

RESEARCH ARTICLE

Toward Improving Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm

AMREEN BATOOL¹ AND YUNG-CHEOL BYUN²¹Department of Electronic Engineering, Institute of Information Science and Technology, Jeju National University, Jeju-si 63243, South Korea²Department of Computer Engineering, Major of Electronic Engineering, Institute of Information Science and Technology, Jeju National University, Jeju-si 63243, South Korea

Corresponding author: Yung-Cheol Byun (ycb@jejunu.ac.kr)

This work was supported in part by the Ministry of Small and Medium-Sized Enterprises (SMEs) and Startups (MSS), South Korea, under the Regional Specialized Industry Development Plus Program (Research and Development) Supervised by the Korea Institute for Advancement of Technology (KIAT) under Grant S3246057; in part by KIAT funded by the Korean Government [Ministry Of Trade, Industry & Energy (MOTIE)] (The Establishment Project of Industry-University Fusion District) under Grant P0016977; and in part by the Regional Innovation Strategy (RIS) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE).

ABSTRACT Over the past decade, breast cancer has been the most common type of cancer in women. Different methods were proposed for breast cancer detection. These methods mainly classify and categorize malignant and Benign tumors. Machine learning is a practical approach for breast cancer classification. Data mining and classification are effective methods to predict and categorize breast cancer. The optimum classification for detecting Breast Cancer (BC) is ensemble-based. The ensemble approach involves using multiple ways to find the best possible solution. This study used the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. We created a voting ensemble classifier that combines four different machine learning models: Extra Trees Classifier (ETC), Light Gradient Boosting Machine (LightGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA). The proposed ELRL-E approach achieved an accuracy of 97.6%, a precision of 96.4%, a recall of 100%, and an F1 score of 98.1%. Various output evaluations are used to evaluate the performance and efficiency of the proposed model and other classifiers. Overall, the recommended strategy performed better. Results are directly compared with the individual classifier and different recognized state-of-the-art classifiers. The primary objective of this study is to identify the most influential ensemble machine learning classifier for breast cancer detection and diagnosis in terms of accuracy and AUC score.

INDEX TERMS Breast cancer, classification, machine learning, voting classifier, ensemble learning.

I. INTRODUCTION

Breast cancer is a disease that grows in the human body through abnormal cells. Men are less affected than women. Breast cancer cases calculated in women are 287,850 in 2022 and 2,710 in men, according to the American Cancer Society (ACS) [1]. However, the breast cancer death rate is also high in women; the death rate in men is 530, and in

women, 43,250. The disease mainly affects women and can be diagnosed at any stage. If diagnosed at an early stage, the survival chances are increased, but in the advanced stage, the survival chances in a breast cancer patient are reduced. There are many types of breast cancer. Breast cancer types also refer to whether it has spread or not and whether it is invasive or non-invasive. Invasive cancer spreads to the lymph nodes or milk ducts. Lobules to other breast tissues, whereas Non-invasive ones cannot invade others. Tissues of breast cancer that are non-invasive are called “in situ” and

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman¹.

may remain dormant for an extended period of a lifetime [2]. Moreover, breast cancer affects 40.3% of the population in Indonesia and dies 16.6% of those diagnosed [3], [4]. Drinking and smoking excessively, as well as an unhealthy diet, increase the risk of breast cancer. It is predicted that breast cancer will increase by 2% in 2030 [5]. Early diagnosis of BC can significantly improve the prognosis and survival probability by allowing patients to receive timely clinical treatment [6]. In recent literature, classification techniques such as RF, SVM, KNN, and XGB classifiers have been used [7]. Several researchers conducted research for the prediction of breast cancer using various machine-learning techniques. Regarding the researcher's concern, the RF and ET strategies use decision trees as proper classifiers to attain the ultimate classification. This work evaluated the quality of each algorithm data classification [8] in terms of efficiency and effectiveness. In [9], the author proposed an ensemble learning-based voting classifier that combines the logistic regression and stochastic gradient descent classifier to detect breast cancer patients accurately.

Moreover, the motivation of this study is that the ensemble classifier method used in the previous research is still limited to detecting and classifying breast cancer. One of the biggest challenges in healthcare research is the timely and accurate detection of various diseases [10], [11], [12]. Breast cancer is one of the significant causes of death for women worldwide, which has prompted a lot of interest in the health field. Detection and classification of breast cancer in its early stages is the primary objective of this study, which uses machine learning methods for accurate classification and evaluation in terms of accuracy.

Therefore, this article compares the performance of different classifiers on the breast cancer dataset. While various machine learning classifiers like RF, SVM, and KNN have been explored, the study introduces a novel ensemble-based approach, including ETC, LightGBM, RC, and LDA. Addressing class imbalance, the research assesses the proposed ensemble against state-of-the-art methods. The study's contributions lie in evaluating strategies, offering a novel ensemble framework, handling class imbalance effectively, and comparing the model's performance. Breast cancer detection is a crucial challenge in healthcare as it is one of the leading causes of death for women worldwide. This study aims to use machine learning to improve early detection and refine breast cancer classification methods. The research will introduce an adaptive voting ensemble algorithm and thoroughly evaluate its performance. This way, it can contribute to better patient outcomes and advance the field of medical decision-making.

The main contributions of the proposed study are given below.

- This study evaluates machine learning approaches and algorithms to determine the best strategy for breast cancer classification.
- Proposed a novel ensemble-based framework for predicting breast cancer. Which includes the Extra Trees

Classifier (ETC), Light Gradient Boosting Machine (LightGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA).

- Breast cancer data often has a class imbalance, with a higher number of benign cases than malignant cases. The proposed study could demonstrate how to handle class imbalance and improve classification performance effectively.
- The proposed study could compare the performance of the voting ensemble model with other state-of-the-art breast cancer classification methods.

The subsequent sections of this manuscript are structured as follows: Section I introduction of the manuscript Section II offers an overview of relevant research in breast cancer classification and ensemble learning. In Section III, we delve into the methodology underpinning our adaptive voting ensemble algorithm, highlighting its adaptive framework and distinctive features. Section IV meticulously describes exploratory data analysis, encompassing dataset details, evaluation metrics, and reference algorithms for performance benchmarking. The ensuing Section V discusses the empirical findings, shedding light on the strengths and limitations observed during the evaluation process. Finally, Section VI encapsulates our conclusions, explores the implications of our research, and outlines potential avenues for future exploration.

II. RELATED WORK

Machine learning algorithms are used to predict an accurate model for breast cancer, but selecting the best classifier is a critical challenge. Data scientists produced excellent outcomes when they applied different algorithms to various medical datasets [13]. Many scientists have worked on designing and assessing breast cancer detection methods. Many researchers predict breast cancer using multiple machine learning algorithms such as Decision Tree [14], NN [9], RF [15], LR [16], Naïve Bayes [17], SVM [18]. In this article, [19] author employed various sorts of classifiers. Author [20] conducted a comparative analysis that included several classifiers and anticipated that the SVM without the rapid co-relation-based Streamlines provides the maximum accuracy of 97%. The author in [21] uses logistic regression for categorization purposes. KNN, SVM, and RFE classifications provide automatic digital data and facts for breast cancer diagnosis [22]: linear Regression algorithm and Machine Learning train modules to classify a breast cancer dataset.

Moreover, in [21], article classification accuracy is 95%, and the author achieved the accuracy using texture classification and maximum perimeter. The authors of [23] and [24] presented a method for detecting and characterizing cell structure. The study [25] on breast cancer categorized as C3 and C4 on fine needle aspiration cytology aims to correlate with the histopathology examination. This [26] study compared different classification and clustering strategies. According to the findings, classification algorithms beat

clustering methods. Similarly, [27], [28], [29], [30], [31], and [32] and Bala et al. [33], [34] have elaborated soft computing, data mining, and machine learning techniques for diabetes and thunderstorm classification, respectively. In [35], the author compared the Bayesian Network, Random Forest, and Support Vector Machine algorithms and found that the Bayesian Network produced the best results. Bhat et al. [36] created an algorithm that allows adaptive resonance theory to be used in breast cancer research. The best-performing models from previous studies using the Wisconsin Breast Cancer Dataset for breast cancer detection are in Table 1.

The table provides a comprehensive overview of studies conducted in breast cancer classification using various machine learning algorithms. Each row in the table corresponds to a specific research, highlighting the year of the study, the algorithms employed, the advantages observed, and the limitations encountered. In 2021, one study evaluated the performance of Support Vector Machines (SVM) and Random Forest (RF) classifiers. The study emphasized using a limited number of features in the classification process. An investigation into ensemble methods was conducted in 2021 using a Stacking Classifier. While this approach showed promise in improving classification outcomes, it was noted that the complexity of the ensemble models was substantial. In 2021, a study explored a combination of classifiers, including Multi-Layer Perceptron (MLP), Sequential Minimal Optimization (SMO), Naïve Bayes (NB), and J48, both individually and as ensembles. While ensembles exhibited good performance, it was observed that complexity increased significantly when using more than two ensemble classifiers for predictions. In 2023, a study focused on the Averaged Perceptron classifier and its impact on false-positive and false-negative predictions. The study highlighted the importance of threshold selection in influencing these prediction outcomes.

Another investigation in 2023 emphasized the challenges posed by imbalanced datasets in logistic regression models for breast cancer classification, which could lead to biased classification results. In 2022, a study explored using K-Nearest Neighbors (KNN), Random Forest (RF), and Naïve Bayes algorithms to detect additional illnesses and provide insights into the nature of breast cancer. However, it was noted that accurately detecting breast cancer remained a challenging task. An approach using AdaBoost and Synthetic Minority Over-sampling Technique (SMOTE) was studied in 2022 to address class imbalance. While effective in dealing with imbalanced classes, the study acknowledged problems related to classification boundary definitions. In 2023, the classification of breast cancer microarray data using Random Forest (RF), Extra Trees (ET), Support Vector Machines (SVM), and Cross-Validation (CV) was explored. This study identified limited optimal features that could lead to improved classification accuracy. Finally, in the same year, a study employed Radial Basis Function (RBF) and Support Vector Machines (SVM) to extract more representative features for breast cancer classification. However, this architecture

was found to be challenging when applied to multi-class classification tasks. These studies contribute insights into the strengths and limitations of various machine learning algorithms for breast cancer classification, addressing issues such as feature selection, class imbalance, and the complexities of ensemble methods.

III. MATERIALS AND METHODS

This section delineates the dataset and classification models employed to enhance classifier compression. The outlined approach, ELRL-E, is illustrated in Figure 1. Our proposed methodology encompasses four key categories: Preprocessing, Training data, Ensemble classifiers, and validation. In the preprocessing, we perform an Exploratory Data Analysis (EDA) analysis process that involves visually and statistically exploring and summarizing the main characteristics, patterns, and relationships within a dataset and extracting features to find a correlation between features and optimized parameters. We used the grid search hyperparameter tuning technique to maximize the model's performance with the right combination of hyperparameters. As part of the training section, we trained four models and adapted the training data results into ensemble models using the voting classifier.

A. DATASET

The dataset is obtained from the Wisconsin Breast Cancer Dataset (WBCD) Diagnostic [45].

This dataset contains 569 patients, each characterized by 32 features. The first feature is a unique identifier, representing the patient ID in the subsequent 31 instances. The enhanced process that applied the dataset reduced the number of features. Accordingly, the top 32 features with considerable weight have been selected, and the other features (redundant and unweighted) need to be addressed. These features provide real-valued measurements that contribute to understanding the cell nuclei's properties. Among the chosen features are key parameters such as radius mean (f1), texture mean (f2), perimeter mean (f3), area mean (f4), smoothness mean (f5), concavity mean (f6), concave point mean (f7), symmetry mean (f8), and fractal dimension mean (f9). These features collectively contribute to a nuanced understanding of the characteristics and behaviors of cell nuclei, providing valuable insights for further analysis. Each instance in the dataset is assigned a label indicating whether the breast mass is classified as benign or malignant. A total of 357 are labeled as benign, indicating non-cancerous conditions. In contrast, the remaining 212 are labeled as malignant, signifying the presence of cancer. Table 4 explains the feature description.

Table 2 provides detailed information about the dataset, including its features and classes. The 70% of the dataset is used for training, while the remaining 30% is kept for testing. This division ensures that the classification models are evaluated fairly and comprehensively. In the training phase, the models study the patterns and relationships within most of the data, which helps them make predictions on new, unseen data. The model's performance is evaluated on

TABLE 1. Summarising literature related work of breast cancer.

Ref	Year	Algorithms	Advantages	Limitations
[37]	2021	SVM, RF	Several machine learning classifiers are evaluated.	A limited number of features are used
[38]	2021	Stacking Classifier	Extensive complexity
[39]	2021	MLP, SMO, NV, J48	Testing of ensembles, as well as individual classifiers, was performed well.	Complexity increases when more than two ensemble classifiers are used for prediction.
[40]	2023	Averaged-perceptron classifier	The investigation has also signified the effect of threshold on false positive and false-negative prediction	Signified the effect of threshold on false negative or false positive prediction
[41]	2023	Logistic regression mode	Logistic regression mode
[42]	2022	KNN, RF, Naïve Bayes	Detection of additional illnesses and providing formation about the nature of cancer	Difficult precisely detecting breast cancer sickness
[43]	2022	AdaBoost, SMOTE	Dealing with imbalanced classes	The problem occurring in classification boundaries accuracy
[44]	2023	RF, ET, SVM, CV	Classify breast cancer microarray data to normal and relapse	Limited optimal features and better classification accuracies
[41]	2022	RBF, SVM	Employed to provide more representative features of breast cancer	This architecture is difficult to predicted multi-class classification tasks

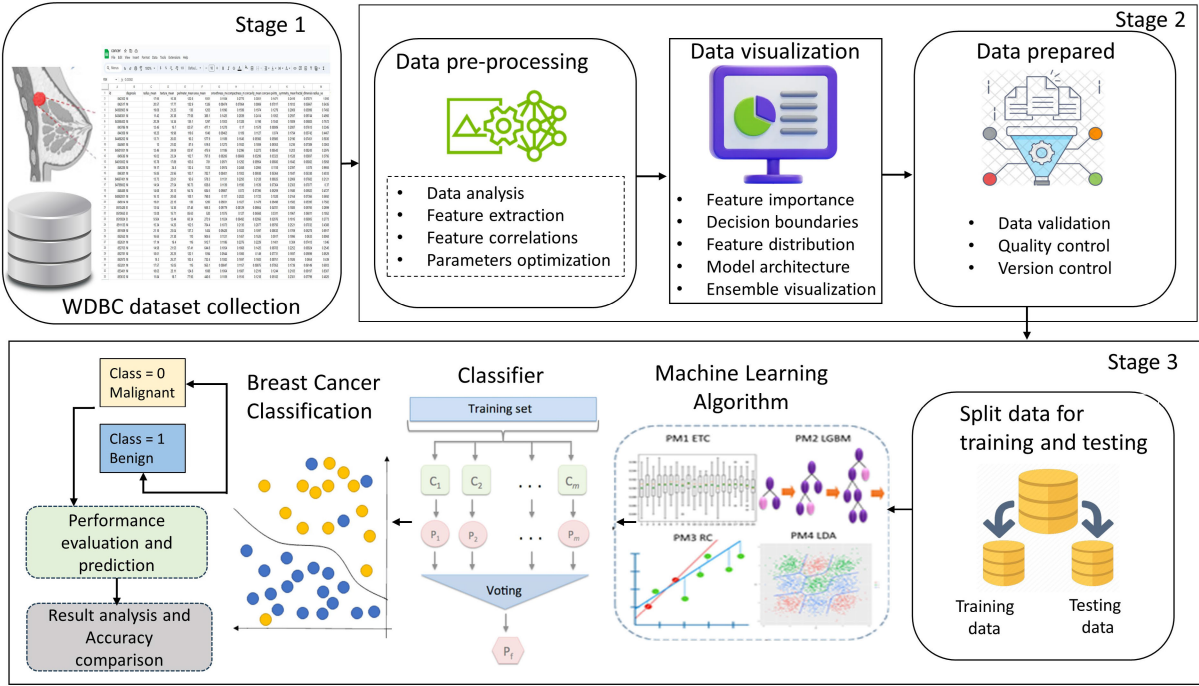


FIGURE 1. Proposed Model Overview: A concise examination of the key components and methodologies employed in the proposed model. This overview provides a high-level understanding of the model's structure, algorithms, and intended contributions to the addressed problem or task.

independent data during testing to determine its effectiveness and generalizability.

Figure 2 shows a correlation bar-plot between diagnosis and dataset attributes. The proposed model has 32 attributes

that correlate with each other. The correlation is individual between the diagnosis outcome and every dataset attribute. Some of the features are negatively correlated. In this bar graph, the four attributes are as *smoothness_se* negatively

TABLE 2. Detailed description of dataset.

Category	Total Sample	Training Sample	Testing Sample	Label
Malignant	212	148	64	1
Benign	357	250	107	2

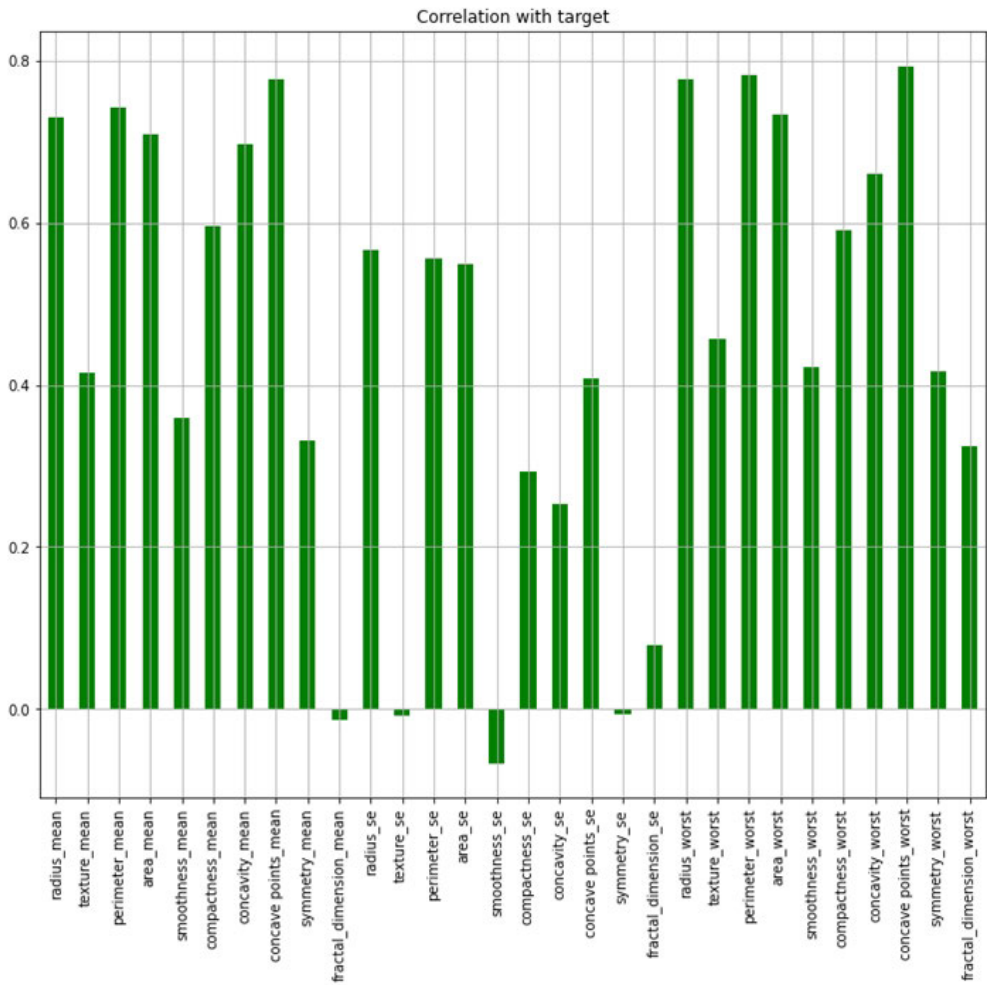


FIGURE 2. Correlation Bar-Plot with Target Features. This visual representation illustrates the correlation between various features.

correlates with the correlation barplot diagnosis, and the *fractal_dimension_mean*, *symenetry_se*, and *symmetry_se* are associated significantly less negatively correlated. Other remaining attributes are highly positively correlated. Afterward, metrics such as standard error, mean, and maximum are calculated for the 10 characteristics, resulting in 30 features. Table 3 presents the computed metrics, which represent the tumor features of the WBCD database. Further information on these features can be found in [46].

IV. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is an important step in understanding and preparing the data for breast cancer classification. Some of the key steps involved in EDA for breast cancer classification include:

- Data Understanding: This step involves understanding the structure of the data and the different types of available variables. It can include tasks such as reviewing the data dictionary and variable definitions.
- Data Cleaning: This step involves identifying and correcting errors and inconsistencies in the data. This can include filling in missing values, removing outliers, and updating data entry errors.
- Data Visualization: This step involves creating visualizations of the data to gain insights and identify patterns. This can include creating histograms, scatter plots, and box plots to visualize the distribution of different variables.
- Data Transformation: This step involves transforming the data to make it suitable for analysis. This can include

TABLE 3. Detailed description of dataset features.

F number	Feature	Mean	S.Error	Max
1	radius_mean	6.98-28.11	0.112-2.873	7.93-36.04
2	texture_mean	9.71-39.28	0.36-4.89	12.02-49.54
3	perimeter_mean	43.79-188.50	0.76-21.98	50.41-251.20
4	area_mean	143.50-2501.00	6.80-542.20	185.20-4254.00
5	smoothness_mean	0.053-0.163	0.002-0.135	0.071-0.223
6	concavity_mean	0.091-0.345	0.000-0.396	0.000-1.252
7	concave_point_mean	0.000-0.427	0.000-0.053	0.000-0.291
8	symmetry_mean	0.106-0.304	0.008-0.079	0.157-0.664
9	fractal_dimension_mean	0.050-0.097	0.001-0.030	0.055-0.208

tasks such as normalizing or scaling the data and creating dummy variables for categorical variables.

- **Feature Selection:** This step involves selecting the most relevant features for the classification task. This can include using correlation and mutual information to identify the features most strongly correlated with the outcome variable.

Initially, the average of the distributions for each feature was used to determine the statistics of 32 features, among which only 9 attributes were selected and extracted. The graph illustrates each attribute's standard deviations and ranges for the most significant characteristics from the real-valued dataset. It demonstrates the distributions of the 9 attributes with the mean. The distribution is fairly normal in most of the dataset. We create a histogram to visualize the correlations between the mean features provided. Fig. 3 displays the relationships between the total and selected attributes. The relationship between radius and perimeter should be linear, while the relationship between radius and area should be polynomial. In addition, other characteristics show linear correlations. We will analyze these characteristics using feature selection to investigate their relationship with diagnosis values. Selecting 9 attributes from a pool of 32 for breast cancer classification involves careful consideration of criteria and methods to ensure the chosen features are relevant and contribute significantly to the classification task. Correlation analysis helps identify attributes with a solid connection to the target variable while avoiding high correlations among selected features to maintain model interpretability. Information gain metrics assist in quantifying the importance of each attribute, and a variance threshold eliminates low-variance features. Additionally, recursive feature elimination (RFE) iteratively selects features based on their impact on model performance. The selection methods include filter methods or tree-based techniques. At this stage, the dataset is split into training and testing the data for calculating the covariance between the models.

A. DATA PROCESSING AND PERFORMANCE METRICS

Data preprocessing is preparing data for use in a machine-learning model. This can include cleaning and formatting the data, filling in missing values, and normalizing the data. Performance metrics are used to evaluate the effectiveness of a machine-learning model. The specific metric used will

depend on the task being performed and the type of model being used. For example, classification models may use metrics such as accuracy, precision, recall, and F1 score, while regression models may use metrics such as AUC-ROC.

B. CLASSIFICATION ALGORITHM

The suggested design aims to enhance machine learning algorithms, establishing an initial breast cancer detection model capable of predicting cancer types as benign or malignant [47]. Recent research underscores the effectiveness of machine and deep learning as precious methods for classifying breast cancer. Using all machine learning classifiers in this experiment produced promising results for predicting breast cancer. Algorithm 1 of this research is presented below for better understanding.

Algorithm 1 Working Procedure of Breast Cancer Prediction

Input: UCIMachine LearningRepository Breast Cancer Dataset

Output: Predicted value Malignant or Benign

1. Begin
2. data \leftarrow load dataset
 - a. if data. value is equal to NaN or empty
 - b. replace NaN or missing_value
3. pre-processing:
 - a. if data. target is equal to M
 - b. replace M with 0
 - d. else
 - e. replace B with 1
4. x \leftarrow data.drop[target]
5. y \leftarrow data.target
6. x1, x2, y1, y2 \leftarrow split_data of x and y
7. model \leftarrow train_model using x1 and y1
8. predict \leftarrow testing_model using x2 and y2
9. s_x \leftarrow scaling data of x
10. s_x \leftarrow compress s_x data
11. apply hyperparameter tuning for each classifier
12. classifier \leftarrow train_model using s_x
13. model \leftarrow apply_voting_classifier using classifier
14. predict \leftarrow cross_validation with model
15. computer performance evaluation metrics

End

TABLE 4. Wisconsin breast cancer dataset features.

No	Features	Details
1	<i>Radius_mean</i>	Mean distance from the centre perimeter points of the cell
2	<i>texture_mean</i>	Standard deviation of gray-scale
3	<i>perimeter_mean</i>	Cell nuclei parameters
4	<i>area_mean</i>	The total area of cell nuclei
5	<i>smoothness_mean</i>	Variation in radius length
6	<i>concavity_mean</i>	Variation in radius length
7	<i>concave_point_mean</i>	$\text{Perimeter}^2/\text{area} - 1$
8	<i>symmetry_mean</i>	Symmetry of the cell nuclei
9	<i>fractal_dimension_mean</i>	Approximation coastline-1

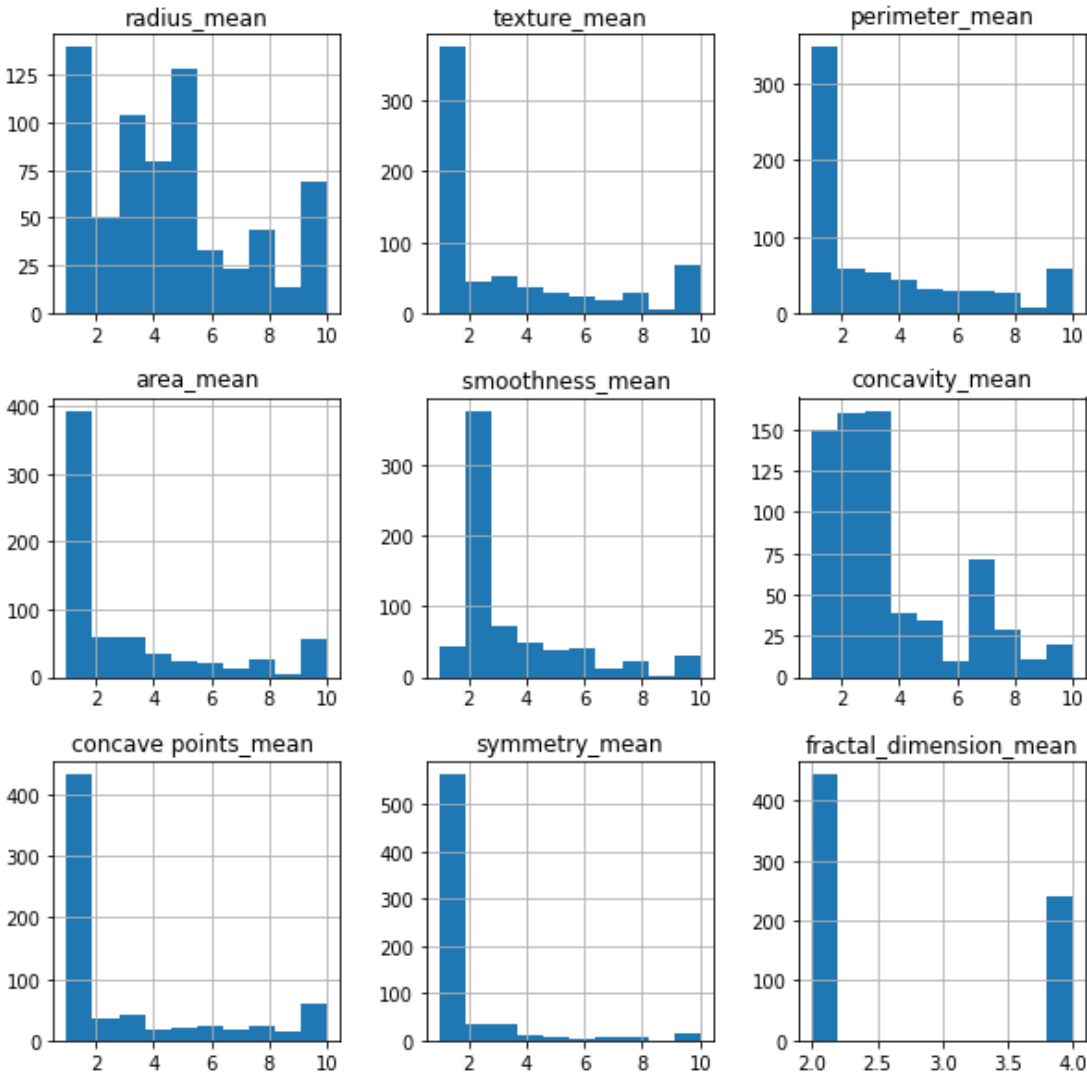


FIGURE 3. Visualizing the Distribution Histogram of Mean Features in the Dataset. This histogram provides insights into the distribution patterns of the dataset’s mean features, offering a comprehensive overview of the central tendencies and variations within the dataset.

1) EXTRA TREE)

Extra trees are the large number of decision trees generated from the training data. A split rule for the root is considered randomly from the root node features k subset, and a partially

random cut point [48]. The parent node is divided randomly into two selected child nodes. Each child node is repeated until the leaf node is reached. The majority votes determine final predictions. The user selects the top k features used in

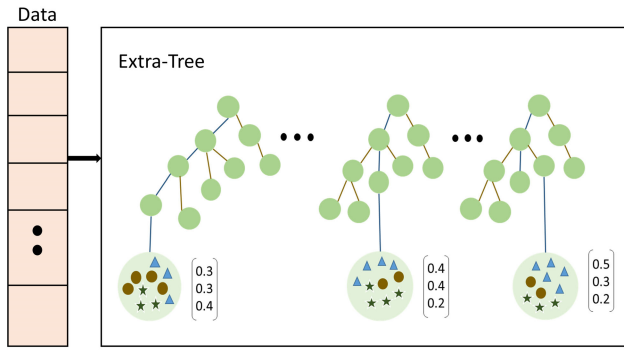


FIGURE 4. Exploring classification with extremely randomized trees.

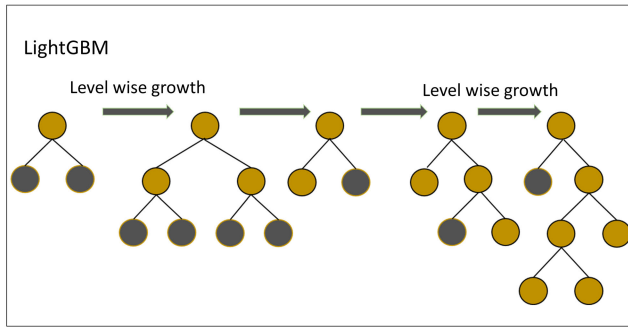


FIGURE 5. Illuminating predictive power: A closer examination of the light gradient boosting method.

the classification model as a final step. Extra tree predicts the decision in cases of regression or classification shown in Figure 4.

- Regression: Averaged predictions based on decision trees.
- Classification: Tree-based predictions based on majority voting.

2) LIGHT GRADIENT BOOSTING MACHINE (LIGHTGBM)

LightGBM is a gradient-boosting algorithm based on decision trees. A regression analysis was used to classify data rank. In training and separating the data from each decision tree, two strategies can be used: one that focuses on the level of the tree and the other that focuses on the tree's leaves. A level-wise approach grows the tree while maintaining its balance, whereas a leaf-wise method keeps splitting the leaves and reduces the loss, as shown in Figure 5. The leaf-wise growing tree structure of LightGBM selects and splits losses in a specific branch based on their contribution to the overall loss. A growing tree-based model with a low error rate typically learns more quickly [49]. The mainly horizontal growth of the LightGBM model prevents over-learning. As a result, large datasets produce better results [50].

3) RIDGE CLASSIFIER (RC)

Ridge classification is a machine-learning technique for analyzing linear discriminant models. It is a type of

regularization in which model coefficients are penalized to prevent over-fitting. This classifier converts the target values to -1 and $+1$ before treating the problem as a regression in training data.

4) LINEAR DISCRIMINANT ANALYSIS (LDA)

Linear discriminant analysis is used for classification, and dimensional reduction is used for supervised classification problems. The primary purpose of LDA is to maximize between-class variance and minimize within-the-class variance through linear discriminant function. In other words, it is based on the search for variables in a linear combination to ensure the best distinguishing characteristics for multi-class labels [51].

$$z = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_d x_d \quad (1)$$

$$s(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T c \beta} \quad (2)$$

$$S(\beta) = \frac{z_1 - z_2}{z(\text{variance in the group})} \quad (3)$$

$$\beta = C^{-1} (\mu_1 - \mu_2) \quad (4)$$

$$C = \frac{1}{(N_1 + N_2)} (N_1 C_1 + N_2 C_2) \quad (5)$$

The following equations are estimated as the linear coefficients and maximize the discriminant function score. In the equations, the c represents the linear model of the coefficient, the β , the covariance matrix of the function, and the μ shows the average vector of the function.

5) VOTING CLASSIFIER

Voting classifier is a machine-learning model that trains an ensemble of various models. The finding of each classifier passed into the voting classifier and predicted the output class based on the highest voting majority. Voting ensemble techniques are used in ensemble machine learning models to combine predictions from multiple models [52]. In our research, we applied the hard voting method, which identifies the class with the highest votes based on the combined predictions of each classifier, as shown in 6. Voting ensemble classifiers are used in the context of breast cancer classification to improve the accuracy and robustness of the classification. In some breast cancer datasets, one class may have many more instances than another. This can make it difficult for a single classifier to predict both classes accurately. By combining the predictions of multiple classifiers, the voting ensemble classifier can provide a more balanced and accurate prediction. In this study, the voting ensemble model uses four base classifiers. The Extra Tree (ET) is employed as a meta-classifier. Basic classifiers were initially trained on the base model's whole training input data set. The meta-model classifier takes the prediction from each base model as its input. The adaptive voting ensemble classifier can improve outliers and noisy data robustness. This is because multiple classifiers are trained on the same dataset,

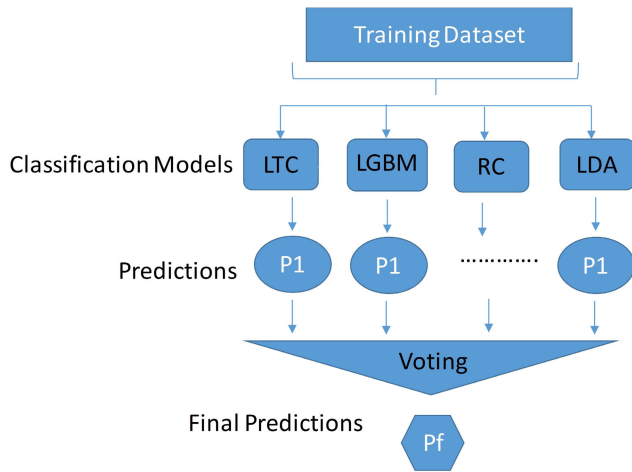


FIGURE 6. A voting-based ensemble classifier is compared to the performance of multiple classifiers combined into one model.

and their predictions are combined in a way that minimizes the impact of outliers and noisy data [53]. Moreover, adaptive voting ensemble classifiers in breast cancer classification can result in improved accuracy, robustness, and balance in the predictions, making it a useful tool in the analysis and diagnosis of breast cancer.

Algorithm 2 Pseudocode of the Voting Ensemble Learning Algorithm

1. Input: Input Breast cancer training data
2. Base level classifiers = (ET, LGBM, RC, LDA)
 - a. Meta Level Classifier ET
3. Output: Trained ensemble classifier
4. Step 1: Train base learner by applying classifiers to dataset
 - a. For training set of classifiers, use cross validation(k-fold)
 - b. for $k = 1$ to k_{max} , do; where $k_{max} = 10$
 - c. $\beta = ()$
5. Step 2: learn a classifier from
6. end for
7. Step 3: Training set for the meta-level classifier ET;
 - a. Train meta classifier ET;
 8. $\beta = ()$
9. Return
10. END

6) ENVIRONMENT SETUP

To carry out our research accurately and efficiently, We created specific environments for completing this research work. We have provided a detailed presentation of our environment setup in Table 5, which includes all the intricate details. This approach helped us conduct a thorough exploration and analysis during the research process, enhancing our findings' reliability and validity.

TABLE 5. Configuration of the proposed model system's environment involves the setup process.

Resource	Details
CPU	Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz 3.19 GHz
RAM	32.0 GB
GPU	NVIDIA GeForce GTX 1060 6Gb
Software	Jupyter Notebook
Language	Python
System	64-bit operating system, x64-based processor

V. RESULT AND DISCUSSION

A. PERFORMANCE METRICS

The confusion matrix is the best method for evaluating a classification model. Observations that the model correctly predicts are True positive and True negative, while False positive and False negative are minimized [54].

Accuracy

In training models, accuracy represents the degree of correctness. In other words, it is the ratio of correct predictions to all predictions.

$$\text{Accuracy(success rate)} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

Recall

A false negative is a difference between True Positives and False Negatives. The recall equation is shown below:

$$\frac{T_P}{T_P + F_N}$$

Precision

Precision measures accuracy in determining the proportion of True Positives to all positive predictions. Here is the precision equation below:

$$\frac{T_P}{T_P + F_P}$$

F1 Measure

In terms of precision and recall, F1 is a harmonic mean. Hence, it considers false positives and false negatives. It is often more helpful than accuracy in cases where class distributions are uneven.

$$F_1 = 2 * \frac{\text{Precision}(P) * \text{Recall}(R)}{\text{Precision}(P) + \text{Recall}(R)}$$

AUC It provides an overall performance measure across all classification criteria. In other words, ROC/AUC measures a classifier's ability to distinguish between classes.

Additionally, the true positive rate $\left(TPR = \frac{T_P}{P}\right)$, true negative rate $\left(TNR = \frac{T_N}{N}\right)$, false positive rate $\left(FPR = \frac{F_P}{F_P + T_N}\right)$, and false negative rate $\left(FNR = \frac{F_N}{F_N + T_P}\right)$ are used to examine the proposed approach.

B. EXPERIMENTAL RESULTS

The experiment conducted for breast cancer diagnosis involves the use of the Wisconsin Breast Cancer dataset (WBCD), which is split into two subsets: 70% of the data

is allocated for training. In contrast, the remaining 30% is reserved for testing. The proposed classification model undergoes evaluation based on performance metrics such as accuracy, f-score, recall, and precision. The emphasis is on predicting optimal features for effective breast cancer detection.

A confusion matrix is employed to assess the accuracy of the classification model and identify potential issues. This matrix is beneficial when dealing with datasets with uneven class distributions, preventing misleading interpretations of classification accuracy. The evaluation involves analyzing Figure 7, which shows four confusion matrices for different machine learning classifiers: Extra Trees Classifier, LGBM Classifier, Ridge Classifier, and Linear Discriminant Analysis. A confusion matrix is a table often used to describe the performance of a classification model on a set of test data for which the actual values are known. Each matrix has two rows and two columns, representing the counts of true negatives, false positives, false negatives, and true positives. These counts are used to calculate performance metrics such as accuracy, precision, recall, and F1 score. The matrices suggest a binary classification problem with two classes (0 and 1). For instance, the Extra Trees Classifier correctly predicted 60 instances of class 0 (true negatives) and 106 instances of class 1 (true positives) while incorrectly predicting 3 instances as class 1 (false positives) and 2 instances as class 0 (false negatives).

Notably, the proposed model correctly classifies 106 benign breast cancer samples, contributing significantly to overall accuracy. Moreover, compared to other models, it exhibits fewer errors, highlighting its effectiveness in improving the breast cancer detection process.

Classification models are used to predict the best feature. This model reduces the number of features and can handle extensive data for a more accurate prediction of breast cancer. The evaluation analysis indicates that in Tab 6, the proposed approach ELRL-E achieved 97.6% testing accuracy, 96.46% precision, 100% recall, and 98.1% F1 score, which indicates that the proposed approach was significantly better and outperformed existing ML and ensemble models

the evaluation results Of the proposed approach with baseline learning models, specifically Extra Trees (ET), LightGBM, Ridge Classifier (RC), and Linear Discriminant Analysis (LDA). The analysis provides detailed metrics for the ET and LightGBM classifiers, showcasing their accuracy and F1 scores.

The Extra Trees (ET) classifier achieved an accuracy of 96.49%, indicating the percentage of correctly classified instances and an F1 score of 97.24%. The F1 score is a metric that balances precision and recall, providing a comprehensive measure of a model's performance.

Similarly, the LightGBM classifier demonstrated an accuracy of 95.99% and an F1 score of 96.86%. These metrics collectively convey the model's effectiveness in accurately classifying instances and balancing precision and recall.

To visually represent the comparison of the proposed approach with these baseline models, Figure 8 is provided. This figure likely depicts a graphical representation, such as a bar chart or line graph, illustrating the overall accuracy of each model. The analysis in this figure allows for a quick and intuitive comparison of the performance of the proposed approach against the baseline learners.

In evaluating model performance on an imbalanced dataset, ROC curves were employed as specific metrics due to their effectiveness in assessing the ability to detect false positives and negatives. The ROC curve is particularly well-suited for such evaluations. Figure 9, illustrates the ROC curves and confusion matrix for the proposed ensemble categorization and an additional ensemble model. Confusion Matrix and ROC (Receiver Operating Characteristic) graph for ensemble classifiers. The Confusion Matrix illustrates the performance of these classifiers by depicting the counts of true positive, true negative, false positive, and false negative instances, offering a detailed breakdown of their classification accuracy. Simultaneously, the ROC graph provides a graphical portrayal of the classifiers' ability to discriminate between different classes, offering insights into their overall performance and trade-offs between sensitivity and specificity across various classification thresholds. These visualizations comprehensively assess the ensemble classifiers' effectiveness in handling classification tasks. Notably, our results indicate that the proposed model achieved the highest Area Under the Curve (AUC) value, reaching a perfect score of 1.00.

C. DISCUSSION

In recent years, many researchers have explored different techniques and methodologies to analyze breast cancer. Based on our comparative analysis, demonstrated in Table 7, we propose a better method than previous research on the same WBCD dataset, where we employ a sophisticated voting ensemble classifier, termed ELRL-E, comprising four integrated machine learning models: Extra Trees Classifier (ETC), Light Gradient Boosting Machine (LightGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA). Our results demonstrate the promising performance of the ELRL-E approach, achieving an accuracy of 97.6%, precision of 96.4%, recall of 100%, and an F1 score of 98.1%. These metrics surpass the performance of previously employed machine learning and ensemble models. Our methodology excels in feature optimization and the strategic use of relevant features, addressing a critical aspect often overlooked in prior studies. Compared to well-known classifiers, such as k-NN, NB, and SVM, evaluated by Acquisition et al. (2019) using Weka, our approach circumvents challenges in cross-language implementation and integrates seamlessly into the existing architecture. Furthermore, we contribute to the discourse on ethical implications and reliability in healthcare applications, as emphasized by Commission et al. (2019). Our work aims for accuracy and

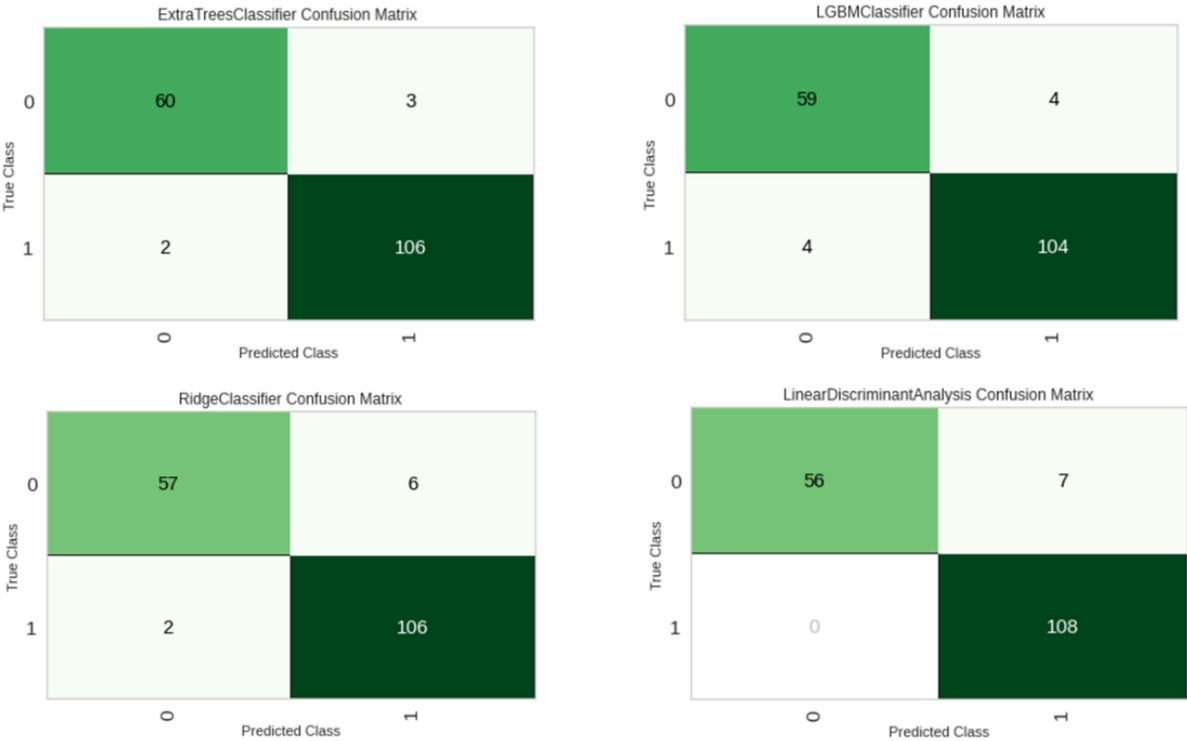


FIGURE 7. Exploring the performance through a detailed Confusion Matrix Comparison for four distinct classifiers: Extra Trees (ET), LightGBM (LGBM), Ridge Classifier (RC), and Linear Discriminant Analysis (LDA).

TABLE 6. Comparison evaluation performance of proposed and baseline ML models.

Models	Accuracy	Sensitivity	Specificity	F1score
ET	96.49%	97.98%	96.62%	97.24%
LightGBM	95.99% %	97.98%	95.88%	96.86%
RC	95.72%	9888%	94.82%	96.70%
LDA	95.48%	99.29%	94.07%	96.53%
Our ELRL-E Model	97.66%	100%	96.43%	98.18%

TABLE 7. Comparison of baseline ML models with existing predicted models.

Year	Model	Dataset	Instances	Features	Accuracy
[55]	KNN	UCI	699	10	96.85%
[56]	SVM	WBCD	699	10	96.72%
[57]	KNN	UCI	569	32	94.35%
[58]	BN, RBF	WBCD	699	11	97.42%
[59]	DT	WBCD	569	31	92.53%
[60]	t-SNE	WBCD	699	11	86.6%
[61]	XGboost	WBCD	569	30	96%
Proposed Model	ELRL-E	WBCD	569	9	97.6%

a comprehensive evaluation of the ensemble model’s efficacy in the critical context of breast cancer detection and diagnosis. While Assegie et al. [57] highlighted the significance of parameter tuning in a K-Nearest Neighbor (KNN) model, our approach builds on this foundation by integrating multiple classifiers to enhance performance. Jabbar et al. [58] achieved a remarkable accuracy of 97%, and we acknowledge their contribution. Still, our study goes beyond by providing a robust comparative analysis, showcasing the strengths of the

ELRL-E approach against existing state-of-the-art classifiers. Sharma et al. [60] utilized t-SNE and snapshot ensembling, acknowledging potential limitations. Sara et al. (2023) This paper introduces a machine learning CAD system for breast cancer classification, leveraging feature selection, PCA, and seven ML classifiers. The XGboost model achieved high recall for the Mammographic Mass dataset, while AdaBoost with S-LR excelled for the WBCD dataset. The stacking with the logistic regression ensemble model demonstrated

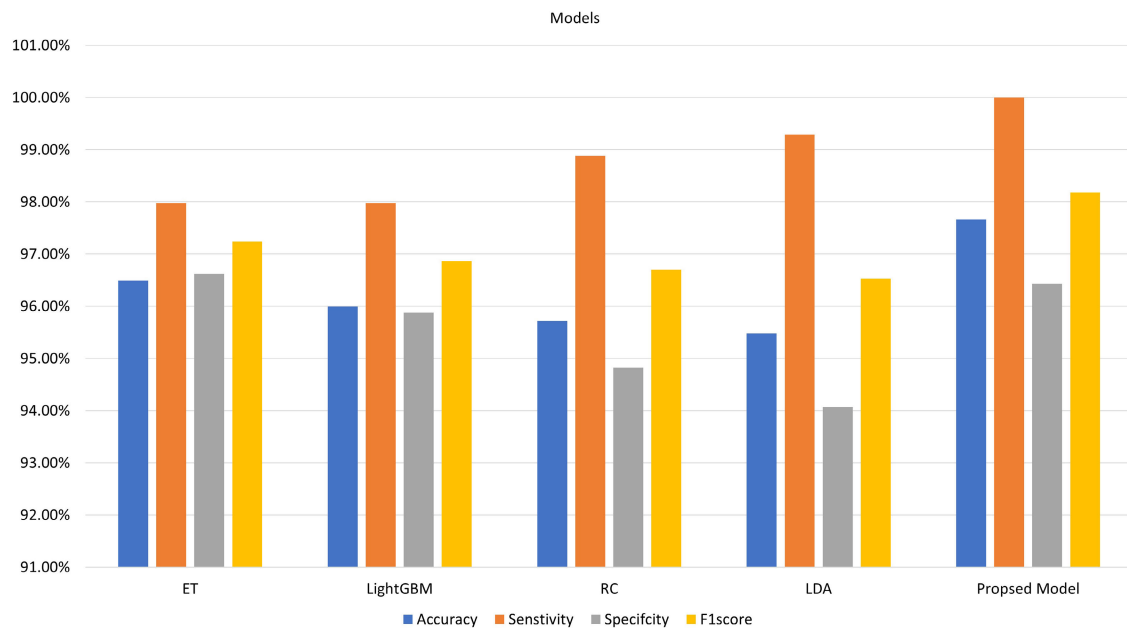


FIGURE 8. A comprehensive examination of the evaluation scores for the proposed approach is conducted in contrast to those of established baseline learners. This analysis delves into the numerical metrics and performance indicators of the proposed model, scrutinizing how it compares to the baseline models—namely, Extra Trees (ET), LightGBM, Ridge Classifier (RC), and Linear Discriminant Analysis (LDA).

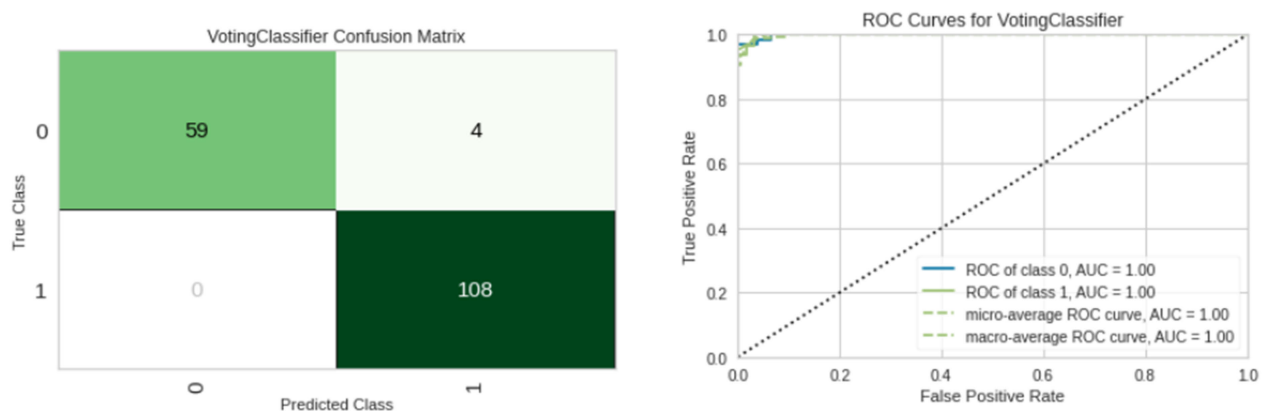


FIGURE 9. This evaluation incorporates a Confusion Matrix to assess the classification accuracy, along with an ROC graph, providing a visual representation of the ensemble classifiers' ability to discriminate between classes.

the highest accuracies. However, limitations include dataset-specificity, potential challenges in clinical implementation, and the simplification of complex decision-making processes in the ML application. The comprehensive evaluation, strategic feature selection, and integration of advanced classifiers in ELRL-E substantiate its superiority, addressing limitations and significantly advancing breast cancer classification.

VI. CONCLUSION AND FUTURE WORK

Breast cancer is one of the leading causes of death in women; thus, early identification is critical. Implementing robust machine learning classifiers can improve early breast cancer tumor identification. Predictive performance enhancement

depends on a range of model factors. Ensemble learning generally outperforms a single-base classifier because it combines several independent learning algorithms. Consequently, it has gained popularity and proven an effective machine-learning method. One of the most significant issues is finding a way to combine the most accurate base classifiers. To solve these issues, we propose applying a unique model known as the ELRL-E model. To select the most practical combination of base classifiers, ELRL-E uses various Machine Learning algorithms, including ET, LightGBM, RC, and LDA, to classify breast cancer tumors accurately. In addition, we used a voting classifier to analyze the significance and effectiveness of the proposed ELRL

E model. The experiment results show that the proposed approach ELRL-E achieves the highest accuracy of 97.66%, a precision of 96.43%, and a recall of 100%—F1 score of 98.18% compared to the other implemented ensemble models. Furthermore, the experiment results indicate that the proposed ELRL improved the accuracy compared to the ET, LightGBM, RC, and LDA models. Combining models can increase diagnosis quality and provide a significant advantage over previous work.

Moving forward, our future research aims to assess the applicability of the proposed model on diverse disease datasets for comprehensive validation. Acknowledging current limitations, such as evaluating a relatively small dataset, it is crucial to extend validation efforts to significantly more extensive datasets. Moreover, we aim to improve the model's performance by refining hyperparameters and exploring optimization algorithms considering hyperparameters like learning rate, tree depth, and regularization, addressing challenges in tuning for robustness and scalability on larger datasets.

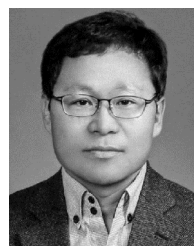
REFERENCES

- [1] *Breast Cancer Statistics*, NBCC, 2022.
- [2] B. He, H. Sun, M. Bao, H. Li, J. He, G. Tian, and B. Wang, "A cross-cohort computational framework to trace tumor tissue-of-origin based on RNA sequencing," *Sci. Rep.*, vol. 13, no. 1, p. 15356, Sep. 2023.
- [3] W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, "Semi-supervised vision transformer with adaptive token sampling for breast cancer classification," *Frontiers Pharmacol.*, vol. 13, Jul. 2022, Art. no. 929755.
- [4] S. Chen, Y. Chen, L. Yu, and X. Hu, "Overexpression of SOCS4 inhibits proliferation and migration of cervical cancer cells by regulating JAK1/STAT3 signaling pathway," *Eur. J. Gynaecol. Oncol.*, vol. 42, no. 3, pp. 554–560, 2021.
- [5] I. A. for Research on Cancer, W. H. Organization, *Globocan 2012: Cervical Cancer-Estimated Incidence, Mortality and Prevalence Worldwide in 2012*, 2018.
- [6] Z.-R. Jiang, L.-H. Yang, L.-Z. Jin, L.-M. Yi, P.-P. Bing, J. Zhou, and J.-S. Yang, "Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients," *Frontiers Oncol.*, vol. 12, Sep. 2022, Art. no. 988680.
- [7] N. Kumar Sinha, "Developing a web based system for breast cancer prediction using XGboost classifier," *Int. J. Eng. Res.*, vol. V9, no. 6, pp. 852–856, Jun. 2020.
- [8] M. Umer, M. Naveed, F. Alrowais, A. Ishaq, A. A. Hejaili, S. Alsubai, A. A. Eshamawi, A. Mohamed, and I. Ashraf, "Breast cancer detection using convoluted features and ensemble machine learning algorithm," *Cancers*, vol. 14, no. 23, p. 6015, Dec. 2022.
- [9] B. He, Y. Zhang, Z. Zhou, B. Wang, Y. Liang, J. Lang, H. Lin, P. Bing, L. Yu, D. Sun, H. Luo, J. Yang, and G. Tian, "A neural network framework for predicting the tissue-of-origin of 15 common cancer types based on RNA-Seq data," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 737, Aug. 2020.
- [10] S. A. Yazdan, R. Ahmad, N. Iqbal, A. Rizwan, A. N. Khan, and D.-H. Kim, "An efficient multi-scale convolutional neural network based multi-class brain MRI classification for SaMD," *Tomography*, vol. 8, no. 4, pp. 1905–1927, Jul. 2022.
- [11] Imran, N. Iqbal, S. Ahmad, and D. H. Kim, "Health monitoring system for elderly patients using intelligent task mapping mechanism in closed loop healthcare environment," *Symmetry*, vol. 13, no. 2, p. 357, Feb. 2021.
- [12] J. Zhang, Q. Shen, Y. Ma, L. Liu, W. Jia, L. Chen, and J. Xie, "Calcium homeostasis in Parkinson's disease: From pathology to treatment," *Neurosci. Bull.*, vol. 38, no. 10, pp. 1267–1270, Oct. 2022.
- [13] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, Sep. 2023, Art. no. 100804.
- [14] Y. Li, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl. Comput. Math.*, vol. 7, no. 4, pp. 212–216, 2018.
- [15] B. He, C. Dai, J. Lang, P. Bing, G. Tian, B. Wang, and J. Yang, "A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation," *Biochim. et Biophys. Acta (BBA)-Mol. Basis of Disease*, vol. 1866, no. 11, 2020, Art. no. 165916.
- [16] A. Bharat, N. Pooja, and R. A. Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in *Proc. 3rd Int. Conf. Circuits, Control, Commun. Comput. (I4C)*, Oct. 2018, pp. 1–4.
- [17] K. Jain and M. S. Sharma, "Breast cancer diagnosis using machine learning techniques," *Int. J. Innov. Sci., Eng. Technol.*, vol. 5, no. 5, 2018.
- [18] A. S. Elkorany, M. Marey, K. M. Almustafa, and Z. F. Elsharkawy, "Breast cancer diagnosis using support vector machines optimized by whale optimization and dragonfly algorithms," *IEEE Access*, vol. 10, pp. 69688–69699, 2022.
- [19] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *Proc. Int. Conf. Robots Intell. Syst. (ICRIS)*, May 2018, pp. 157–160.
- [20] V. Amudha, R. G. Babu, K. Arunkumar, and A. Karunakaran, "Machine learning-based performance comparison of breast cancer detection using support vector machine," in *Proc. AIP Conf.*, vol. 2519, no. 1, 2022, Art. no. 050011.
- [21] E. Halim, P. P. Halim, and M. Hebrard, "Artificial intelligent models for breast cancer early detection," in *Proc. Int. Conf. Inf. Manage. Technol. (ICIMTech)*, Sep. 2018, pp. 517–521.
- [22] B. S. Abunasser, M. R. J. Al-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Breast cancer detection and classification using deep learning Xception algorithm," *Breast Cancer*, vol. 13, no. 7, 2022.
- [23] D. Albashish, "Ensemble of adapted convolutional neural networks (CNN) methods for classifying colon histopathological images," *PeerJ Comput. Sci.*, vol. 8, p. e1031, Jul. 2022.
- [24] M. Dabass, S. Vashisth, and R. Vig, "A convolution neural network with multi-level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105680.
- [25] P. Goyal, S. Sehgal, S. Ghosh, D. Aggarwal, P. Shukla, A. Kumar, R. Gupta, and S. Singh, "Histopathological correlation of atypical (C3) and suspicious (C4) categories in fine needle aspiration cytology of the breast," *Int. J. Breast Cancer*, vol. 2013, Sep. 2013, Art. no. 965498.
- [26] R. Mitchell and E. Frank, "Accelerating the XGBoost algorithm using GPU computing," *PeerJ Comput. Sci.*, vol. 3, p. e127, Jul. 2017.
- [27] D. Sharma, P. Jain, and D. K. Choubey, "A comparative study of computational intelligence for identification of breast cancer," in *Proc. Int. Conf. Mach. Learn., Image Process., Netw. Secur. Data Sci.* Springer, 2020, pp. 209–216.
- [28] D. K. Choubey, P. Kumar, S. Tripathi, and S. Kumar, "Performance evaluation of classification methods with PCA and PSO for diabetes," *Netw. Model. Anal. Health Informat. Bioinf.*, vol. 9, no. 1, pp. 1–30, Dec. 2020.
- [29] D. K. Choubey, S. Tripathi, P. Kumar, V. Shukla, and V. K. Dhandhanian, "Classification of diabetes by kernel based SVM with PSO," *Recent Adv. Comput. Sci. Commun.*, vol. 14, no. 4, pp. 1242–1255, Jul. 2021.
- [30] D. K. Choubey, M. Kumar, V. Shukla, S. Tripathi, and V. K. Dhandhanian, "Comparative analysis of classification methods with PCA and LDA for diabetes," *Current Diabetes Rev.*, vol. 16, no. 8, pp. 833–850, Sep. 2020.
- [31] M. O. Adebiyi, M. O. Arowolo, M. D. Mshelia, and O. O. Olugbara, "A linear discriminant analysis and classification model for breast cancer diagnosis," *Appl. Sci.*, vol. 12, no. 22, p. 11455, Nov. 2022.
- [32] O. J. Egwom, M. Hassan, J. J. Tanimu, M. Hamada, and O. M. Ogar, "An LDA–SVM machine learning model for breast cancer classification," *BioMedInformatics*, vol. 2, no. 3, pp. 345–358, Jun. 2022.
- [33] K. Bala, D. K. Choubey, and S. Paul, "Soft computing and data mining techniques for thunderstorms and lightning prediction: A survey," in *Proc. Int. Conf. Electron., Commun. Aeronaut. Technol. (ICECA)*, vol. 1, Apr. 2017, pp. 42–46.
- [34] K. Bala, D. K. Choubey, S. Paul, and M. G. N. Lala, "Classification techniques for thunderstorms and lightning prediction: A survey," in *Soft-Computing-Based Nonlinear Control Systems Design*. Hershey, PA, USA: IGI Global, 2018, pp. 1–17.
- [35] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: A comparative study using machine learning techniques," *Social Netw. Comput. Sci.*, vol. 1, no. 5, pp. 1–14, Sep. 2020.

- [36] A. Bhardwaj, H. Bhardwaj, A. Sakalle, Z. Uddin, M. Sakalle, and W. Ibrahim, "Tree-based and machine learning algorithm analysis for breast cancer classification," *Comput. Intell. Neurosci.*, vol. 2022, Jul. 2022, Art. no. 6715406.
- [37] M. R. Basunia, I. A. Pervin, M. Al Mahmud, S. Saha, and M. Arifuzzaman, "On predicting and analyzing breast cancer using data mining approach," in *Proc. IEEE Region 10 Symp. (TENSYP)*, Jun. 2020, pp. 1257–1260.
- [38] G. I. Salama, M. Abdelhalim, and M. A. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," *Breast Cancer (WDBC)*, vol. 32, no. 569, p. 2, 2012.
- [39] V. Birchha and B. Nigam, "Performance analysis of averaged perceptron machine learning classifier for breast cancer detection," *Proc. Comput. Sci.*, vol. 218, pp. 2181–2190, 2023.
- [40] V. A. M. De Barros, H. M. Paiva, and V. T. Hayashi, "Using PBL and agile to teach artificial intelligence to undergraduate computing students," *IEEE Access*, vol. 11, pp. 77737–77749, 2023.
- [41] V. D. P. Jasti, A. S. Zamani, K. Arumugam, M. Naved, H. Pallathadka, F. Sammy, A. Raghuvanshi, and K. Kaliyaperumal, "Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis," *Secur. Commun. Netw.*, vol. 2022, Mar. 2022, Art. no. 1918379.
- [42] F. Budiman, I. A. Saputro, P. Purwanto, and P. N. Andono, "Optimization of classification results by minimizing class imbalance on decision tree algorithm," in *Proc. Int. Seminar Mach. Learn., Optim., Data Sci. (ISMODE)*, Jan. 2022, pp. 6–11.
- [43] N. Mohd Ali, R. Besar, and N. A. A. Aziz, "A case study of microarray breast cancer classification using machine learning algorithms with grid search cross validation," *Bull. Electr. Eng. Informat.*, vol. 12, no. 2, pp. 1047–1054, Apr. 2023.
- [44] F. Atban, E. Ekinci, and Z. Garip, "Traditional machine learning algorithms for breast cancer image classification with optimized deep features," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104534.
- [45] S. W. Wolberg and M. O. William, "Breast cancer Wisconsin (diagnostic)," UCI Machine Learning Repository, Tech. Rep., 1995.
- [46] L. Dora, S. Agrawal, R. Panda, and A. Abraham, "Optimal breast cancer classification using Gauss–Newton representation based algorithm," *Expert Syst. Appl.*, vol. 85, pp. 134–145, Nov. 2017.
- [47] S. Sharma and S. Deshpande, "Breast cancer classification using machine learning algorithms," in *Proc. Mach. Learn. Predictive Anal. (ICTIS)*. Springer, 2021, pp. 571–578.
- [48] G. Alfian, M. Syafrudin, I. Fahrurrozi, N. L. Fitriyani, F. T. D. Atmaji, T. Widodo, N. Bahiyah, F. Benes, and J. Rhee, "Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method," *Computers*, vol. 11, no. 9, p. 136, Sep. 2022.
- [49] Q. Chen, Z. Meng, and R. Su, "WERFE: A gene selection algorithm based on recursive feature elimination and ensemble strategy," *Frontiers Bioeng. Biotechnol.*, vol. 8, p. 496, May 2020.
- [50] S. Akbulut, I. Balıkcı Cicek, and C. Colak, "Classification of breast cancer on the strength of potential risk factors with boosting models: A public health informatics application," *Med. Bull. Haseki*, vol. 60, no. 3, pp. 196–203, Jun. 2022.
- [51] H. Mandelkow, J. A. de Zwart, and J. H. Duyn, "Linear discriminant analysis achieves high classification accuracy for the BOLD fMRI response to naturalistic movie stimuli," *Frontiers Hum. Neurosci.*, vol. 10, p. 128, Mar. 2016.
- [52] H. Jamil, F. Qayyum, N. Iqbal, F. Jamil, and D. H. Kim, "Optimal ensemble scheme for human activity recognition and floor detection based on AutoML and weighted soft voting using smartphone sensors," *IEEE Sensors J.*, vol. 23, no. 3, pp. 2878–2890, Feb. 2023.
- [53] S. Chatterjee and Y.-C. Byun, "Voting ensemble approach for enhancing Alzheimer's disease classification," *Sensors*, vol. 22, no. 19, p. 7661, Oct. 2022.
- [54] A. Batool and Y.-C. Byun, "An ensemble architecture based on deep learning model for click fraud detection in Pay-Per-Click advertisement campaign," *IEEE Access*, vol. 10, pp. 113410–113426, 2022.
- [55] B. Akbugday, "Classification of breast cancer data using machine learning algorithms," in *Proc. Med. Technol. Congr. (TIPTKNO)*, Oct. 2019, pp. 1–4.
- [56] Md. I. H. Showrov, M. T. Islam, Md. D. Hossain, and Md. S. Ahmed, "Performance comparison of three classifiers for the classification of breast cancer dataset," in *Proc. 4th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Dec. 2019, pp. 1–5.
- [57] T. A. Assegie, "An optimized K-nearest neighbor based breast cancer detection," *J. Robot. Control (JRC)*, vol. 2, no. 3, pp. 115–118, 2021.
- [58] M. A. Jabbar, "Breast cancer data classification using ensemble machine learning," *Eng. Appl. Sci. Res.*, vol. 48, no. 1, pp. 65–72, 2021.
- [59] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, p. 184, 2021.
- [60] N. Sharma, K. P. Sharma, M. Mangla, and R. Rani, "Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding," *Multimedia Tools Appl.*, vol. 82, no. 3, pp. 4011–4029, Jan. 2023.
- [61] "An improved breast cancer disease prediction system using ml and pca multimedia tools and applications," 2023.



AMREEN BATOOL received the bachelor's degree from GC University, Pakistan, the M.C.S. degree from the Virtual University of Pakistan, and the M.S. degree in computer science and technology from Tiangong University, Tianjin, China, in 2021. She is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Jeju National University, Republic of Korea. She is also a Project Coordinator with EUT Global Ltd. Her primary role is coordinating with clients and field engineers to plan project delivery. Her research interests include machine learning, deep learning, and blockchain technology.



YUNG-CHEOL BYUN received the B.S. degree from Jeju National University, in 1993, and the M.S. and Ph.D. degrees from Yonsei University, in 1995 and 2001, respectively. He was a Special Lecturer with SAMSUNG Electronics, in 2000 and 2001. From 2001 to 2003, he was a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI). He was promoted to join Jeju National University as an Assistant Professor, in 2003, where he is currently an Associate Professor with the Computer Engineering Department. His research interests include the areas of AI machine learning, pattern recognition, blockchain and deep learning-based applications, big data and knowledge discovery, time series data analysis and prediction, image processing and medical applications, and recommendation systems.

• • •