Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/121223

Utilizing LASSO for Breast Cancer Prediction: A Hyper Machine Learning Technique with Significant

Rawia Elarabi[1,], Najla Babiker[1], Awatef Balobaid[1] and Walaa M. Abd-Elhafiez[1,2]*

[1]College of Computer Science & Information Technology, Jazan University, Jazan, Kingdom of Saudi Arabia
[2]Computer Science Department, Faculty of Computers and Artificial Intelligence, Sohag University, Sohag, Egypt

Received: 30 Oct. 2023, Revised: 28 Nov. 2023, Accepted: 29 Nov. 2023.
Published online: 1 Dec. 2023.

Abstract: Cancer is a dangerous disease that greatly impacts people's lives, with breast cancer being the most common form in women. Detecting and predicting cancer accurately is crucial for a healthy life. This paper aims to achieve the highest accuracy in classifying breast cancer using various classifiers. Machine learning models and LASSO feature selection were employed, and the performance of different classifiers was compared using precision, accuracy, recall, F1 score, and ROC-AUC metrics. The results showed that the proposed model with SVM and LASSO achieved the highest accuracy.

Keywords: Breast cancer; Random Forest Classifier; Support Vector Machines; AdaBoost Classifier; classification; machine learning, KNN.

1 Introduction

Breast cancer can be another critical factor in the death of women. According to the World Health Organization site, around 685,000 people around the globe died in 2020 because of breast cancer, affecting 2.3 million women. Breast cancer was the most common public cancer on the globe at the end of 2020, having been identified in 7.8 million women in the prior five years. Malignant and benign tumors can be distinguished to diagnose this disease [1] [2] [3]. Tumors needed an accurate diagnosis to differentiate between malignant and benign tumors. Breast cancer has four stages that vary from stage 0 to stage 4. Stage 0 breast cancer is a normal cell that is not spreading outside the ducts into the surrounding breast cancer tissue. The cancer spreads to other body parts in stage 4 breast cancer, which is also termed metastatic breast cancer. As a result, early detection and diagnosis slow the disease's spread. Previously, doctors analyzed factors impacting breast cancer survival rates using basic software programs [4]. These traditional statistical techniques are not very versatile when discovering new variables or creating creative and integrative visualizations [5]. Since the manual conclusion of this disease requires long hours and specialists, computer-supported frameworks have been proposed and developed to minimize the time taken for analysis and decrease the spread of cancer. A disease can be detected and diagnosed via machine learning (ML) and deep learning (DL) techniques.

This research's intended contribution is the development of a professional healthcare system based on machine learning for use in the detection of breast cancer. We have chosen the most highly associated characteristics that significantly affect the predicted value of the target using the selection techniques Least Absolute Shrinkage and Selection Operator LASSO; this also helps to resolve machine learning overfitting and underfitting issues. In this study, a variety of supervised models, including Support Vector Machines, KNN, Random Forest classifiers, Gaussian Process classifiers, and others, are used. Cross-validation and grid search over a parameter grid are used to optimize the parameters of the estimator used to implement these algorithms. Using execution time, RMSE, and classifier performance assessment parameters, including classification accuracy, specificity, sensitivity, and the F1 Score of our model, individual results are presented for comparison.

The structure of this article is as follows: Section 2 presents the related work. Section 3 contains a description of the dataset and methods used in this study. Section 4 presents the results and discussion. The conclusion and potential future work are presented in Section 5.

2 Related Works

Liu Y. and Cheng W. [6] tested predictive power of features in different models to determine the importance of predictor variables; by producing upsampling method to improve the predictive performance of machine learning models. The author in ref. [7] proposed a method to classify breast cancer ultrasound images according to benign,

*Corresponding author e-mail: relarbi@jazanu.edu.sa

malignant, or normal status. Which different deep learning models have been utilized. Three models ResNet50, ResNeXt50 and VGG16 were evaluated for classification and achieved an accuracy of 85.4%, 85.83%, and 81.11%, respectively. In ref. [8] the author used ML methods for detecting as well as visualizing substantial prognostic indicators of breast cancer survival rate. In [9], Charan S. et al. [10] applied convolutional neural networks to mammograms to find unusual mammograms. On the MIAS dataset, this strategy was tested. A modified channel measure and preprocessing techniques were used to remove the noise component that was targeted to increase the precision of the overall model. NB, the REF network, and J48 are three data mining approaches recently employed to forecast breast cancer. A dataset of 683 samples from three continents— Latin America, Africa, and Asia—was used. Furthermore, the models were judged by a 10-fold cross-validation method with respective accuracy of 97.36%, 96.77%, and 93.94%. In comparison to the REF network as well as J48 models, the NB model is the most accurate predictor [11].

Data mining as well as ML approaches have been used to improve the efficiency as well as precision of diagnosis [12], [13]. But, increasing precision to decrease the RMSE error rate is thought to be difficult, as well as this research is being explored. This article analyzes the useful indices that can be used to create ML models and compares the various models for breast cancer prediction. So, in this work, we provide a multilayer model and assess how well-known and significant models perform. This work aims to achieve the best level of accuracy for the different classifiers utilized in this work. The precision of the various classifiers is also compared to determine which classifier is best for classifying breast cancer. The total precision and the time spent creating the framework are used to rate all classifiers and their types. When compared to various methods, the suggested methodology, combined with K-Nearest Neighbors (KNN), is proven to be accurate in breast cancer disease prediction, with a 10% increase in prediction accuracy. Compared to the suggested work, the presented technique performs better than other current techniques by accuracy, recall, precision, F1 score, and training loss. The results indicate that, with accuracy values of 98.25%, 97.37%, 97.37%, and 97.37%, respectively, the Gaussian Process model outperforms Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbors (KNN), and NB as the best predictor.

3 Materials and Methods

3.1 Materials/Dataset

This study's hybrid strategy improves the accuracy of categorization, prediction, and diagnosis in Wisconsin breast cancer datasets. The dataset used was in the UCI ML repository, and the CSV file [14]. The following is a summary of the dataset's characteristics. The dataset comprises 569 patterns (357 benign and 212 malignant), each of which has three classes (ID number, malignant and benign) and 33 columns of characteristics. A digitized image of a sufficient needle aspirate (FNA) for the breast mass is given record the features.

Furthermore, breast cancer diagnoses in Wisconsin use multi-ML approaches to classify malignant and benign tumors. This approach gathers the total values of malignant and benign tumors from an available dataset. This project aims to compare multiple classifiers and determine which one has the best accuracy. Also, evaluate the algorithms' efficiency and efficacy by sensitivity, accuracy, specificity, as well as precision. The UCI machine learning repository offered the Wisconsin Diagnostic Breast Cancer dataset [15]. This dataset comprises of 569 instances along with 33 features that contain 357 cases of benign breast cancer as well as 212 cases of malignant breast cancer shown in Figure 1. The first column is the ID number; the second is the diagnostic result column (malignant or benign), trailed by the standard deviation, mean, as well as mean of the worst measurements for ten traits. No values were lost. The features are not phenotypically independent. The relationship among features of the Wisconsin breast cancer diagnosis data set is illustrated in the feature heatmap in Figure 2, which was determined utilizing the Pearson correlation coefficient. The heat map scale signifies correlation degree between traits, with values ranging from -0.20 to 1.00, indicating the same degree of association. Figure 3 shows the dataset distribution.

3.1.1 Data Preprocessing

Data set preprocessing is a fundamental phase that could be taken before using ML methods. It entails searching within the data set for duplicate or missing data, which can be handled using real-time preprocessed or standardized in various ways. We employed label encoding for category features and the deletion of irrelevant features like ["id, "Unnamed: 32"] to minimize the disproportionate impact of disease categories.

3.1.2 Feature Selection Algorithms

ML requires using feature selection techniques to extract the optimal classification features, which will speed up classification and decrease execution time. To estimate the relief feature and the absolute most negligible shrinkage determination factor, we used the Least Absolute Shrinkage.

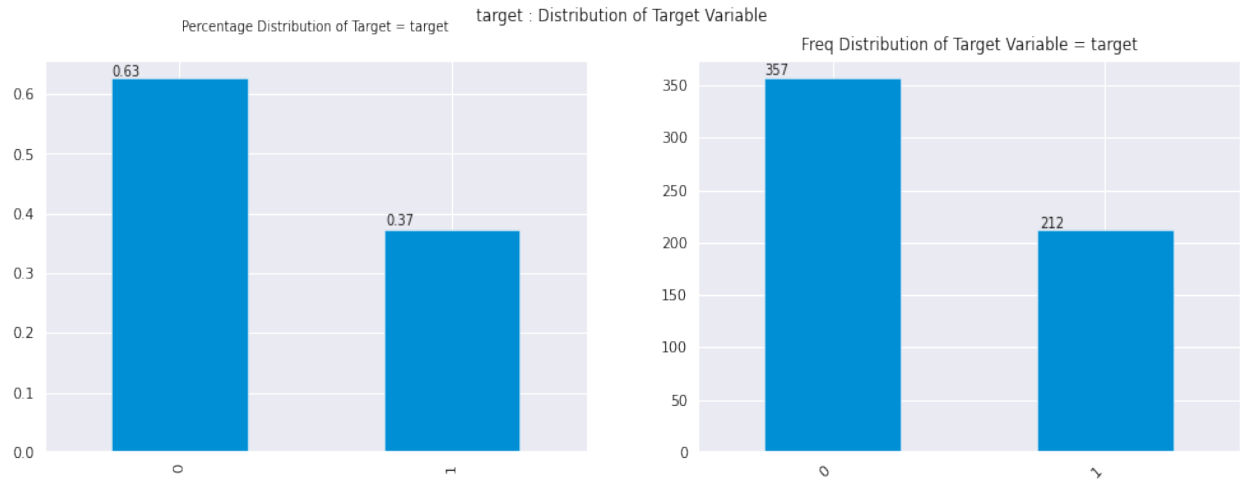

Fig. 1: Target distribution and distribution ratio

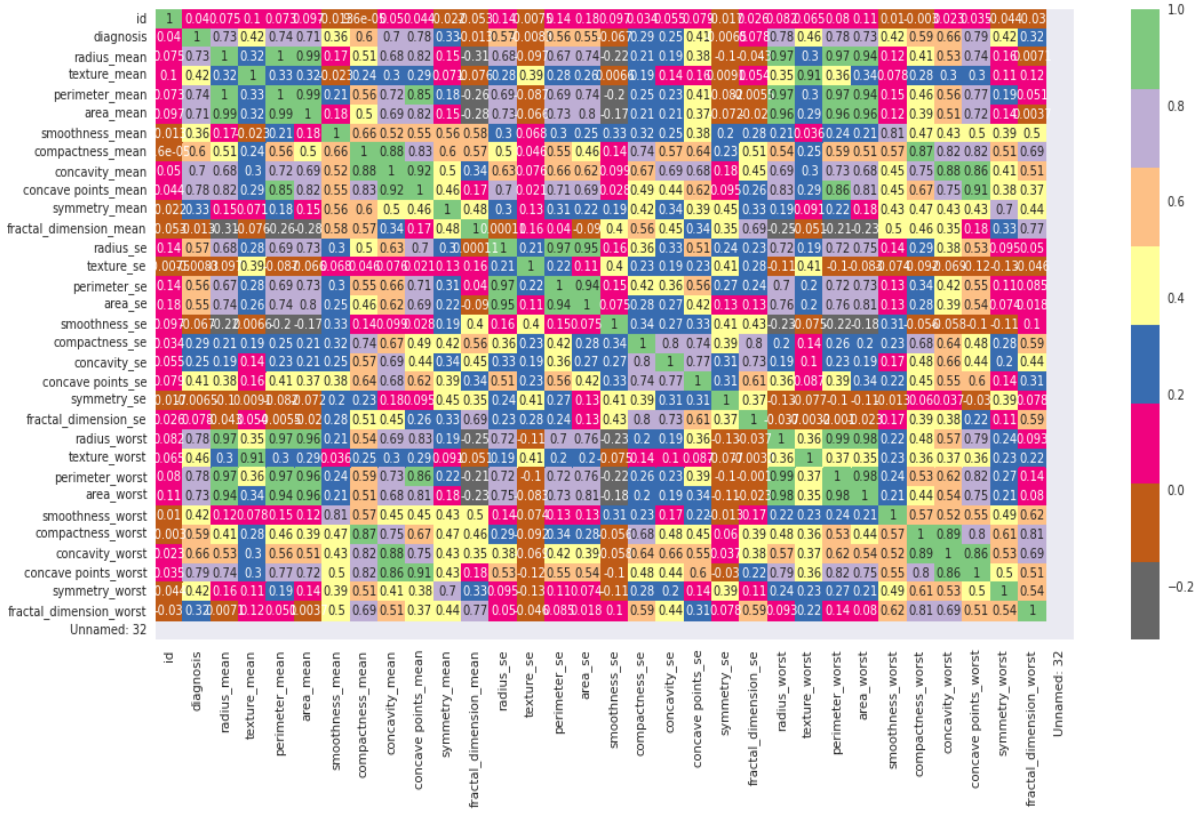

Fig. 2: Heatmap of features of the Wisconsin Breast Cancer Diagnostic Dataset derived using Pearson correlation coefficient.

3.1.3 Least Absolute Shrinkage and Selection Operator (LASSO)

The absolute value of the feature coefficient has been updated, which will help to select LASSO features. The zero-coefficient features are eliminated from the subset of features when some feature coefficient values reach zero. The LASSO executes brilliantly with low coefficient feature values, and the selected feature subsets will contain feature subsets with high coefficient values. A subset of the selected features might be used in LASSO, and some irrelevant features could be chosen [16]. We obtained the 13th most important value using LASSO, as shown in Figure 5. The LASSO method automatically eliminates features that are not needed and keeps only the most useful ones. When LASSO was used, the highest negative ordered score was achieved by perimeter worst (-0.782914), as shown in Table 1.

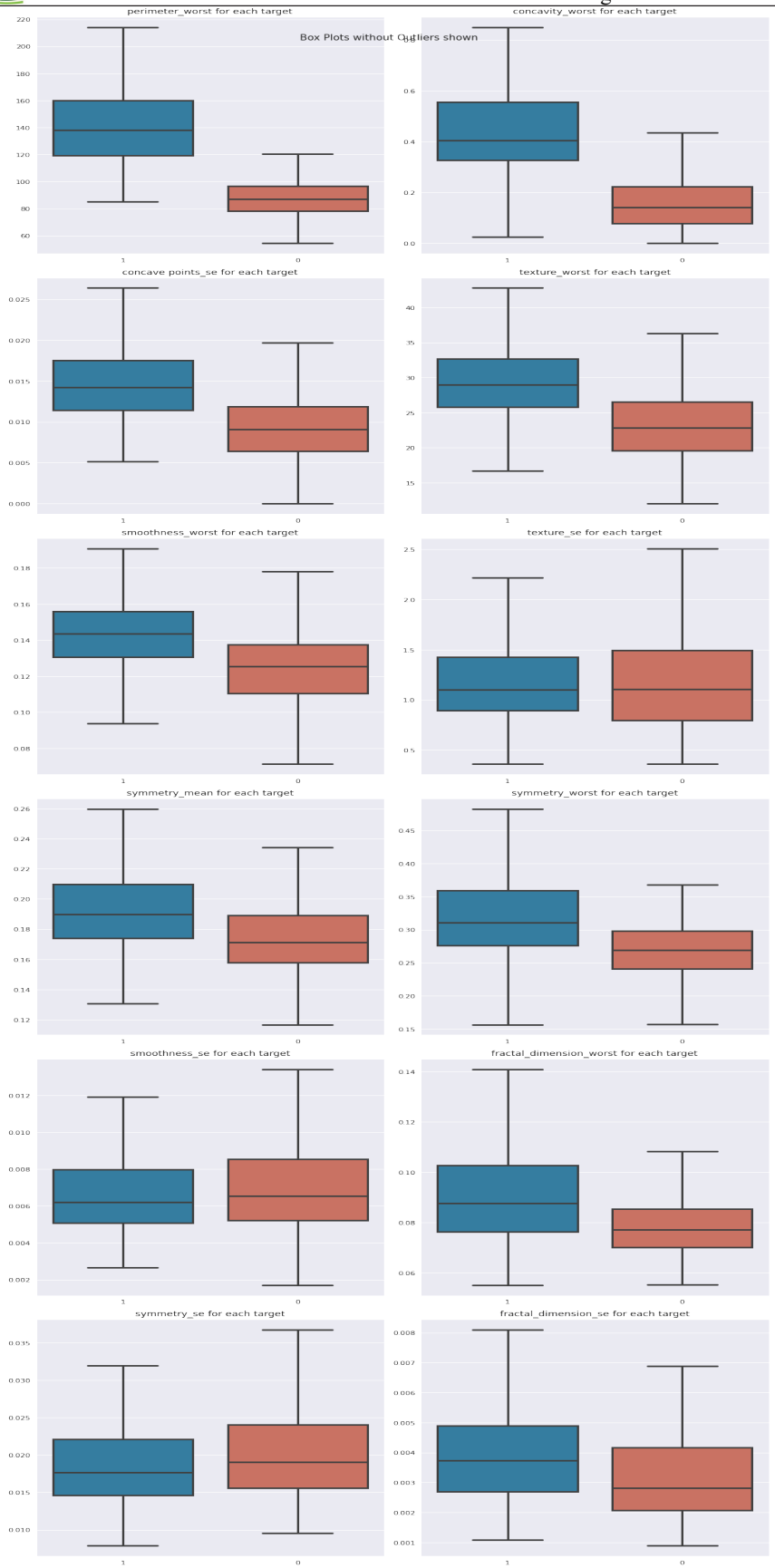

Fig. 3: Dataset distribution

3.2 Methods

In recent years, researchers started to use artificial intelligence applications, such as ML as well as DL, for medical treatment. ML achieved a high degree of predicting, diagnosing, and analyzing different types of diseases like diabetes, cancer, pneumonia, arthritis, heart disease, COVID-19 and many more disease [17]. Because it is a matter of life and death, making the correct diagnosis at the appropriate moment is the key to successful treatment in the medical world. The researcher must select the proper tools and algorithms that operate at a high level of accuracy. There are three different kinds of ML algorithms. Classification, regression, and forecasting all employ supervised learning. Semi-supervised learning, which uses labelled and unlabeled data, is analogous to supervised learning. Unsupervised learning, such as clustering and dimension reduction, is a type of ML where machines analyze the provided data by finding correlations and relationships without human instruction.

The optimum outcome is achieved by Reinforcement learning since it uses previous experiences to teach the machine and then adjusts its strategy in response to the circumstance. This allows the machine to be taught through trial and error. Various variables, including data size, diversity, quality accuracy, and training duration, influence the selection of the best ML algorithms. The Wisconsin Breast Cancer Dataset (WBCD) and other ML classifiers have been used to create an integrated intelligent model for predicting breast cancer in this study. There were ten different classification models utilized, together with DT, KNN, Artificial Neural Networks (ANN), SVM, Logistic Regression, NB, Gradient Boosting (GB), Random Forest, and AdaBoost Classifier. Feature selection/feature extraction techniques are the other contributions that touch all disciplines that require knowledge discovery from big data. An overview of the research materials and algorithms used in this work was provided in the following subsections.

3.2.1 Logistic Regression

The popular ML technique "logistic regression "is a predictive analysis approach based on probability and is utilized for classification and regression analysis problems. Binary and multi-linear function failure-class logistic regression are the two main types available. It is one of the most widely employed tools for discrete data analysis and applied statistics [18]. The bond among a dependent variable along with many independent variables has been evaluated using regression analysis via a logistic function to calculate the probability [19]. For a model having two predictor variables (x_1 and x_2) as well as one response variable Y, the relationship among them is given in the following form:

$$l = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (1)$$

Where p represents the event probability, and β_i is the model parameters. Once the values of the βi are fixed, the probability, Y=1 or Y=0, can be estimated.

3.2.2 Support Vector Machines (SVM)

This study uses a SVM classifier as a most commonly utilized approaches to detect breast cancer. It uses for resolving linear and non-nonlinear problems. One supervised ML technique that can use to learn classification, regression, or ranking functions is the SVM. The hyperplanes that defined the location of decision boundaries could be determined using SVM algorithm employing statistical learning theory and structural risk minimization. This approach produces the optimal separation of classes that can work with linear and nonlinear and support different kernel functions simultaneously. It returns a sole result to a problem. The SVM has many advantages, such as being active in high dimensional space and versatile because diverse kernel functions can be stated for the decision function [20] and find the better decision border, which demonstrates the most remarkable determination among the classes. SVM is divided into two types: linear and nonlinear SVM uses a single straight line to classify data into two classes; if data is nonlinearly separated, it cannot be classified by means of a straight line, so the classifier used is termed a nonlinear SVM classifier [21].

If we have N training data points $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_N, y_N)\}$, suppose a hyper-plane definite by (w, b), w is a weight vector, and b is a bias. New object classification is given by:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} w\, x + b\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i\, (x_i x) + b\right) \quad (2)$$

Where $x_i \in R\, d$ and $y \in \{+1, -1\}$, x_i is the training vectors, as well as α_i is the Lagrangian multiplier.

3.2.3 K-Nearest Neighbors (KNN)

The KNN classifier is the most focal ML strategy in classification [22]. KNN is a non-parametric sluggish learning approach employed for classification. This classifier groups the things using their 'k' closest neighbors, and it depends on the neighbors around the thing, not the vital data allocation. KNN is a simple and easy-to-implement algorithm under supervised ML types. It can be used in classification and predictive regression issues, mainly in classification problems. In KNN or Lazy algorithms, as it is named time, the concept of Majority voting is used, in which the 'K' parameter is

used to determine the number of nearest neighbors.

3.2.4 AdaBoost Classifier (AD)

Authors in [23] introduced the adaptive boosting approach (AdaBoost) that is used to maintain weights set over training data as well as adaptively modify them after each cycle of weak learning to create a group of poor learners. Training samples that the present weak learners incorrectly classify will have their weights increased, whereas training samples that are correctly classified will have their weights decreased.

3.2.5 Gradient Boosting (GB)

The ML-based Gradient Boosting technique has achieved notable success in various real-world applications [24]. This is adaptable to the operation's specific needs and can be taught concerning different loss functions.

3.2.6 Random Forest (RF)

The RF approach is a supervised classification approach. It depends on many self-learning DT, like the forest. These trees robotically state rules at each node rooted on a training dataset. RF pursues to minimize the heterogeneity of the two resulting subsets of the data shaped by the own rule, as shown in Figure 4. When compared to a single decision tree, the idea behind many decision trees is that many learners can reach a single solid and robust decision.

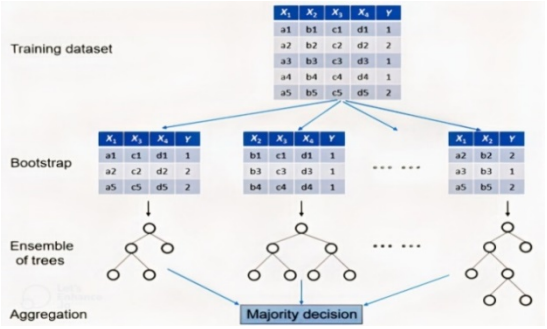

Fig. 4: Schematic structure of the RF approach

RF performs tree correlation to achieve better accuracy and less overfitting by using a large number of trees, with the error as well as converge into some generalized value, the function of the margin for sampling classifier $T_1(x)$, $T_2(x)$,...$T_I(x)$ with random training data got from vectors X as well as Y is represented as:

$$mg(X,Y) = av_b P(T_1(X) = Y) - \max_{j \neq Y} av_b(T_1(X) = j \quad (3)$$

Where j≠Y, P is the indicator function, avb the average, with $T_i(X)$=Y the classification result, Y is the class prediction and $T_i(X)$=j is the classification result with j [25].

3.2.7 Decision Tree (DT)

A model diagram with nodes and edges represents the DT classifier employed in this work. This classifier is recursively splitting the example space [26]. It is an analytical pattern that turns as a map of the object features along with its values [27]. It constantly splits all implicit result of the data into portions.

3.2.8 Naïve Bayes (NB)

The NB classifier plays a prominent part owing to its simplicity [28], efficiency along with tractability The NB implementation is significantly facilitated by the implicit guess of independent features conditioned on the class, which enables the decomposition of sample likelihood into a product of univariate marginal. A supervised learning algorithm is a NB classifier; it is a probabilistic classifier rooted on the Bayes theorem and is suitable for large datasets. The NB classifier plays a prominent part as it is simple [28], tractable, and efficient. The implicit statement of independent features conditioned on the class eases the NB operation knowingly as it lets the decomposition of sample likelihood into a univariate marginal product. Bayes' theorem's general formula is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4)$$

Where P(A) and P(B) denote, the probability of the hypothesis prior to and following the observation of the evidence, in contrast, p (A|B) and P(B|A) are the probabilities of hypothesis A on observed event B, or the likelihood that the evidence supports a hypothesis given its probability of being true.

3.2.9 Gaussian Process Classification

The nonparametric classification strategy is based on a Bayesian technique [29]. It expects a few prior distributions on the primary probability densities that ensure a few smoothness properties. At that point, the final classification is given that gives an excellent fit for the observed 1 data while at the same time assuring smoothness. Typically, this is accomplished by accounting for the smoothness of the prior while factoring in the observed classification of the training data classifier among all.

3.3 Methodology of the Proposed System

The proposed approach was created to classify cancer cells as benign (B) or malignant (M). For breast cancer diagnosis, various ML classification and prediction algorithms are applied. Two experiences were used to test the effectiveness of several machines learning predictive models on the WBCD's breast cancer diagnostic data: one used all features, while the other used LASSO's thirteen features. The LASSO feature selection algorithm was used to identify essential features. Figure 5 shows the 13 crucial features selected by LASSO that are used in our study. There are no missing or duplicate values. The data set is separated into two parts: a training as well as validation set with 80% for model training along with a validation set with 20%. The performance of classifiers with the attributes defined by these techniques is checked. All models of groups and traditional classifiers were implemented in the system. Model validation along with various performance evaluation metrics were obtained, and all algorithms implemented in the system were compared. Hyper ML such as the LR classifiers, AdaBoost classifier, K-NN, SVM, ANN, NB and DT were utilized in the system. The model's confirmation along with performance were calculated. This was formerly constructed into the framework. The workflow is depicted in Figure 6.

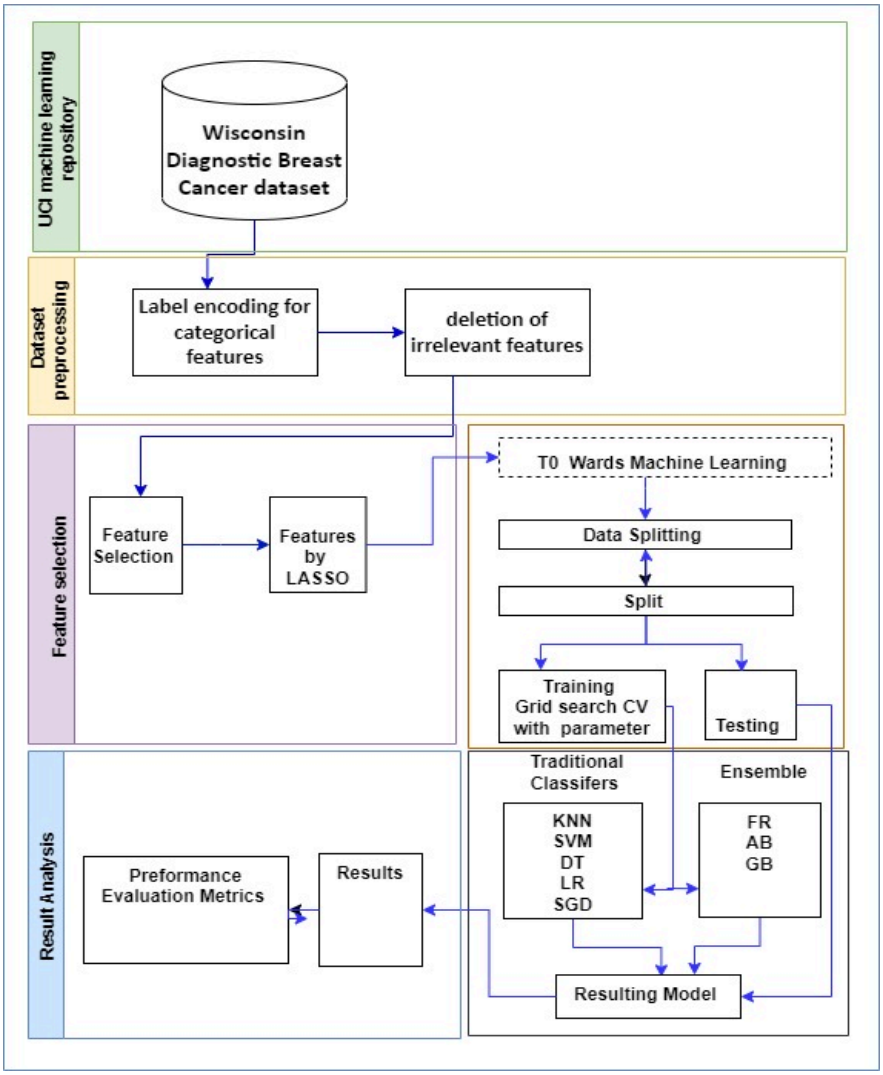

Fig. 5: Framework predicting breast cancer.

Table 1: Features selected by LASSO algorithms and their rankings.

Features name	Score
perimeter_worst	-0.78
concavity_worst	-0.66
concave points_se	-0.41
texture_worst	-0.46
smoothness_worst	-0.42
texture_se	0.0083
symmetry_mean	-0.33
symmetry_worst	-0.42
smoothness_se	0.067
fractal_dimension_se	-0.32
symmetry_se	0.0065
fractal_dimension_se	-0.078
target	1

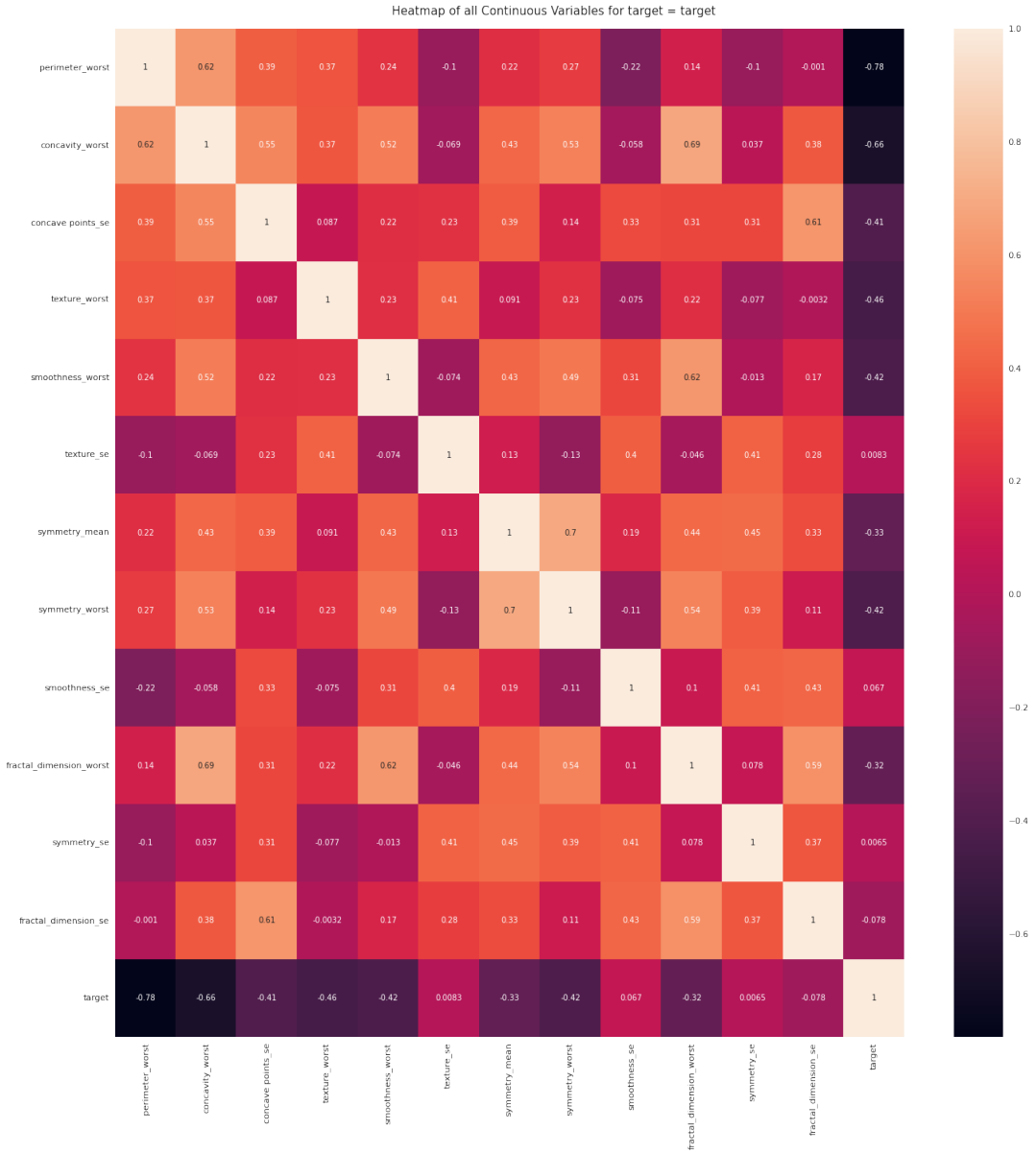

Fig. 6: Heatmap of 13 important features selected by LASSO derived using Pearson correlation coefficient.

4 Model Performance and Validation

The implemented model is written in a Collaboratory or "colab notebook", a product of Google Research. Colab is written in Python and uses simple libraries such as Pandas, NumPy, Seaborn, Pyplot, and Scikit-learn libraries. This research aims to achieve the highest level of accuracy for the various classifiers utilized in this work. In addition, the accuracy of diverse classifiers is compared to determine which classifier is finest for breast cancer classification. The accuracy signifies the accurate classification of normal subjects as well as breast cancer patients collectively and is depicted statistically by confusion matrix components. The generated confusion matrix provides an understanding of the ML approach's learning potential and classification accuracy. True positive (TP), true negative (TN), false positive (FP), as well as false negative (FN) are the primary components of the confusion matrix. The total accuracy and the time taken to develop the models are used to rate all classifiers and their types, as shown in the following equation:

$$\text{F1score} = \frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \times 100\%, \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\%, \tag{6}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \times 100\%, \tag{7}$$

- Root mean square error (RMSE): The RMSE stands for the standard deviation of the prediction errors. Prediction errors, also known as residuals, are a way to measure the difference between the finest fit line and the actual data points. The following is the error rate calculated by taking the square root of MSE [30][31]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{8}$$

- ROC-AUC: The optimistic receiver curves examine the ML classifiers' prediction ability. The figure of merit is represented by the positive of the false rate as well as the true rate within the classification results of a ML system for comparison using ROC analysis. AUC is a classifier's ROC measure. The higher the AUC value, the more influential the classifier's performance will be.

4.1 Comparison of various algorithms on the different features

Table 2 compares various classifiers concerning the accuracy shows that when using 13 features with LASSO, SVM achieves more than 99% accuracy compared to all features.

With accuracy, RMSE, precision, F1 score, and recall, the SVM outperformed the other ten classifiers, as shown in Table 3. When using SVM with the LASSO feature, we have 100% precision, a 98.95% F1 score, and a 97.92% recall. Logistic Regression and Random Forest Classifier were the second most crucial classifier, with a specificity of 100% precision, a 97.87% F1 score, and a 95.83%recall. In terms of precision, F1 score and recall, Naive Bayes performed the lowest out of the ten classifiers, with 91.67%, 91.67%, and 91.67%, respectively. Figures 7 illustrate the ROC-AUC values of classifiers for 13 features. Compared to other classifiers, the ROC-AUC values of LR, SVM, DT and RF classifiers were 99%, 95%, and 98%.

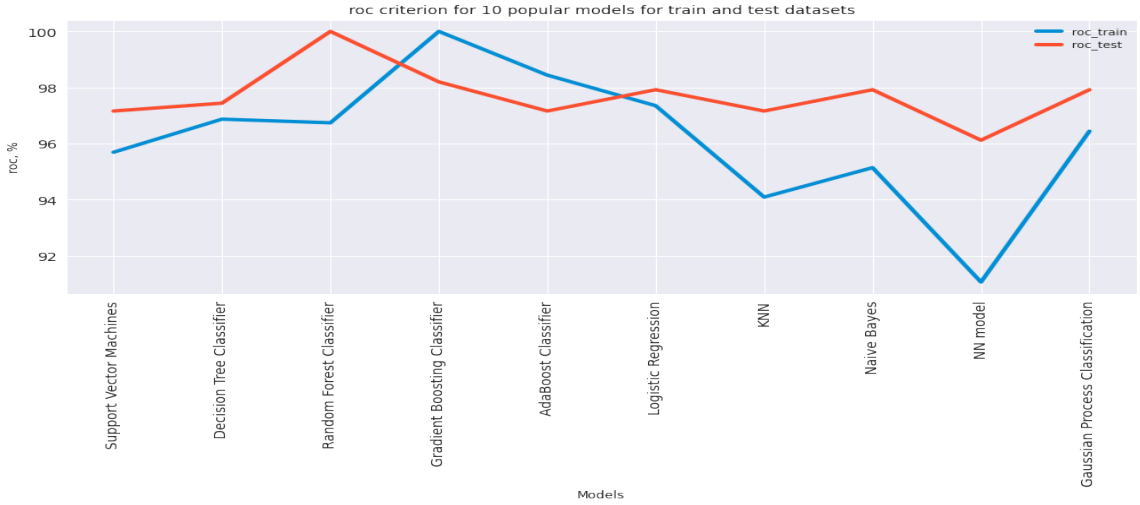

Fig. 7: Comparison between diverse methods based on ROC-AUC area which found feature by LASSO.

Table 2: Comparison between different models used features selected by lasso and all features selected based on accuracy.

Predictive Models	13 Feature with Lasso		All Features	
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
Support Vector Machines	99.12	99.12	98.90	95.61
KNN	98.68	98.25	96.26	97.37
Random Forest Classifier	96.26	98.25	96.48	92.98
Gaussian Process Classification	99.56	97.37	99.78	98.25
AdaBoost Classifier	98.68	97.37	98.68	94.74
Logistic Regression	97.36	96.49	98.90	97.37
Gradient Boosting Classifier	100	95.61	100.00	94.74
Decision Tree Classifier	98.46	95.61	99.34	97.37
Naive Bayes	97.8	94.74	94.07	97.37
NN model	93.19	92.98	99.78	98.25

In Table 4, Wisconsin Dataset is more extensively used to train learning models than MIAS Dataset. When comparing models trained on Wisconsin Dataset to models trained on MIAS Dataset, the Wisconsin Dataset models are considerably more accurate. Diverse models trained on the Wisconsin Dataset [31] [32] [33] [34] [35] [36] [37] have achieved nearly 97% accuracy. The findings show that if a suitable dataset is employed to train the models, ML methods can successfully help in identifying breast cancer. The trained models can be used to quantify the malignancy severity on a calibrated number scale. Compared to other classifiers, the accuracy of the proposed method was 99.12% on selected features by LASSO.

Table 3: Classification performance estimation of different classifiers on breast cancer dataset on features selected by LASSO and all features.

Predictive Models		RMSE	Precision	F1score	Recall
Support Vector Machines	All Features	20.94	94.87	93.67	95.44
	13 Feature with LASSO	9.37	100.0	98.95	97.92
KNN	All Features	16.22	97.37	94.83	96.77
	13 Feature with LASSO	22.94	95.65	93.62	94.32
Random Forest Classifier	All Features	26.49	92.31	90.0	92.82
	13 Feature with LASSO	13.25	100.0	97.87	95.83
Gaussian Process Classification	All Features	16.22	95.0	96.2	97.44
	13 Feature with LASSO	16.22	97.87	96.84	95.83
AdaBoost Classifier	All Features	22.94	90.24	92.5	94.87

	13 Feature with LASSO	18.73	97.83	95.74	93.75
Logistic Regression	All Features	16.22	97.37	96.1	96.77
	13 Feature with LASSO	13.25	100.0	97.87	95.83
Gradient Boosting Classifier	All Features	22.94	90.24	92.5	94.77
	13 Feature with LASSO	20.94	93.88	94.85	95.83
Decision Tree Classifier	All Features	16.22	95.0	97.38	97.44
	13 Feature with LASSO	20.94	97.78	94.62	91.67
Naive Bayes	All Features	16.22	95.0	96.2	97.44
	13 Feature with LASSO	26.49	91.67	91.67	91.67

Table 4: Diverse prediction models comparison

	Algorithm	Accuracy
Ref. [29]	NB and KNN	94.42%, 94.28%
Ref. [30]	NB	92%
Ref. [31]	Weighted NB classifier and Domain based weights	92%, 90%
Ref. [32]	SVM, Decision Tree and Bayesian classifier	96%
Ref. [33]	NB, RBF Network	97.36%, 96.77%
Ref. [34]	NB and KNN	96.19%, 97.51%
Ref. [35]	Fuzzy inference System	93%
Proposed model	RF with 13 features by LASSO	99.12%

5 Conclusion

Image analysis can be used to diagnose breast cancer with high accuracy. The training data set as well as features examined for examining breast cancer impose limitations on the accuracy given by ML algorithms. A reliable dataset for training ML models is the WBCD. This dataset is commonly used since it contains many essentially noise-free cases. All the tested ML techniques have resulted in more than 99.12 % prediction accuracy when utilizing thirteen features selected by the LASSO method. This result shows that ML techniques can efficiently forecast breast cancer. The promising results show that developing a dependable system for high-accuracy breast cancer diagnosis has many potentials. To discover models with improved performance and resilience, more hybrid and ensemble ML models and comparative analysis with DL models are proposed for future study.

Funding Statement: *The authors received no specific funding for this study.*

Conflicts of Interest: *The authors have no conflicts of interest.*

References

[1] C. C. Ponnuraja, B. Lakshmanan, V. Srinivasan, and B. K. Prasanth, (2017). Decision tree classification and model evaluation for breast cancer survivability: a data mining approach. *Biomed Pharmacol J*, **10**, 281–289, (2017).

[2] Hazra, R., Banerjee, M., Badia, L. (2020, November). Machine learning for breast cancer classification with ann and decision tree. In 2020 11th IEEE Annual Information Technology. *Electronics and Mobile Communication Conference (IEMCON)*, pp. 0522-0527. IEEE, (2020).

[3] Zorluoglu, G., & Agaoglu, M. (2017). Diagnosis of breast cancer using ensemble of data mining classification methods. International Journal of Oncology and Cancer Therapy, 2.

[4] N. Bhoo-Pathy et al. (2015) Trends in presentation, management and survival of patients with de novo metastatic breast cancer in a Southeast Asian setting. *Sci. Rep.*, **5**(1), 1–8, (2015).

[5] C. B. Pearce, S. R. Gunn, A. Ahmed, and C. D. Johnson, (2006). Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C-reactive protein. *Pancreatology*,

6(1-2), 123–131, (2006).

[6] Tran, T., Le, U., & Shi, Y. (2022). An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis. Plos one, 17(5), e0269135.

[7] Uysal, F., Köse, M.M. (2022). Classification of Breast Cancer Ultrasound Images with Deep Learning-Based Models. *Eng. Proc*. **31(8)**, (2022).

[8] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, (2019). Predicting factors for survival of breast cancer patients using machine learning techniques, *BMC Med. Inform. Decis. Mak.*, **19(1)**, 1–17, (2019).

[9] R. M. M. Al-Tam, (2020). Effective Multimedia Framework to Detect and Diagnose Breast Cancer Using Machine Learning, (2020).

[10] S. Charan, M. J. Khan, and K. Khurshid, (2018). Breast cancer detection in mammograms using convolutional neural network. in 2018 *International Conference on Computing, Mathematics and Engineering Technologies* (iCoMET), 1–5, (2018).

[11] V. Chaurasia, S. Pal, and B. B. Tiwari, (2018). Prediction of benign and malignant breast cancer using data mining techniques, *J. Algorithm. Comput. Technol.*, **12(2)**, 119–126, (2018).

[12] Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., Bhardwaj, A. (2020). Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications*, **140**, 112866, (2020).

[13] Chidambaram, S., Ganesh, S. S., Karthick, A., Jayagopal, P., Balachander, B., & Manoharan, S. (2022). Diagnosing breast cancer based on the adaptive neuro-fuzzy inference system. Computational and Mathematical Methods in Medicine, 2022.

[14] Y. Hailemariam, A. Yazdinejad, R. M. Parizi, G. Srivastava, and A. Dehghantanha, (2020). An empirical evaluation of AI deep explainable tools, in 2020 *IEEE Globecom Workshops* (GC Wkshps), 1–6, (2020).

[15] A. Frank and A. Asuncion, (2010). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California, *Sch. Inf. Comput. Sci.*, *213(2)*, 2010.

[16] McEligot, A. J., Poynor, V., Sharma, R., & Panangadan, A. (2020). Logistic LASSO regression for dietary intakes and breast cancer. *Nutrients*, **12(9)**, 2652, (2020).

[17] B. Krithiga, P. Sabari, I. Jayasri, and I. Anjali, (2021). Early detection of coronary heart disease by using naive bayes algorithm, in *Journal of Physics*: Conference Series, **1717(1)**, 12040, (2021).

[18] S. le Cessie and J. C. Van Houwelingen, (1994). Logistic regression for correlated binary data, *J. R. Stat. Soc. Ser. C* (Applied Stat., **43(1)**, 95–108, (1994).

[19] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, (2002). An introduction to logistic regression analysis and reporting, *J. Educ. Res.*, **96(1)**, 3–14, (2002).

[20] Mao, Q., Chen, Y., Duan, P., Zhang, B., Hong, Z., Wang, B. (2022). Privacy-preserving classification scheme based on support vector machine. *IEEE Systems Journal*, **16(4)**, 5906-5916, (2022).

[21] S. A. Mulay, P. R. Devale, and G. V Garje, (2010). Intrusion detection system using support vector machine and decision tree, *Int. J. Comput. Appl.*, **3(3)**, 40–43, (2010).

[22] X. Li, L. Wang, and E. Sung, (2005). A study of AdaBoost with SVM based weak learners, in Proceedings. 2005 *IEEE International Joint Conference on Neural Networks*, **1**, 196–201, (2005).

[23] C. Aroef, Y. Rivan, and Z. Rustam, (2020). Comparing random forest and support vector machines for breast cancer classification, *Telkomnika*, **18(2)**, 815–821, (2020).

[24] M. A. Mohammed, M. K. Abd Ghani, R. I. Hamed, and D. A. Ibrahim, (2017). Analysis of an electronic methods for nasopharyngeal carcinoma: Prevalence, diagnosis, challenges and technologies, *J. Comput. Sci.*, **21**, 241–254, (2017).

[25] R. L. De Mántaras, (1991). A distance-based attribute selection measure for decision tree induction, *Mach. Learn.*, **6(1)**, 81–92, (1991).

[26] D. J. Hand and K. Yu, (2001). Idiot's Bayes—not so stupid after all?, *Int. Stat. Rev.*, **69(3)**, 385–398, (2001).

[27] H. Nickisch and C. E. Rasmussen, (2008). Approximations for binary Gaussian process classification, *J. Mach.*

Learn. Res., 9(10), 2035–2078, (2008).

[28] Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, **15(14)**, 5481-5487, (2022).

[29] R. Elarabi, F. Alqahtani, A. Balobaid, H. Zain, and N. Babiker, (2021). COVID-19 Analysis and Predictions Evaluation for KSA Using Machine Learning, in 2021 *International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management* (ICSECS-ICOCSIM), 261–266, (2021).

[30] S. Kharya and S. Soni, (2016). Weighted naive bayes classifier: a predictive model for breast cancer detection, *Int. J. Comput. Appl.*, 133(9), 32–37, (2016).

[31] L. Hussain, W. Aziz, S. Saeed, S. Rathore, and M. Rafique, (2018) Automated breast cancer detection using machine learning techniques by extracting different feature extracting strategies, in 2018 *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering* (TrustCom/BigDataSE), 327–331, (2018).

[32] M. S. Yarabarla, L. K. Ravi, and A. Sivasangari, (2019). Breast cancer prediction via machine learning, in *2019 3rd International Conference on Trends in Electronics and Informatics* (ICOEI), 121–124, (2019).

[33] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances, *IEEE Trans. Inf. Technol. Biomed.*, **13(2)**, 236–251, (2009).

[34] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, (2018). Breast cancer classification using machine learning, in *2018 electric electronics, computer science, biomedical engineerings' meeting* (EBBT), 1–4, (2018).

[35] B. M. Gayathri and C. P. Sumathi, (2015). Mamdani fuzzy inference system for breast cancer risk detection, in *2015 IEEE International Conference on Computational Intelligence and Computing Research* (ICCIC), 1–6, (2015).