov. 2022, doi: 10.3390/su142113998.

[8] U. Naseem et al., "An Automatic Detection of Breast Cancer Diagnosis and Prognosis Based on Machine Learning Using Ensemble of Classifiers," IEEE Access, vol. 10, no. July, pp. 78242–78252, 2022, doi: 10.1109/ACCESS.2022.3174599.

[9] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," J. Imaging, vol. 6, no. 6, 2020, doi: 10.3390/JIMAGING6060039.

[10] V. Nemade and V. Fegade, "Machine Learning Techniques for Breast Cancer Prediction," in Procedia Computer Science, Elsevier B.V., 2022, pp. 1314–1320. doi: 10.1016/j.procs.2023.01.110.

[11] M. T. Ramakrishna, V. K. Venkatesan, I. Izonin, M. Havryliuk, and C. R. Bhat, "Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data," Entropy, vol. 25, no. 2, Feb. 2023, doi: 10.3390/e25020245.

[12] N. A. Mashudi, S. A. Rossli, N. Ahmad, and N. M. Noor, "Comparison on Some Machine Learning Techniques in Breast Cancer Classification," in Proceedings - 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences, IECBES 2020, 2020, pp. 499–504. doi: 10.1109/IECBES48179.2021.9398837.

[13] M. Momtahen, S. Momtahen, R. Remaseshan, and F. Golnaraghi, "Early Detection of Breast Cancer using Diffuse Optical Probe and Ensemble Learning Method," in 2023 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization, NEMO 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 139–142. doi: 10.1109/NEMO56117.2023.10202520.

[14] Q. H. Nguyen et al., "Breast Cancer Prediction using Feature Selection and Ensemble Voting," in Proceedings of 2019 International Conference on System Science and Engineering, ICSSE 2019, IEEE, 2019, pp. 250–254. doi: 10.1109/ICSSE.2019.8823106.

[15] V. Chaurasia and S. Pal, "Applications of Machine Learning Techniques to Predict Diagnostic Breast Cancer," SN Comput. Sci., vol. 1, no. 5, 2020, doi: 10.1007/s42979-020-00296-8.

[16] M. A. Naji, S. El Filali, M. Bouhlal, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier," Procedia Comput. Sci., vol. 191, pp. 481–486, 2021, doi: 10.1016/j.procs.2021.07.061.

[17] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning," Eng. Appl. Sci. Res., vol. 48, no. 1, pp. 65–72, 2021, doi: 10.14456/easr.2021.8.

[18] T. Srinivas et al., "Novel Based Ensemble Machine Learning Classifiers for Detecting Breast Cancer," Math. Probl. Eng., vol. 2022, 2022, doi: 10.1155/2022/9619102.

[19] T. R. Mahesh, V. Vinoth Kumar, V. Muthukumaran, H. K. Shashikala, B. Swapna, and S. Guluwadi, "Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer," J. Sensors, vol. 2022, 2022, doi: 10.1155/2022/4649510.

[20] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," Int. J. Electr. Comput. Eng., vol. 13, no. 1, pp. 736–745, 2023, doi: 10.11591/ijece.v13i1.pp736-745.

[21] A. Khalid et al., "Breast Cancer Detection and Prevention Using Machine Learning," Diagnostics, vol. 13, no. 19, pp. 1–21, 2023, doi: 10.3390/diagnostics13193113.

[22] T. Islam, A. B. Akhi, F. Akter, M. N. Hasan, and M. A. Lata, "Prediction of Breast Cancer using Traditional and Ensemble Technique: A Machine Learning Approach," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, pp. 867–875, 2023, doi: 10.14569/IJACSA.2023.0140692.

[23] M. Gupta and B. Gupta, "An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP)," 2018 11th Int. Conf. Contemp. Comput. IC3 2018, pp. 1–3, 2018, doi: 10.1109/IC3.2018.8530572.

[24] A. Al Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," Int. J. Mach. Learn. Comput., vol. 9, no. 3, pp. 248–254, Jun. 2019, doi: 10.18178/ijmlc.2019.9.3.794.

[25] Naveen, R. K. Sharma, and A. Ramachandran Nair, "Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models," 2019 4th IEEE Int. Conf. Recent Trends Electron. Information, Commun. Technol. RTEICT 2019 - Proc., pp. 100–104, 2019, doi: 10.1109/RTEICT46194.2019.9016968.

[26] A. K. Jakhar, A. Gupta, and M. Singh, "SELF: a stacked-based ensemble learning framework for breast cancer classification," Evol. Intell., pp. 0–29, 2023, doi: 10.1007/s12065-023-00824-4.

[27] M. S. Al Reshan et al., "Enhancing Breast Cancer Detection and Classification Using Advanced Multi-Model Features and Ensemble Machine Learning Techniques," Life, vol. 13, no. 10, p. 2093, 2023, doi: 10.3390/life13102093.

[28] V. Rupapara, F. Rustam, A. Ishaq, E. Lee, and I. Ashraf, "Chi-Square and PCA Based Feature Selection for Diabetes Detection with Ensemble Classifier," Intell. Autom. Soft Comput., vol. 36, no. 2, pp. 1931–1949, 2023, doi: 10.32604/iasc.2023.028257.

[29] B. R. Roy, M. Pal, S. Das, and A. Huq, "Comparative study of machine learning approaches on diagnosing breast cancer for two different dataset," 2020 2nd Int. Conf. Adv. Inf. Commun. Technol. ICAICT 2020, no. November, pp. 29–34, 2020, doi: 10.1109/ICAICT51780.2020.9333507.

[30] M. Abdar et al., "A new nested ensemble technique for automated diagnosis of breast cancer," Pattern Recognit. Lett., vol. 132, pp. 123–131, 2020, doi: 10.1016/j.patrec.2018.11.004