Healthcare Analytics 6 (2024) 100353

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction

Md. Shazzad Hossain Shaon, Tasmin Karim, Md. Shahriar Shakil, Md. Zahid Hasan *

Health Informatics Research Lab, Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Birulia, Dhaka, 1216, Bangladesh

ARTICLE INFO

Handling editor: Madijd Tavana

Keywords:
Breast cancer
Machine learning
Diagnostic tools
Feature selection
Meta-models
SHapley additive exPlanations (SHAP)

ABSTRACT

In recent decades, breast cancer has become the most prevalent type of cancer that impacts women in the world, which shows a significant risk to the death rates of women. Early identification of breast cancer might drastically decrease patient mortality and greatly improve the chance of an effective treatment. In modern times, machine learning models have become crucial for classifying cancer and strengthening both the accuracy and efficiency of diagnostic and medical treatment strategies. Therefore, this study is focused on early detection of breast cancer using a variety of machine learning algorithms and desires to identify the most effective feature selection process with an amalgamated dataset. Initially, we evaluated five traditional models and two meta-models on separate datasets. To find the most valuable features, the study used the Least Absolute Shrinkage and Selection Operator (LASSO) as well as SHapley Additive exPlanations (SHAP) selection methods and analyzed them through a wide range of performance regulations. Additionally, we applied these models to the combined dataset and observed that the mergeddataset was significantly beneficial for breast cancer diagnosis. After analyzing the feature selection strategies, it was demonstrated that the majority of models performed more accurately when utilizing SHAP methodologies. Notably, three traditional models and two meta-classifiers obtained an accuracy of 99.82%, demonstrating superior performance compared to state-of-the-art methods. This advancement holds a crucial role as it lays the foundation for refining diagnostic tools and enhancing the progression of medical science in this field.

1. Introduction

Breast cancer occurs due to uncontrolled expansion within the cell cycle in the breast tissues. Globally, this cancer is the second most prevalent malignancy among women [1]. The World Health Organization (WHO) states that more than 2 million women have been diagnosed with breast cancer and 685000 women have been deceased in 2020 [2]. According to the World Cancer Research Fund International (WCRFI), more than eight nations have experienced an alarming increase in cancer mortality and diagnoses, which is calculated by the Age-Standardized Rates (ASR). Among them, Belgium has the highest prevalence of this cancer with 113.2 ASR stating 11,737 women being affected. The Netherlands has the second highest prevalence with 100.09 ASR and 15,725 women have been affected by this tumor. On the other side, the maximum death rate was discovered in Barbados, which had 42.2 ASR with 111 women deceased. Following that, 184 women died in Fiji with a 41.0 ASR in 2020 [3]. By 2030, this cancer rate is anticipated to rise by 27 million people [4]. Breast cancer can be identified by categorizing cells. As demonstrated in breast cancer occurrences, there are different kinds of tumors: both malignant and benign. Malignant tumors spread around the cells quickly posing more of a great threat than benign ones. There are nevertheless instances of both rapidly expanding benign cancers and slowly developing malignant tumors. This disease proportion has been increasing continuously for many reasons. Most women are unaware of the dangers of cancer, and economic disparity, insufficient detection, and lack of integrity in healthcare are impacted by the result of cancer [5]. Therefore, in the modern-era, it is necessary to build an Autonomous application process for reliable cancer detection. Many specialists have suffered a lot for the accurate detection of using outdated technology. Machine learning is a tremendous platform for diagnosing breast cancer, and it would be efficient to build an artificial system based on machine learning

* Corresponding author.
E-mail addresses: shazzad15-3404@diu.edu.bd (Md.S.H. Shaon), tasmin15-2920@diu.edu.bd (T. Karim), shakil.cse@diu.edu.bd (Md.S. Shakil), zahid.cse@diu.edu.bd (Md.Z. Hasan).

https://doi.org/10.1016/j.health.2024.100353
Received 1 February 2023; Received in revised form 20 May 2024; Accepted 24 June 2024
Available online 25 June 2024

methods. Hence, the purpose of this study is to develop an enhanced approach to address the challenges associated with breast cancer prediction. The contributions of this study are defined as follows.

a) Integration of two datasets optimized for breast cancer diagnosis using LASSO and SHAP feature selection methods, which resulted in increasing models' performance.
b) The addition of meta-models such as HVC and SVC to standard frameworks for enhancing testing accuracy and reducing bias.
c) Our proposed method, especially when integrating the SHAP feature-based approach, demonstrated superior performance with 99.82% accuracy compared to existing models, showcasing accelerated diagnostic capabilities for breast cancer detection.

We believe that our methodology would be beneficial in the medical sector for detecting breast cancer. In the future, developers will be able to easily construct a robust web server utilizing feature selection techniques. This advancement is beneficial for the future of technology as it allows for the development of an advanced framework with significant features. The relevance derives from the ability to save lives with early and accurate diagnosis tools. The method excels at accurately discriminating between positive and negative classes, resulting in beneficial and life-saving outcomes for patients. In the following section, we have described several existing methods based on breast cancer disease.

2. Related work

Machine learning is the most successful way for identifying breast cancer, using several models to assess variables and identify the most significant aspects for reliable identification. Various models have been presented in previous studies that have utilized several features, where authors applied different algorithms and methodologies for detecting tumors. According to the study, the most significant challenge in breast cancer detection is the scarcity of massive data, lack of feature selections, and insufficient performances of the models. In this section, we identified relevant research on breast cancer detection based on computational approaches.

Assegie et al. [6] proposed adapting method (AB) for identification with 96.50% accuracy. However, the researchers used the Wisconsin Breast Cancer Dataset (WBCD1) with 30 features, and they did not mention the model's parameters and tuning processes. In another study, El Massari et al. [7] developed their model based on machine learning approach with 683 instances and obtained 96.90% accuracy. According to the study, they used a 50% split in the training dataset which provided less data for the models to learn during the training time and the authors did not mention the model's parameters, which will give ambiguity while building the framework in the future. Chaurasia et al. [8] proposed an LR model based on the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia's datasets which have 286 instances with an accuracy of 74.47 %. However, the authors used a smaller dataset for the detection, the model would be trained better with the larger dataset.

Ahmet Mert et al. [9] have evaluated ANN, KNN, RBFN, and SVM on the Wisconsin breast cancer dataset. The authors acquired 93.03 % with 10 cross-validations in KNN. The authors did not compare their model with state-of-the-art methods and the result is not sufficient to detect breast cancer more accurately. In another study, Abdullah-Al Nahid et al. [10] used SVM and CNN models, where they obtained a better accuracy with CNN-based feature extraction approach. [10]. R. Alam-Khan et al. [11] used cross-validation on the Wisconsin-Madison Prognostic Breast Cancer (WPBC) dataset on the Generating Genetic Algorithm (GGA), Another Genetic Algorithm (AGA), Yet another Generating Genetic Algorithm (YAGGA) and YAGGA2 feature selection approaches. The authors attained 93.62% accuracy with DT. However, the authors obtained lower accuracy with lots of feature selection methods, and the detection of breast cancer with feature selection methods still has room for improvement. Lorencin et al. [12] developed an Artificial Neural Network (ANN) model on the Breast Cancer Wisconsin Diagnostic (BCWD) dataset with a 0.95 AUC score. Lavanya at el [13]. used DTC-based classification and regression tree (CART) methods on the BCWD dataset. The CART method demonstrated higher accuracy with 94.72%, with feature selection approaches. In recent years, in 2023, Elsadig et al. [14] proposed a SVM machine learning method with 97.7% accuracy on five distinct feature selection methods, where there is still potential for enhancement of the model's performances with the feature selection method, moreover, the authors did not mention that how the model was built. In the same year, another study was constructed for the detection of breast classification, where Sahu et al. [15] used numerous datasets for the deep-learning and machine-learning approaches, in the machine learning method, they used the WBDC1 dataset and obtained 99.10% accuracy. Melekoodappattu et al. [16] proposed extreme learning machines with the Fruitfly Optimization Algorithm (ELM-FOA) and obtained 99.04% accuracy. The ensemble method with the optimization process is more complex to develop in future. et al. [17] applied a hybrid machine-learning model with 99.65% accuracy. Kadhim et al. [18] proposed GB model for the detection of breast cancer classes. However, they failed to describe the model's architecture and the parameters in which way they were implemented.

However, according to the previous outcomes, several methodologies have been used to evaluate breast cancer prediction on various data repositories. Many researchers applied various types of strategies, however, there are still a few difficulties while detecting, such as insufficient datasets, feature selection, and implementations of classification methods and there is still a question remaining on which methods are more efficient for the detection of breast cancer. To deal with these challenges, we have implemented two feature selection approaches to elucidate the most impactful features and also used ensembled approaches to acquire a more improved and balanced performance while distinguishing malignant and benign tumors.

3. Research methodology

3.1. Dataset overview

In this work, a comprehensive suite of approaches is built to attain remarkable accuracy in early cancer diagnosis. The analysis includes the employing of the WBCD1 and WBCD2 datasets, both of which are accessible for free on Kaggle, to ensure scientific rigor and availability [19,20].

According to Fig. 1, a total of 569 cases have been analyzed with 357 benign occurrences (62.74%) and 212 malignant instances (36.26%) of the entire dataset. WBCD1 produced 30 features, whereas WBCD2

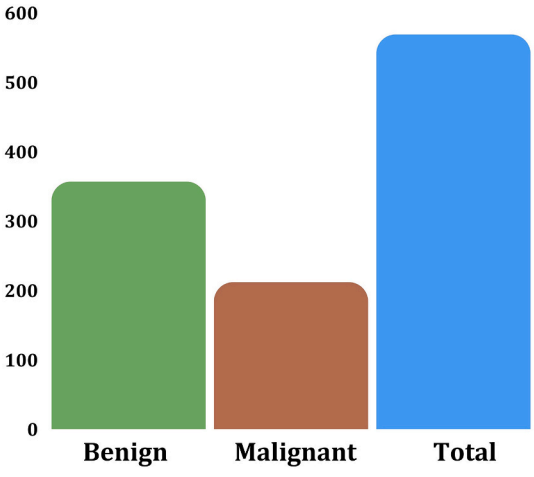

Fig. 1. Class distribution of the dataset of breast cancer.

provided 9 different features. To increase dataset size and accuracy, WBCD1 and WBCD2 were merged, leading to the collection of 39 aspects for the study.

3.2. Data preprocessing

In this paper, we used a variety of preprocessing strategies to ensure the robustness of our models. In the beginning, we thoroughly explored the dataset to identify and rectify any missing values that were absent in the dataset. Afterward, we determined that the dataset's features possessed an abundance of data types, with 10 features containing integer values, 30 appearing floating-point values, and one categorized as an object. For standardization of these data types, the study opted for the label encoding approach, which essentially converts all values into numerical representations for consistency and compatibility among the models [21]. In implementing the label encoding method to refine the collected dataset, the fit transform approach has been utilized. During the "fit" step, the label encoding method was modified according to the categorical variable's various classes, delivering numerical labels to every category. Throughout the transformation process, these collected values were replaced with their corresponding numerical representation. Following that, we minimized the dataset through the elimination of extraneous data (notably lowering the 'id' column). Particularly, we used the standard deviations approach to find outliers in the rigorously curated dataset, we found there were no noisy numbers [22]. This procedure guaranteed that the information sets were streamlined and suitable for successful model training, resulting in valid samples for improved performance. For comprehensive insight, all feature range values have been detailed in Table 1.

3.3. Overview of the proposed approach

This study conducted an extensive analysis to improve the efficiency of breast cancer detection. Several stages, as shown in Fig. 2, were thoroughly executed to determine the best procedure. These processes began with the implementation of various organizing phases to the integrated WBCD1 and WBCD2 datasets, according to section 3.2 on data preprocessing. After that, five traditional machine learning models: Logistic Regression (LR), Decision Tree Classifier (DTC), Extremely Randomized Tree Classifier (ERT), Adaboost Boosting Classifier (ABBC), Linear Discrimination Analysis (LDA), and two meta models: Hard-Voting Classifier (HVC), and Soft-Voting-Classifier (SVC) developed employing a hyperparameter tuning approach. These models were rigorously evaluated using a variety of performance criteria on the WBCD1 and WBCD2 datasets independently. Following that, the most appropriate features were selected using LASSO and SHAP algorithms. All models were thoroughly evaluated using a 10-fold cross-validation approach. However, to delve deeper into model performance, extensive analyses were carried out on the combined dataset, encompassing all features, LASSO-selected features, and SHAP-derived features. Incredibly, the majority of models showed improved accuracy and performance metrics on the merged dataset. The findings highlighted the efficacy of integrating SHAP-based methods, ultimately boosting the accuracy and reliability of breast cancer detection models.

3.4. Application of the proposed method

The proposed strategy's broad suitability improves the system's capacity for dealing with emergencies in the real world. This section elaborates on this concept. As illustrated in Fig. 3, the requisite procedures illustrate the implementation of the technique within a healthcare setting. In the beginning, patient record information is submitted into the database, and then relevant attributes are extracted from these entries to create inputs for the constructed approaches. Then, the model applies predetermined criteria to generate outputs, denoted as either 0 or 1, where 0 signifies the absence of cancer and 1 indicates the malignancy.

Table 1
Summary of dataset information post-preprocessing stages.

Features Name	Data Types	Value Range
Class	Integer	0 and 1
Radius mean	Integer	0 to 455
Texture mean	Integer	0 to 478
Perimeter mean	Integer	0 to 521
Area mean	Integer	0 to 538
Smoothness mean	Integer	0 to 473
Compactness mean	Integer	0 to 536
Concavity mean	Integer	0 to 536
Concave points mean	Integer	0 to 541
Symmetry mean	Integer	0 to 431
Fractal dimension mean	Integer	0 to 498
Radius se	Integer	0 to 539
Texture se	Integer	0 to 518
Perimeter se	Integer	0 to 532
Area se	Integer	0 to 527
Smoothness se	Integer	0 to 546
Compactness se	Integer	0 to 540
Concavity se	Integer	0 to 532
Concave points se	Integer	0 to 506
Symmetry se	Integer	0 to 497
Fractal dimension se	Integer	0 to 544
Radius worst	Integer	0 to 456
Texture worst	Integer	0 to 510
Perimeter worst	Integer	0 to 513
Area worst	Integer	0 to 543
Smoothness worst	Integer	0 to 410
Compactness worst	Integer	0 to 528
Concavity worst	Integer	0 to 538
Concave points worst	Integer	0 to 491
Symmetry worst	Integer	0 to 499
Fractal dimension worst	Integer	0 to 534
Clump thickness	Integer	0 to 9
Cell Size	Integer	0 to 9
Cell Shape	Integer	0 to 9
Marginal Adhesion	Integer	0 to 9
Signle Cell size	Integer	0 to 9
Bare Nuclei	Integer	0 to 9
Bland	Integer	0 to 9
Nucleoli	Integer	0 to 9
Mitoses	Integer	0 to 9

3.5. Explanations of applicable features

Based on the analysis of the datasets, it has been established that the majority of features exhibit a high correlation with the target feature (class). As depicted in Fig. 4, all correlation values are depicted within the range of positive 1 to negative 1 (1 to −1), with darker red shades representing highly correlated features (close to 1) and dark blue shades indicating less correlation (closer to −1). Features associated with the target factor are highlighted in red, signifying their relevance, while those lacking correlation with the target variable appear in blue.

It has been determined that 12 features have a minor relationship to the target feature. These features include smoothness mean, symmetry mean, fractal dimension mean, radius se, smoothness se, compactness se, symmetry se, fractal dimension se, smoothness worst, symmetry worst, fractal dimension worst, and mitoses.

Furthermore, 27 features exhibit correlation with the target feature, excluding the "class" feature. Nine of these features exhibit a strong association, including parameter worst, concave points mean, radius worst, area worst, concave points worst, cell size, cell shape, bare nuclei, and bland. Specifically, bare nuclei appear as the most closely connected feature, with a correlation value of 0.9 out of 1. Moreover, all eight features have correlation values of 0. 8. Aside from these nine variables, 19 additional features indicate a frequent connection, with 10 having coefficients of 0.7, four with values of 0.6, and the remaining four with coefficients of 0.5 out of 1.

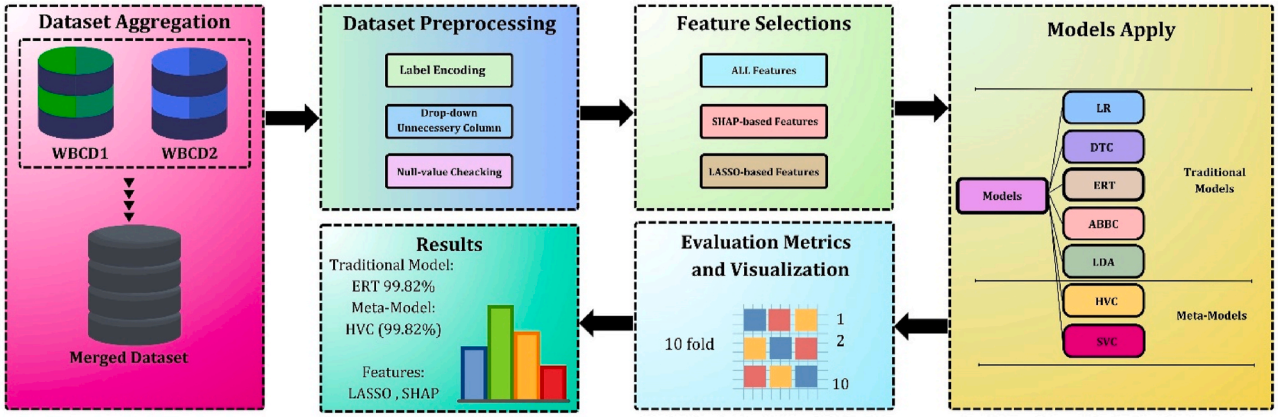

Fig. 2. Methodology of the study depicted here includes data collection, preprocessing, feature selection, model utilization, evaluation metrics, and results analysis.

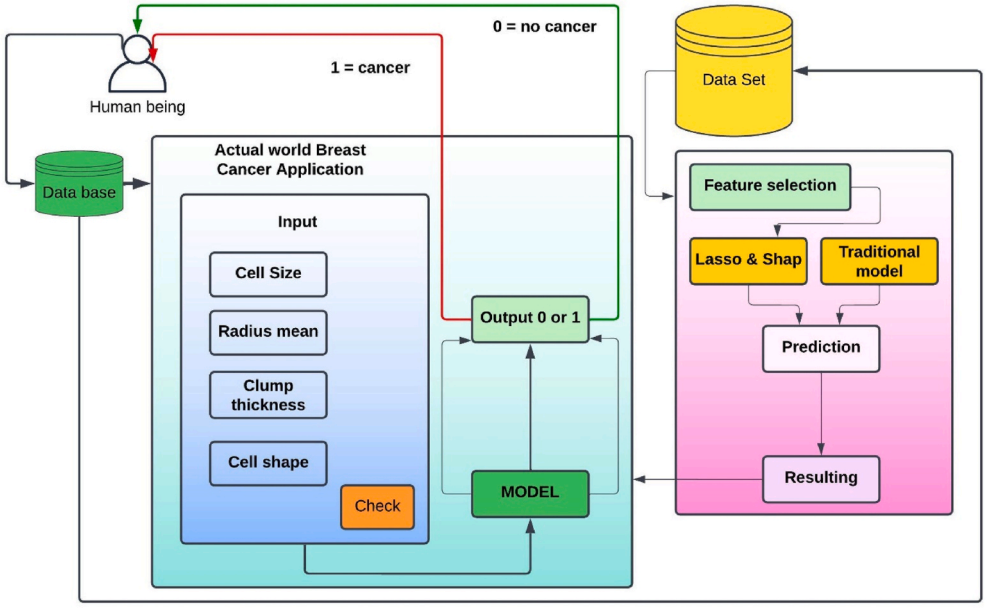

Fig. 3. Application structure of the proposed approach based on real-world scenarios, where users input the data and predict the outcomes.

4. Supervised learning libraries and feature selection methods

Supervised machine learning plays an essential role in the medical sector because it enables the creation of prediction models which could significantly enhance patient care and operational efficiency. These models can discover patterns and predict patient diagnoses, treatment results, and disease progression by training algorithms on labeled datasets in which inputs have connections with recognized outcomes. This predictive potential has significance for early diagnosis of diseases including cancer when prompt intervention can significantly increase survival rates according to the best-fit features. Therefore, we used LASSO and SHAP feature selection approaches. Since our dataset possessed two classes, we used a variety of supervised machine-learning algorithms. We used five standard models and combined them to create two metamodels. Our technique incorporates many models, including LR, ERT, DTC, ABBC, and LDA. These models were fine-tuned with soft and hard voting classifiers. The next subsections provide a detailed description of the performance of each model and the related meta-models.

4.1. Logistic Regression (LR)

LR is among the most popular and common models in machine learning. LR is used to highlight dummy variables that accurately represent a result of yes or no, while yes, is represented by 1 and no is defined as 0. This classifier generally implies that conventional techniques are acceptable [23–25]. Fig. 5 illustrates the procedure of the LR model.

4.2. Decision Tree (DTC)

One of the most powerful algorithms in machine learning techniques is the DTC. This strategy includes several vertices like a tree, as shown in Fig. 6. Every branch of vertices has a specific category that belongs to an evaluating feature and is relevant to the outcomes [26]. While splitting the tree, selecting the optimal feature, and setting it at the vertices of the tree assumption is important. After that, it produces excellent elements and generates an effective result [27,28].

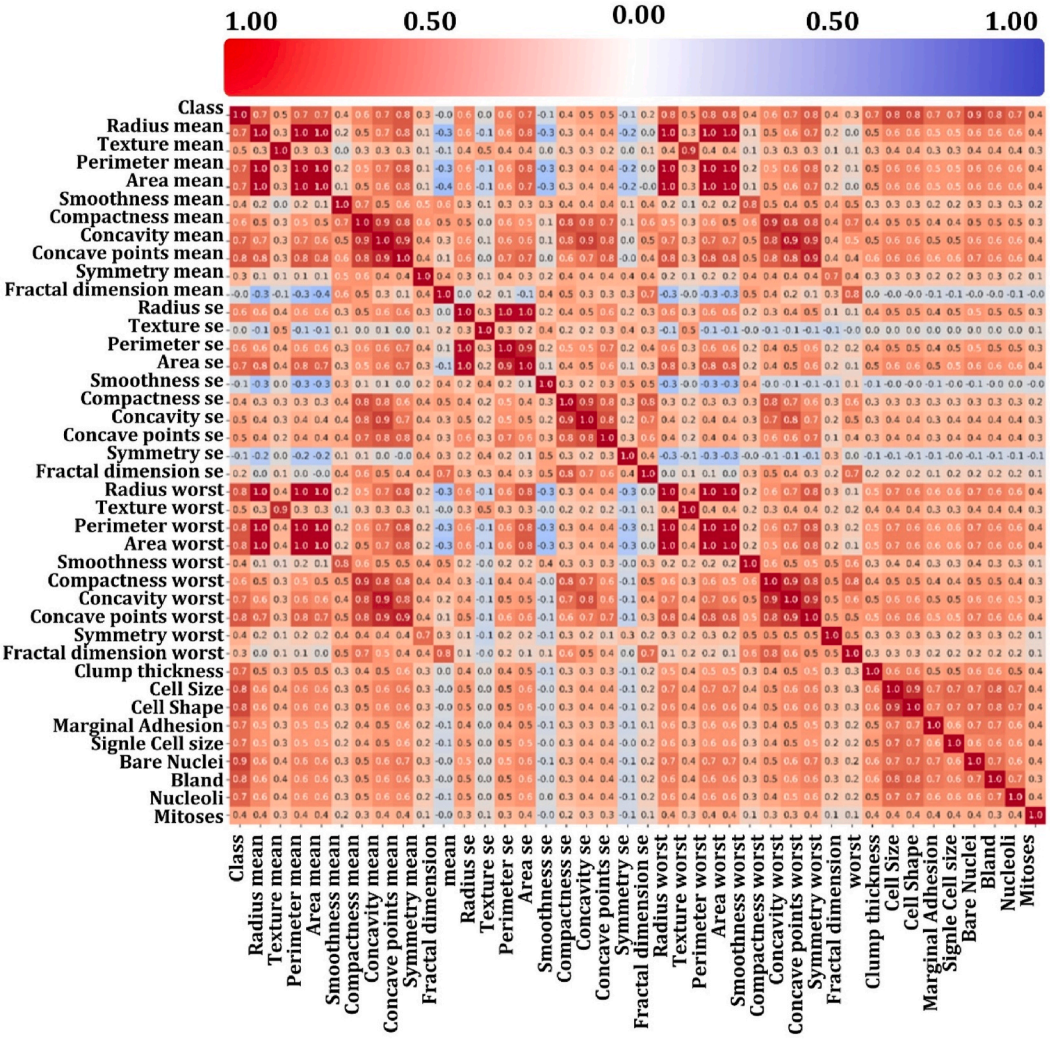

Fig. 4. Representation of highly correlated features in the combined dataset, indicating a strong association between variables.

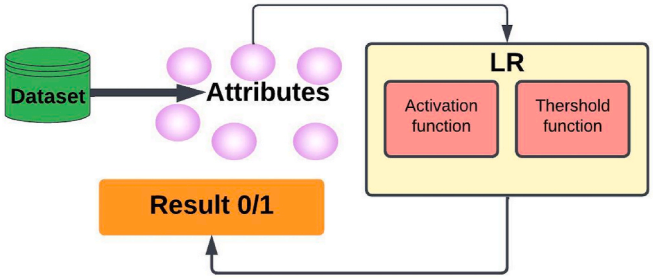

Fig. 5. Logistic Regression algorithm's working process with two functions and attributes for the actual binary classification.

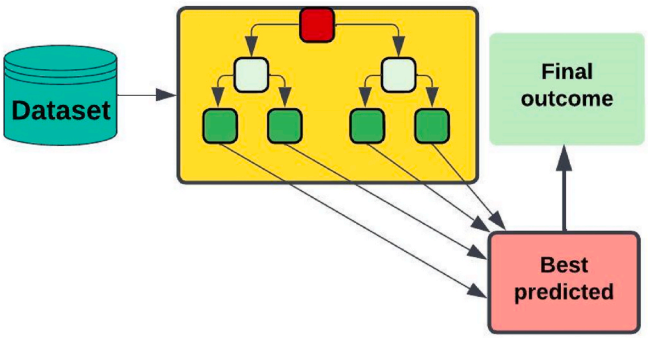

Fig. 6. Decision Tree algorithm's overall overview, where the models predict the whole tree values and provide the outcome.

4.3. Extremely Randomized Trees (ERT)

ERT is an ensemble approach that is useful in a variety of DTC-developed methods [29]. This method is relatively like Random Forest (RF) approaches because this method detects the feature as a randomized state and decreases variance as possible [30]. Therefore, this randomized state generated the best-split features and produced the potential result. Besides that, this approach has a significantly higher deviation [31]. Fig. 7 demonstrates the overall procedure of the ERT.

4.4. Adaboost Boosting Classifier (ABBC)

Adaboost is a boosting ensemble approach that is capable of boosting the insufficient to powerful and effective classifier [32]. That approach has the benefit of becoming applicable to practically all existing supervised learning models for boosting the outcome as well as Several classifications' determinations could minimize the lower threshold of classifier [33,34]. Then it provides a better outcome according to Fig. 8.

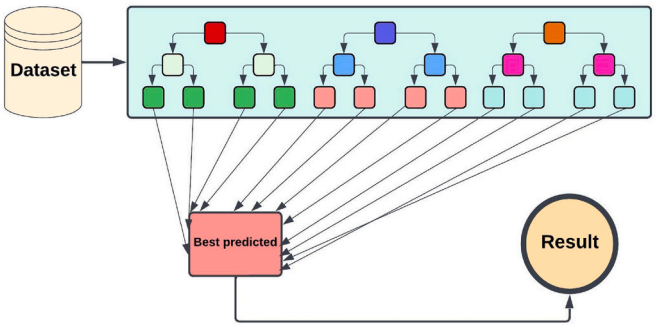

Fig. 7. General overview of the Extremely Randomized Trees model, illustrating outcome predictions based on the optimal trees.

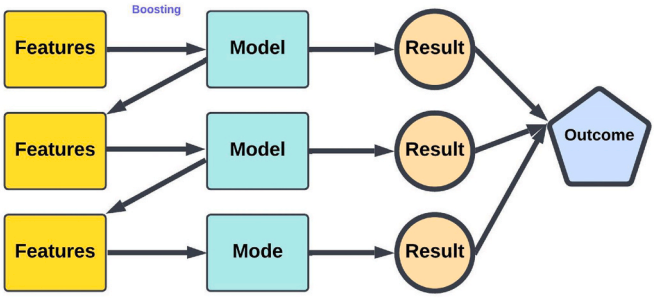

Fig. 8. Working strategy of the Adaboost Boosting Classifier algorithm, where enhancing weak Features to strengthen and deliver predicted outcomes.

4.5. Linear Discrimination Analysis (LDA)

LDA is an efficient technique in the field of machine learning. This approach is popular for decreasing the dimensionality of the entire dataset in a pre-processing phase for distinct classification [35]. Following that, LDA selects the best components from the datasets produces the distinctive best features, and generates the effective result [36]. Fig. 9 depicts the overall workflow of this model.

4.6. Meta-models (Voting classifiers)

A comprehensive review of "meta-model" is commonly associated with what the phrase "meta" indicates in a diversity of many other, accurate instances [37,38]. This meta-model integrated and finalized the output results of various models. In this paper, we used two meta-models the soft and hard voting classifiers as shown in Fig. 10, which have been merged with the proposed traditional models. In hard voting, each

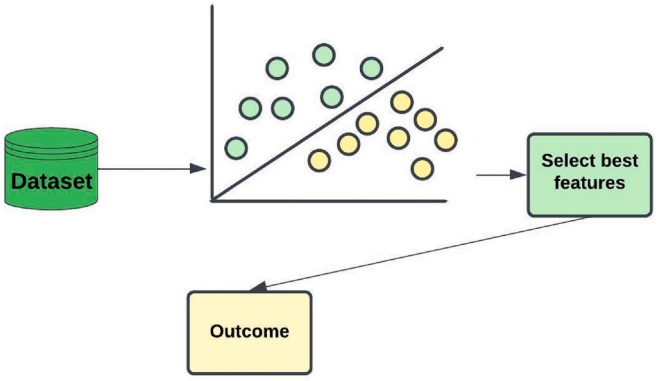

Fig. 9. Linear Discrimination Analysis model's overall method, where the best-fit line discrimination the classes.

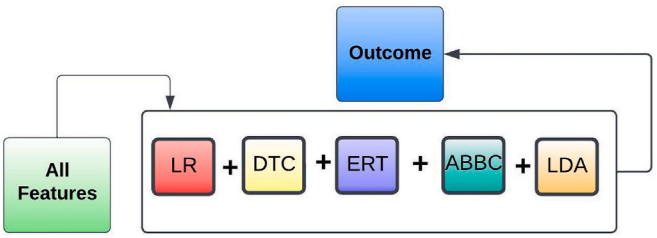

Fig. 10. HVC and SVC-based meta-models overview of the study incorporating all features and five traditional models.

model in the ensembles estimates the class label for a given input, and the outcome of the forecast is decided by an overwhelming decision. Soft voting is frequently used with classifiers that can generate estimates of probability for all classes and combine them to make a final decision. The ensemble voting Classifier is a meta-classifier that integrates machine learning approaches and voting for the best outcomes. Both hard voting (HVC) and soft voting (SVC) are enabled by the voting classifier [39–41]. Where HVC predicts the most votes in this classifier and SVC predicts the best possibilities from the datasets. In this paper, we used (LR + DTC + ERT + ABBC LDA) for HVC and SVC outcomes.

4.7. LASSO feature selection technique

LASSO is a powerful feature selection approach used to decrease dimensionality. Several methods are used to reduce complexity [42]. LASSO finds significant parameter correlation coefficients among derived feature subsets while identifying and discarding insufficient features [43]. In this work, we used this selection technique, which led to the discovery of 31 essential characteristics. Then we used our standard and meta-model techniques. Fig. 11 shows the entire process of the LASSO feature selection approach.

4.8. SHAP feature selection technique

SHAP methodology is one of the most effective feature selection methods. SHAP provides a relevance score for each parameter in every forecast. In 2017, Lundberg and Lee proposed SHAP, emphasizing that it provides a consistent framework for understanding predictions [44]. In this study, we used SHAP as a feature selection approach to determine the most frequent variables across all models. From Fig. 12, this technique generated 20 common characteristics, which were then utilized to produce a dataset and evaluated the outcomes of this dataset using meta-models and other classical models.

4.9. Performance analysis method

This study used seven evaluation performance metrics to detect breast cancer more with balancing outcomes such as Precision (PRE), Recall (RE), F1 score (F1S), Cohen Kappa Score (K), Matthew's Correlation Coefficient (MCC), Execution time and Accuracy score. These analytical methods provided reliable estimation. These methods formulation can be stated as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1\ Score = \frac{2(Precision \times Recall)}{(Precision + Recall)} \tag{4}$$

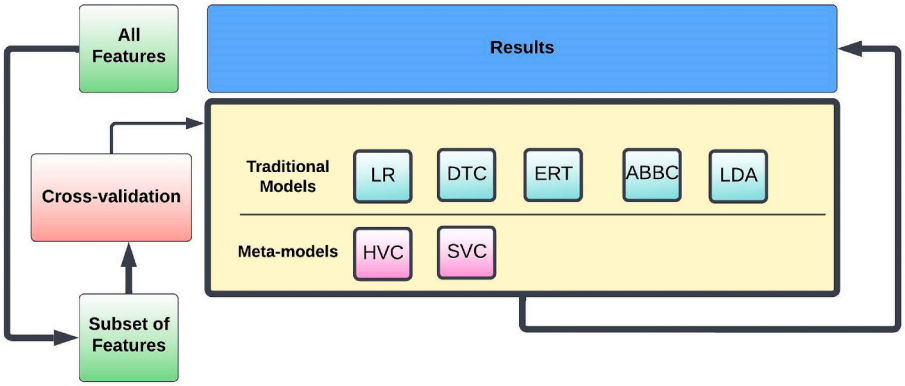

Fig. 11. Application of models on the LASSO-based feature selection procedure.

$$\mathrm{MCC} = \frac{TP.TN - FP.FN}{\sqrt{TP + FP.(TP + FN).(TN + FP)}} \tag{5}$$

$$Kappa = \frac{2(TP.TN - FP.FN)}{(TP + FP)\cdot(FP + TN) + (TP + FN)\cdot(FN + TN)} \tag{6}$$

All of the methods utilize the initials TP, TN, FP, and FN, which stand for True Positive, True Negative, False Positive, and False Negative, respectively. TP refers to properly recognized positive aspects, whilst TN refers to correctly detected negative examples. In contrast, FP shows the misclassification of negative occurrences as positive, and FN suggests the inaccurate identification of positive examples as negative [45,46]. However, with these mathematical formulations, we tested our models with several hyperparameters of each model, because hyperparameters play an important role in fine-tuning machine learning models to improve performance, stability, and flexibility to specific necessities or constraints. In Table 2, this study demonstrates the hyper-tuning process of our applied machine-learning models.

5. Result and discussion

Machine learning techniques serve as crucial to finding breast cancer because of their ability to analyze difficult patterns in health information, and produce accurate diagnoses based on significant feature selection methods, which play a critical role in improving model performance to determine the most relevant features for categorization. Therefore, in the present study, we employed SHAP and LASSO-based feature selection approaches with the WBCD1 and WBCD2 datasets. Our study found that integrating these data sets results in enhanced breast cancer diagnosis. Importantly, we used all validated medical data in our study. Our selection regarding maintaining the original information despite sampling over had been deliberate, facilitating us to evaluate model behavior with unbalanced datasets. Considering having a slightly smaller number of malignant cases compared to benign instances, our models performed satisfactorily without excessive sampling. This demonstrates the effectiveness of our preprocessing method, which effectively addressed the challenges encountered with unbalanced datasets. Accordingly, in this work, we proposed the SHAP-based methods because of their importance in producing interpretable model results. This strategy was chosen because SHAP not only provides a clear perspective of model predictions but also consistently performs well across a variety of models. In this section, we presented the overall outcomes of the feature selection methods and model performances of the study.

5.1. The outcomes of feature selection procedures

In this phase, we provide the top features identified through correlation analysis, LASSO, and SHAP feature selection techniques. Tables 3 and 4 present comprehensive feature assessments.

LASSO feature selection identified 18 features that are highly linked with the target variable. Notably, "bare nuclei" had the highest correlation coefficient, 0.851. This characteristic is critical in breast cancer diagnosis since it shows the presence or lack of nuclei in cells, which is a reliable sign of malignancy. Cells with greater "bare nuclei" levels frequently exhibit aberrant cell proliferation, a characteristic of malignant tissue. "Clumps thickness" and "cell size" follow closely after, with coefficients of 0.832 and 0.831, respectively. "Clump thickness" measures the compactness of cell clusters, with higher values indicating more tumor malignancy. "Cell size" is another important indication, as bigger cell sizes are commonly associated with malignant tumors, suggesting fast cell proliferation. Similarly, SHAP feature selection identified 16 aspects that are highly linked with the target variable. Furthermore, "bare nuclei" was the highest-ranked distinctive, with a score of 0.851. Its significance stems from its capacity to provide insights into cellular abnormalities, acting as a crucial biomarker for malignant growth. Following "bare nuclei," "cell size" received notable attention, with a score of 0.832. This trait emphasizes the relevance of cellular morphology in cancer detection, as higher cell sizes are generally symptomatic of malignancy.

These top features obtained through correlation analysis, LASSO, and SHAP feature selection approaches provide important insights into the underlying properties of breast cancer tumors.

5.2. Models performances

In this section, we provide a comprehensive overview of the performance metrics of the applied machine learning models across different feature sets, such as all features, LASSO-selected features, and SHAP-selected features, analyzed separately for the WBCD1 and WBCD2 datasets, as well as their combined dataset. Tables 5–7 illustrate the models' performance based on an extensive analysis of all characteristics and specific features. These tables provide crucial performance parameters that provide information on the models' ability to distinguish between benign and malignant breast cancers across different feature sets.

Table 5 evaluates the performance of seven machine learning models on three distinct datasets: Merged set, WBCD1, and WBCD2, taking into account all attributes. These datasets contain various subsets or combinations of data related to breast cancer diagnosis. In the Merged set, all conventional models show great accuracy, ranging from 98.59% to 99.82%. Notably, the ERT model had the best accuracy at 99.82%. Among the meta-models, HVC and SVC had the best accuracy of 99.82%. In the WBCD1 dataset again, the ERT and meta models (HVC and SVC) had a maximum accuracy of 99.82%. The WBCD2 dataset shows consistent performance with high accuracy rates across all models. The ERT model and meta models (HVC and SVC) attained the maximum accuracy of 99.82%. Overall performance across datasets is very

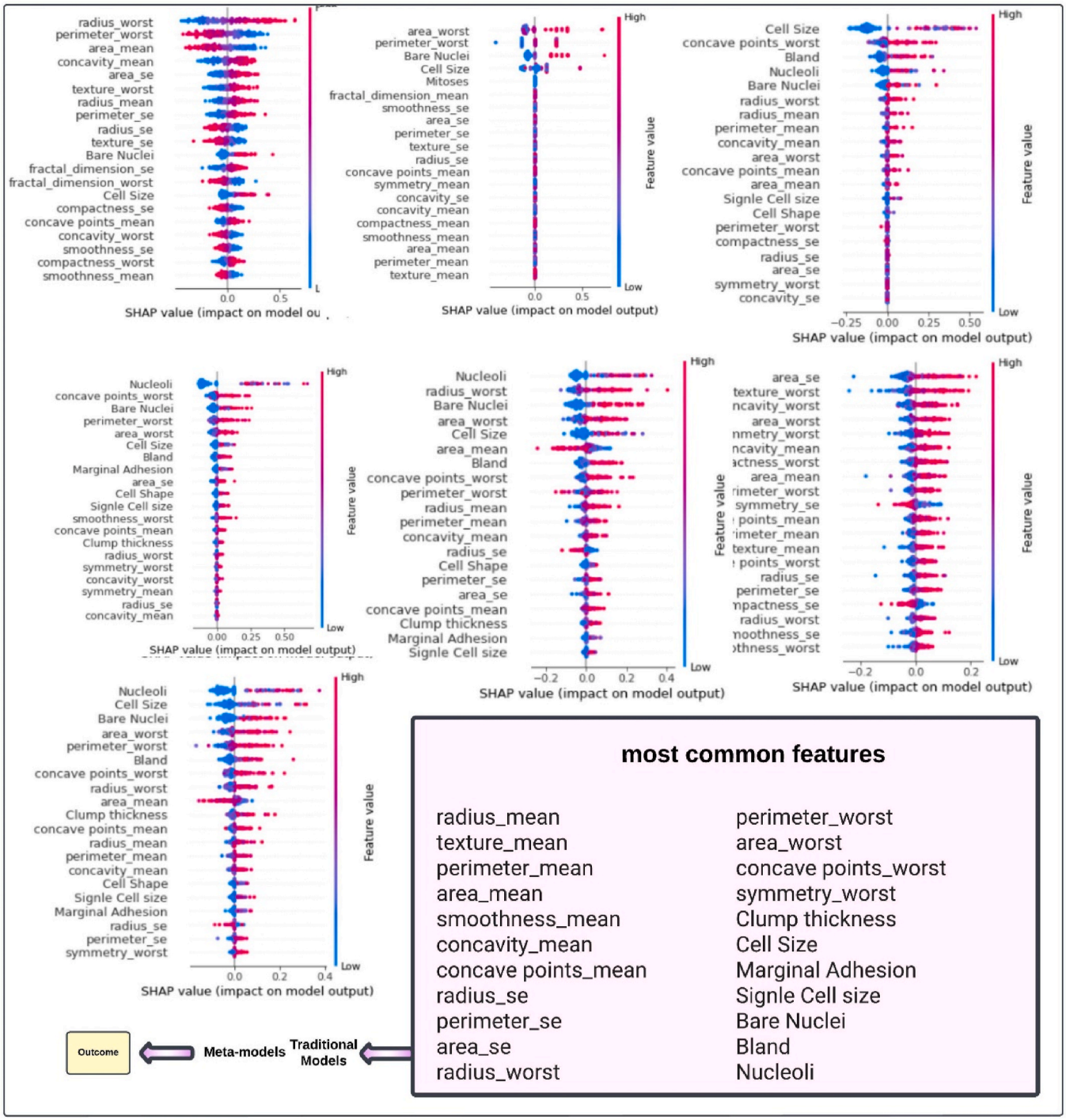

Fig. 12. SHAP-based feature selection procedure, highlighting the extraction of attributes from seven models to select the top 20 common features.

Table 2
Summary of all the hyperparameters of seven models in this study.

Mode	Hyperparameter
LR	random state = 100, penalty = 'l1', solver = 'liblinear'
DTC	random state = 10
ERT	n estimators = 100, random state = 10
ABBC	n estimators = 200, random state = 10, learning rate = 1
LDA	solver = 'eigen', shrinkage = 'auto'
HVC	(LR, DTC, ERT, ABBC, LDA)'s default parameters, voting = 'hard'
SVC	(LR, DTC, ERT, ABBC, LDA)'s default parameters, voting = 'soft'

constant, with the ERT and meta-models consistently having the greatest accuracy rates with a more balancing performance in every evaluation method. The model execution times varied between datasets and classifiers, although usually were modest. Across all datasets, traditional model execution times varied from 0.0135 to 0.9941 s, whereas meta-model execution times ranged from 0.5477 to 0.8346. SVC had minimal execution times, while DTC had the fastest. Despite minor variances, execution durations were consistently low across all models, showing efficient processing and model inference.

From Table 6, machine learning models perform consistently well across various datasets and classifiers, suggesting their efficacy in breast

Table 3
LASSO-based feature ranking.

Features Name	Features Score
Perimeter mean	0.747
Compactness mean	0.606
Concavity mean	0.732
Concave points mean	0.780
Perimeter se	0.633
Area se	0.722
Radius worst	0.734
Perimeter worst	0.786
Area worst	0.786
Compactness worst	0.608
Concave points worst	0.780
Concavity worst	0.705
Clump thickness	0.725
Cell size	0.832
Cell shape	0.831
Marginal adhesion	0.704
Bare nuclei	0.851
Nucleoli	0.739

Table 4
SHAP-based feature ranking.

Features Name	Features Score
Radius mean	0.743
Perimeter mean	0.747
Area mean	0.735
Concavity mean	0.732
Perimeter se	0.633
Radius worst	0.794
Perimeter worst	0.786
Area worst	0.786
Concavity worst	0.705
Clump thickness	0.725
Cell size	0.832
Signle cell size	0.701
Marginal adhesion	0.704
Bare nuclei	0.851
Nucleoli	0.739
Bland	0.795

cancer diagnosis. Across all datasets, the ERT model consistently achieved the best accuracy, reaching 99.82% in the Merged set and WBCD2 dataset. Meta models, such as HVC and SVC, also performed well, with accuracy rates ranging from 97.01% to 99.65% across datasets. In terms of execution time, all models had reasonably low processing durations, with the majority lying between 0.0140 and 1.4070 s. The execution time for the HVC model in the merged dataset was 0.54372 s, while the SVC model's execution time was 0.7820 s. These execution times are reasonable and efficient for practical applications, especially considering the high accuracy achieved by both models, where ERT achieved the highest accuracy in the merged set with 0.2369 s execution times.

From Tables 7 and In the combined dataset, all models, traditional and meta, had extremely high accuracy rates ranging from 99.12% to 99.82%. The LR, ERT, ABBC, LDA, HVC, and SVC models, all had a maximum accuracy of 99.82%. These models also performed well on other measures like MCC, Precision, Recall, and F1-score. On the WBCD1 dataset, conventional models performed worse than the merged dataset, with accuracy rates ranging from 91.21% to 96.31%. However, meta-models HVC and SVC attained 96.66% accuracy rates, demonstrating their efficacy in this dataset. In the WBCD2 dataset, conventional models achieved accuracy rates ranging from 97.19% to 99.47%, with ERT and LR attaining the maximum accuracy of 99.47%. The meta-models HVC and SVC also performed well in this dataset, with accuracy rates of 99.30% and 98.95%, respectively. Overall, the execution times for all models across datasets were within reasonable limits, ensuring timely processing of breast cancer diagnostic classifications. These efficient execution times complemented the high accuracy rates of the models, making them practical choices for real-world applications in breast cancer detection.

5.3. Comparison of different model's performance

This segment presents a comparative performance analysis of the adopted model output in Fig. 13. Six different statistical analyses that are compatible with various sub-visualization portions have been used. Each visualization section includes adopted model performances for all datasets, along with LASSO and SHAP feature extraction. The green bar reflects all features, the blue color represents LASSO based features, and the other one is the SHAP based feature selection's.

In the ACC subsection, LR is top placed for SHAP features with 99.82 %, which indicates an increase in both selections like all features and LASSO features. DTC, ERT, and LDA have stayed constant throughout all segments. ABBC has scored well in SHAP feature procedures, getting 99.82 %, which has been higher than in other categories. HVC and SVC have stayed stable for all features and SHAP features, besides a decrease in LASSO features.

In the MCC subsection, LR has performed well throughout all

Table 5
Performance analysis of seven models using all features on different datasets.

Datasets	Category	Classifier	ACC	MCC	K	PRE	RE	F1S	Execution time
Merged set	Traditional models	LR	98.59%	96.99%	96.98%	98.59%	98.59%	98.59%	0.2772
		DTC	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.0135
		ERT	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.1420
		ABBC	99.64%	99.25%	99.25%	99.65%	99.65%	99.65%	0.9133
		LDA	99.12%	98.13%	98.11%	99.12%	99.12%	99.12%	0.1630
	Meta models	HVC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.5630
		SVC	99.82%	99.25%	99.25%	99.65%	99.65%	99.65%	0.6543
WBCD1	Traditional models	LR	98.95%	97.76%	97.73%	98.95%	98.95%	98.95%	0.4442
		DTC	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.0177
		ERT	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.3013
		ABBC	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.9941
		LDA	99.12%	98.13%	98.11%	99.12%	99.12%	99.12%	0.0235
	Meta models	HVC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.5477
		SVC	99.65%	99.25%	99.25%	99.65%	99.65%	99.65%	0.8320
WBCD2	Traditional models	LR	98.77%	97.37%	97.36%	98.77%	98.77%	98.77%	0.4075
		DTC	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.0181
		ERT	99.65%	99.25%	99.25%	99.65%	99.65%	99.65%	0.5899
		ABBC	99.65%	99.25%	99.25%	99.65%	99.65%	99.65%	0.9640
		LDA	99.12%	98.13%	98.11%	99.12%	99.12%	99.12%	0.0159
	Meta models	HVC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.5668
		SVC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.8346

Table 6
Performance analysis of seven models using LASSO features on different datasets.

Datasets	Category	Classifier	ACC	MCC	K	PRE	RE	F1S	Execution time
Merged set	Traditional models	LR	98.59%	96.99%	96.99%	98.59%	98.59%	98.59%	0.2949
		DTC	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.0148
		ERT	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.2369
		ABBC	99.30%	98.50%	98.49%	99.30%	99.30%	99.30%	0.8634
		LDA	99.12%	98.13%	98.11%	99.12%	99.12%	99.12%	0.2236
	Meta models	HVC	99.65%	99.25%	99.25%	99.65%	99.65%	99.65%	0.54372
		SVC	99.65%	99.25%	99.25%	99.65%	99.65%	99.65%	0.7820
WBCD1	Traditional models	LR	95.78%	90.98%	90.98%	95.78%	95.78%	95.78%	0.3222
		DTC	92.44%	84.08%	84.00%	92.44%	92.44%	92.44%	0.0140
		ERT	97.01%	93.60%	93.57%	97.01%	97.01%	97.01%	0.1512
		ABBC	95.96%	91.33%	91.31%	95.96%	95.96%	95.96%	0.8605
		LDA	97.19%	94.00%	93.93%	97.19%	97.19%	97.19%	0.0256
	Meta models	HVC	97.01%	93.60%	93.58%	97.01%	97.01%	97.01%	0.5509
		SVC	96.66%	92.84%	92.82%	96.66%	96.66%	96.66%	0.4971
WBCD2	Traditional models	LR	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.0324
		DTC	98.24%	96.24%	96.23%	98.24%	98.24%	98.24%	0.0147
		ERT	99.30%	98.50%	98.49%	99.30%	99.30%	99.30%	0.2416
		ABBC	98.95%	97.74%	97.74%	98.95%	98.95%	98.95%	1.4070
		LDA	96.66%	92.97%	92.73%	96.66%	96.66%	96.66%	0.0170
	Meta models	HVC	98.95%	97.74%	97.74%	98.95%	98.95%	98.95%	0.2552
		SVC	98.95%	97.74%	97.74%	98.95%	98.95%	98.95%	0.2576

Table 7
Performance analysis of seven models using SHAP features on different datasets.

Datasets	Category	Classifier	ACC	MCC	K	PRE	RE	F1S	Execution time
Merged set	Traditional models	LR	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.4743
		DTC	99.12%	98.12%	98.12%	99.12%	99.12%	99.12%	0.0175
		ERT	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.2292
		ABBC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.7476
		LDA	99.30%	98.50%	98.49%	99.30%	99.30%	99.30%	0.0327
	Meta models	HVC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.7179
		SVC	99.82%	99.62%	99.62%	99.82%	99.82%	99.82%	0.4499
WBCD1	Traditional models	LR	96.13%	91.71%	91.70%	96.13%	96.13%	96.13%	0.3500
		DTC	91.21%	81.24%	81.24%	91.21%	91.21%	91.21%	0.0074
		ERT	96.31%	92.09%	92.04%	96.31%	96.31%	96.31%	0.1418
		ABBC	96.13%	91.71%	91.68%	96.13%	96.13%	96.13%	0.9191
		LDA	96.31%	92.08%	92.07%	96.31%	96.31%	96.31%	0.0173
	Meta models	HVC	96.66%	92.84%	92.81%	96.66%	96.66%	96.66%	0.3404
		SVC	95.61%	90.57%	90.56%	95.61%	95.61%	95.61%	0.3410
WBCD2	Traditional models	LR	99.47%	98.87%	98.87%	99.47%	99.47%	99.47%	0.0104
		DTC	98.07%	95.86%	95.85%	98.07%	98.07%	98.07%	0.0062
		ERT	99.47%	98.87%	98.87%	99.47%	99.47%	99.47%	0.2282
		ABBC	98.95%	97.74%	97.74%	98.95%	98.95%	98.95%	0.4401
		LDA	97.19%	94.07%	93.89%	97.19%	97.19%	97.19%	0.0091
	Meta models	HVC	99.30%	98.50%	98.50%	99.30%	99.30%	99.30%	0.2900
		SVC	98.95%	97.74%	97.74%	98.95%	98.95%	98.95%	0.2483

strategies with 96.99%. DTC and ERT have stayed constant in all segments with 98.12% and 99.62%. Whereas, in SHAP selection methods, ABBC with 99.62%, LDA with 98.50%, and meta-models with 99.62% outperform the other two selection approaches.

In K, the LR model considers the top in SHAP selection features over other segments which have obtained 99.62%. DTC with 98.12% and ERT models with 99.62% remained constant in the three different selection approaches. ABBC performed better in the SHAP feature procedures than the other two approaches with 99.62%, The LDA model has remained constant in LASSO and all feature selection methods, but in SHAP feature selection techniques this model has been increased suitably, the model has obtained 98.49%. When meta-models were used for all segmentation, HVC and SVC both performed better in SHAP techniques with 99.62%.

Further three statistical analysis methods, such as PRE, RE, and F1S methodologies, LR, LDA, ABBC, and SVC, topped the others with 99.82 %. DTC and ERT have stayed at 99.12% and 99.82%, respectively. HVC decreases in LASSO features and all features but when including SHAP features it appears at 99.82 % accordingly.

5.4. Comparison with the existing tools

Breast cancer is a major anxiety for women, encouraging considerable studies using a variety of datasets. In response, our paper recommends SHAP-based feature selection approaches, which provide an extensive arsenal for assessing sophisticated machine learning models. This technique assists in a more detailed understanding of model outputs in associating predictions to various variables and clarifying their unique contributions to overall predictions. This interpretability is vital in medical environments, where decisions have a significant impact. Table 8 compares the efficacy of our proposed approaches with alternative options. We aspire to enhance the area of breast cancer detection through robust scientific studies, enhancing accuracy in diagnosis and informing clinical decision-making.

Table 8 compares the accuracy of our proposed models with existing predictors for breast cancer detection. Our work provides a distinctive procedure, specifically SHAP-based feature selection methods, that shows significant gains in accuracy when compared to alternative approaches. Across various machine learning models, including LR, DTC, ERT ABBC, and LDA, our proposed models consistently exhibit superior

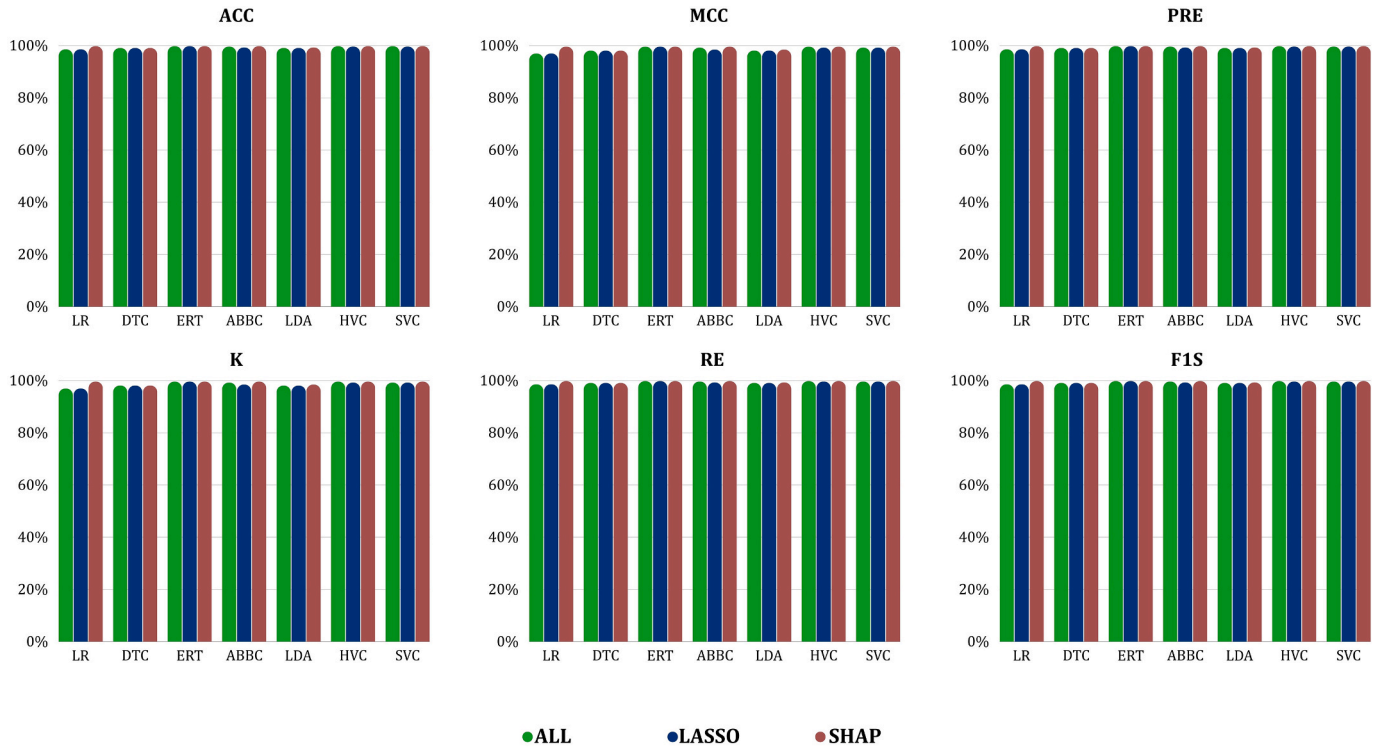

Fig. 13. Models' performance comparison on three modes, All features, LASSO features, and SHAP features methods.

Table 8
Overall comparison between the proposed model's accuracy with existing predictors.

Our works				Others work	Calculate (%) three methods		
Models Name	ALL Accuracy	LASSO Accuracy	SHAP Accuracy	Existing Accuracy	ALL Features	LASSO Features	SHAP Features
LR	98.59%	98.59%	99.47%	94.63% [47]	3.96% increase	3.96% increase	4.84% increase
DTC	99.12%	99.12%	99.12%	88.80% [6]	10.32% increase	10.32% increase	10.32% increase
ERT	99.82%	99.82%	99.82%	97.34% [47]	2.48% increase	2.48% increase	2.48% increase
ABBC	99.65%	99.30%	99.82%	92.53% [47]	7.12% increase	6.77% increase	7.29% increase
LDA	99.12%	99.12%	99.12%	95.33% [47]	3.79% increase	3.79% increase	3.79% increase
Meta-models	99.82%	99.65%	99.30%	98.31% [48]	1.51% increase	1.34% increase	0.99% increase
Meta-models	99.82%	99.65%	99.30%	97.96% [49]	1.86% increase	1.69% increase	1.34% increase
Meta-models	99.82%	99.65%	99.30%	98.49% [50]	1.33% increase	1.16% increase	0.81% increase

accuracy rates. Notably, our models achieve accuracy levels ranging from 98.59% to 99.82%, surpassing existing predictors by significant margins. In particular, in LR, DTC, ERT, ABBC, and LDA models, our suggested approaches improve accuracy by 3.96%, 10.32%, 2.48%, 7.12%, and 3.79%, respectively, over current predictors. These improvements demonstrate the usefulness of our method for improving diagnostic precision and prediction performance. Furthermore, meta-models developed from our proposed approaches demonstrate significant accuracy advancements. Our methodology produces accuracy gains ranging from 0.81% to 1.86% when compared to current predictors, illustrating the resilience and generalizability of our approach across diverse ensemble methods. Overall, our findings make substantial contributions to the area of breast cancer detection by proposing novel feature selection approaches that improve model interpretability and prediction accuracy. Through comparative research, we demonstrate the performance of our suggested models and highlight their potential to transform diagnostic methods in clinical settings. However, there are limitations to this study. We used smaller datasets, and although the malignant dataset is smaller than the benign samples, we did not use any balancing methods. This limitation is due to the research-focused nature of the study, and we still need to provide practical server applications for these models.

Future research in this domain will focus on enhancing the efficacy of classification strategies to estimate various feature selection techniques and predict outcomes using different models. Additionally, efforts will be directed towards leveraging tools to optimize model effectiveness, ensuring improved performance with larger datasets. Deep Learning techniques will be deployed with proper applications or webservers.

6. Conclusion

This study concentrated primarily on improving the accuracy of particular algorithms using two combined datasets, with different statistical assessors evaluating the model's effectiveness. A wide range of statistical studies were performed, demonstrating the effectiveness of the recommended approaches. When applied to all data, the ERT model and meta-models (SVC, HVC) obtained an impressive 99.82% accuracy. However, the SHAP and LASSO feature choices produced divergent results. Notably, only the ERT model outperformed in LASSO, however in SHAP strategies, LR, ABBC, and ERT all performed more effectively, with meta-models (SVC and HVC) delivering a remarkable 99.82% accuracy. When LASSO feature selection techniques were used for all features, LR, DTC, LDA, and ERT produced similar results. However, ABBC and meta-models experienced a reduction in accuracy by 0.35%

and 0.18%, respectively. Moreover, the inclusion of SHAP features resulted in enhanced performance across all models, with LR and ABBC models experiencing an improvement of 0.88 and 0.18 rate, respectively. Consequently, it can be deduced that SHAP feature selection demonstrated the optimal performance in this analysis. This study endeavors to refine existing methodologies through innovative model construction, facilitating the practical deployment of projects in real-world scenarios.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] L. Panigrahi, K. Verma, B.K. Singh, Ultrasound image segmentation using a novel multi-scale Gaussian kernel fuzzy clustering and multi-scale vector field convolution, Expert Syst. Appl. 115 (Jan.2019) 486–498, https://doi.org/10.1016/j.eswa.2018.08.013.

[2] World Health Organization, Switzerland, statistics on breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. (Accessed 20 December 2022).

[3] World Cancer Research Fund International, London, breast cancer statistics. https://www.wcrf.org/cancer-trends/breast-cancer-statistics/. (Accessed 20 December 2022).

[4] Mohammed Abdulrazaq Kahya, Classification enhancement of breast cancer histopathological image using penalized logistic regression, Indonesian Journal of Electrical Engineering and Computer Science 13 (1) (2019), https://doi.org/10.11591/ijeecs.v13.i1.pp405-410.

[5] S.A. Begum, T. Mahmud, T. Rahman, J. Zannat, F. Khatun, K. Nahar, M. Towhida, M. Joarder, A. Harun, F. Sharmin, Knowledge, attitude and practice of Bangladeshi women towards breast cancer: a cross sectional study, Mymensingh Med. J. 28 (1) (2019 Jan) 96–104. PMID: 30755557, https://pubmed.ncbi.nlm.nih.gov/30755557/.

[6] Tsehay Admassu Assegie, R. Lakshmi Tulasi, N. Komal Kumar, Breast cancer prediction model with decision tree and adaptive boosting, IAES Int. J. Artif. Intell. 10 (1) (2021) 184, https://doi.org/10.11591/ijai.v10.i1.pp184-190.

[7] Hakim El Massari, et al., An ontological model based on machine learning for predicting breast cancer, Int. J. Adv. Comput. Sci. Appl. 13 (2022) 7, https://doi.org/10.14569/IJACSA.2022.0130715.

[8] Vikas Chaurasia, Saurabh Pal, Data mining techniques: to predict and resolve breast cancer survivability, International Journal of Computer Science and Mobile Computing IJCSMC 3 (1) (2014) 10–22.

[9] Ahmet Mert, Niyazi Kılıç, Erdem Bilgili, Aydin Akan, Breast cancer detection with reduced feature set, Comput. Math. Methods Med. 2015 (2015), https://doi.org/10.1155/2015/265138. Article ID 265138, 11 pages.

[10] Abdullah-Al Nahid, Yinan Kong, Involvement of machine learning for breast cancer image classification: a survey, Comput. Math. Methods Med. 2017 (2017), https://doi.org/10.1155/2017/3781951. Article ID 3781951, 29 pages.

[11] R. AlamKhan, N. Ahmad, N. Minallah, Classification and regression analysis of the prognostic breast cancer using generation optimizing algorithms, Int. J. Comput. Appl. 68 (25) (Apr.2013) 42–47, https://doi.org/10.5120/11754-7423.

[12] Ivan Lorencin, et al., Dataset size-based approach in design of artificial neural Network for breast cancer diagnosis, World Health 3 (2020) 13–19.

[13] D. Lavanya, K. Dr, Usha Rani. "Analysis of feature selection with classification: breast cancer datasets.", Indian Journal of Computer Science and Engineering (IJCSE) 2 (5) (2011) 756–763.

[14] Muawia A. Elsadig, Abdelrahman Altigani, Huwaida T. Elshoush, Breast cancer detection using machine learning approaches: a comparative study, Int. J. Electr. Comput. Eng. 13 (2023) 1, https://doi.org/10.11591/ijece.v13i1.pp736-745, 2088-8708.

[15] Adyasha Sahu, Pradeep Kumar Das, Sukadev Meher, Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms, Phys. Med. 114 (2023) 103138, https://doi.org/10.1016/j.ejmp.2023.103138.

[16] Jayesh George Melekoodappattu, Perumal Sankar Subbian, Automated breast cancer detection using hybrid extreme learning machine classifier, J. Ambient Intell. Hum. Comput. 14 (5) (2023) 5489–5498, https://doi.org/10.1007/s12652-020-02359-3.

[17] Surjeet Dalal, et al., A hybrid machine learning model for timely prediction of breast cancer, International Journal of Modeling, Simulation, and Scientific Computing 14 (4) (2023) 2341023, https://doi.org/10.1142/S1793962323410234.

[18] Rania R. Kadhim, Mohammed Y. Kamil, Comparison of machine learning models for breast cancer diagnosis, IAES Int. J. Artif. Intell. 12 (1) (2023) 415, https://doi.org/10.11591/ijai.v12.i1.pp415-421.

[19] Wisconsin Breast Cancer Dataset (WBCD).

[20] Breast Cancer Wisconsin (Original) Dataset.

[21] Rabah Guedrez, et al., Label encoding algorithm for MPLS segment routing. 2016 IEEE 15th International Symposium on Network Computing and Applications (NCA), IEEE, 2016, https://doi.org/10.1109/NCA.2016.7778603.

[22] Christophe Leys, et al., Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median, J. Exp. Soc. Psychol. 49 (4) (2013) 764–766, https://doi.org/10.1016/j.jesp.2013.03.013.

[23] Shahadat Uddin, et al., Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 19 (1) (2019) 1–16, https://doi.org/10.1186/s12911-019-1004-8.

[24] Michael P. LaValley, Logistic regression, Circulation 117 (18) (2008) 2395–2399, https://doi.org/10.1161/CIRCULATIONAHA.106.682658.

[25] Todd G. Nick, Kathleen M. Campbell, Logistic regression, Topics in biostatistics (2007) 273–301, https://doi.org/10.1007/978-1-59745-530-5_14.

[26] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, M. Alazab, "A comprehensive survey for intelligent spam email detection,", IEEE Access 7 (2019) 168261–168295, https://doi.org/10.1109/ACCESS.2019.2954791.

[27] S. Hegelich, Decision trees and random forests: machine learning techniques to classify rare events, Eur. Policy Anal. 2 (1) (2016) 98–120, https://doi.org/10.18278/epa.2.1.7.

[28] Shan Suthaharan, Shan Suthaharan, Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 2016, pp. 237–269, https://doi.org/10.1007/978-1-4899-7641-3_10.

[29] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.

[30] Alexandre Abraham, et al., Machine learning for neuroimaging with scikit-learn, Front. Neuroinf. (2014) 14, https://doi.org/10.3389/fninf.2014.00014.

[31] E. Eslami, A.K. Salman, Y. Choi, A. Sayeed, Y. Lops, A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks, Neural Comput. Appl. 6 (Mar) (2019), https://doi.org/10.1007/s00521-019-04287-6.

[32] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (Aug. 1997) 119–139, https://doi.org/10.1006/jcss.1997.1504.

[33] H. Xu, H. Yuan, An SVM-based AdaBoost cascade classifier for sonar image, IEEE Access 8 (2020) 115857–115864, https://doi.org/10.1109/ACCESS.2020.3004473.

[34] Cao Ying, et al., Advance and prospects of AdaBoost algorithm, Acta Autom. Sin. 39 (6) (2013) 745–758, https://doi.org/10.1016/S1874-1029(13)60052-X.

[35] L. Xu, J. Raitoharju, A. Iosifidis, M. Gabbouj, Saliency-based multilabel linear discriminant analysis, IEEE Trans. Cybern. 52 (10) (Oct. 2022) 10200–10213, https://doi.org/10.1109/TCYB.2021.3069338.

[36] Petros Xanthopoulos, et al., Linear discriminant analysis. Robust Data Mining, 2013, pp. 27–33, https://doi.org/10.1007/978-1-4419-9878-1_4.

[37] M. Fayad, D.C. Schmidt, Object-oriented application frameworks, Commun. ACM 40 (10) (1997) 32–38, https://doi.org/10.1145/262793.262798.

[38] U. Karthik Kumar, M.B. Sai Nikhil, K. Sumangali, Prediction of breast cancer using voting classifier technique. 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), IEEE, 2017, https://doi.org/10.1109/ICSTM.2017.8089135.

[39] Saloni Kumari, Deepika Kumar, Mamta Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, International Journal of Cognitive Computing in Engineering 2 (2021) 40–46, https://doi.org/10.1016/j.ijcce.2021.01.001.

[40] Al-Zadid Sultan Bin Habib, Tanpia Tasnim, An ensemble hard voting model for cardiovascular disease prediction. 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), IEEE, 2020, https://doi.org/10.1109/STI50764.2020.9350514.

[41] Sandeep Trivedi, Nikhil Patel, The determinants of AI adoption in healthcare: evidence from voting and stacking classifiers, ResearchBerg Review of Science and Technology 1 (1) (2021) 69–83, https://doi.org/10.17613/9ses-ah53.

[42] H. Zhao, L. Lv, Y. Ji, Improved redundant rule-based stochastic gradient algorithm for time-delayed models using lasso regression, IEEE Access 10 (2022) 3336–3342, https://doi.org/10.1109/ACCESS.2021.3138641.

[43] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retro-spective,", J. Roy. Stat. Soc. B Stat. Methodol. 73 (3) (Jun. 2011) 273–282, https://doi.org/10.1111/j.1467-9868.2011.00771.x.

[44] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. (2017) 4765–4774. https://dl.acm.org/doi/10.5555/3295222.3295230.

[45] H. Cramér, Mathematical Methods of Statistics, vol.43, Princeton Univ. Press, Princeton, NJ, USA, 1999, https://doi.org/10.4236/jmp.2016.79090.

[46] H.C. Kraemer, Kappa Coefficient, Wiley StatsRef: Statistics Reference Online, NewYork,NY,USA:Wiley, 2014, pp. 1–4, https://doi.org/10.1002/9781118445112.stat00365.pub2.

[47] Habib Dhahri, et al., Automated breast cancer diagnosis based on machine learning algorithms, Journal of healthcare engineering 2019 (2019), https://doi.org/10.1155/2019/4253641.

[48] L.K. Singh, M. Khanna, R. Singh, An enhanced soft-computing based strategy for efficient feature selection for timely breast cancer prediction: Wisconsin Diagnostic

Breast Cancer dataset case, Multimed. Tool. Appl. (2024) 1–66, https://doi.org/10.1007/s11042-024-18473-9.

[49] L.K. Singh, K. Shrivastava, An enhanced efficient approach for feature selection for chronic human disease prediction: a breast cancer study, Heliyon (2024) e26799, https://doi.org/10.1016/j.heliyon.2024.e26799.

[50] L.K. Singh, M. Khanna, A novel enhanced hybrid clinical decision support system for accurate breast cancer prediction, Measurement 221 (2023) 113525, https://doi.org/10.1016/j.measurement.2023.113525.