Breast Cancer Subtypes Classification with Hybrid Machine Learning Model

Suvobrata Sarkar[1] Kalyani Mali[2]

[1] Department of Computer Science and Engineering, Dr. B.C. Roy Engineering College, Durgapur, West Bengal, India
[2] Department of Computer Science and Engineering, University of Kalyani, Kalyani, West Bengal, India

Address for correspondence Suvobrata Sarkar, Department of Computer Science and Engineering, Dr. B.C. Roy Engineering College, Durgapur, West Bengal 713206, India
(e-mail: suvobrata.sarkar@gmail.com).

Methods Inf Med

Abstract

Background Breast cancer is the most prevailing heterogeneous disease among females characterized with distinct molecular subtypes and varied clinicopathological features. With the emergence of various artificial intelligence techniques especially machine learning, the breast cancer research has attained new heights in cancer detection and prognosis.

Objective Recent development in computer driven diagnostic system has enabled the clinicians to improve the accuracy in detecting various types of breast tumors. Our study is to develop a computer driven diagnostic system which will enable the clinicians to improve the accuracy in detecting various types of breast tumors.

Methods In this article, we proposed a breast cancer classification model based on the hybridization of machine learning approaches for classifying triple-negative breast cancer and non-triple negative breast cancer patients with clinicopathological features collected from multiple tertiary care hospitals/centers.

Results The results of genetic algorithm and support vector machine (GA-SVM) hybrid model was compared with classics feature selection SVM hybrid models like support vector machine-recursive feature elimination (SVM-RFE), LASSO-SVM, Grid-SVM, and linear SVM. The classification results obtained from GA-SVM hybrid model outperformed the other compared models when applied on two distinct hospital-based datasets of patients investigated with breast cancer in North West of African subcontinent. To validate the predictive model accuracy, 10-fold cross-validation method was applied on all models with the same multicentered datasets. The model performance was evaluated with well-known metrics like mean squared error, logarithmic loss, F1-score, area under the ROC curve, and the precision–recall curve.

Conclusion The hybrid machine learning model can be employed for breast cancer subtypes classification that could help the medical practitioners in better treatment planning and disease outcome.

Keywords
- triple-negative breast cancer
- clinicopathological parameters
- hybrid machine learning models
- classification
- genetic algorithm
- support vector machine

Introduction

Breast cancer, the world most prevalent cancer is caused due to the anomalous growth of breast cells leading to the evolution of malignant lump. According to GLOBOCAN 2020,[1] breast cancer has become the most prevalent diagnostic cancer ahead of lung cancer with an estimate of 2.3 million new cases and death count of 685,000 globally.[2]

received
December 31, 2021
accepted after revision
May 11, 2022

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0042-1751043.
ISSN 0026-1270.

Breast cancer prevalence rate has increased rapidly worldwide, however, early diagnosis and proper treatment outcome can decrease the mortality rate considerably. There are certain well-established methods for distinguishing breast cancer into distinct subtypes like histopathological classification on morphological features, expression profile of immunohistochemical (IHC) markers like estrogen receptor, progesterone receptor, and human epidermal growth factor receptor (HER-2). Each of these subtype Luminal A, Luminal B, HER2-positive, triple negative and normal-like have varied clinical outcome, response to therapy, and patients' survival rate. Luminal A and luminal B are responsive to endocrine therapy with better prognosis and survival rate than any other subtypes. HER2-positive breast cancer is sensitive to targeted therapies with poor prognosis and is likely to metastasis auxiliary lymph nodes. Triple-negative, the most aggressive basal-like subtype is identified by the absence of estrogen receptor, progesterone receptor, and HER-2. It is associated with high grade tumor, early development of recurrence within first 1 to 3 years of follow-up treatment, poor prognosis, and no precise therapeutic options. Thus, the diverse behavior of these subtypes compels us to perform classification for diagnosis and appropriate treatment outcome.

The capability of machine learning (ML) to detect unknown patterns and to establish relationship between them from a complex dataset can be utilized for predicting cancer types. Many research articles have started to come out recently which involves development of breast cancer prognosis evaluation models with ML approaches.[3–5] With digitization of medical records in the recent decade, medical data are now easily available to clinical practitioners based on which the medical decisions are revamped into a data driven machine. Clinicians usually collected the data from several sources as medical records and proper analysis of these heterogeneous data may yield diagnostic accuracy and prognosis evaluation. Some of the factors that affect the breast cancer prognosis include clinicopathological features (like age, tumor size, tumor grade, lymph node status) and molecular biology features (like HER-2 and hormonal receptors). Understanding the breast cancer subtypes with the predictive potency of ML can help the doctors in determining suitable treatment, thereby reducing the side effect of unnecessary treatment and financial loss of the patient party.[6] The human error caused by an inexperienced expert for cancer diagnosis can also be minimized with the aid of ML for automatic and accurate diagnostic prediction. Moreover, ML provides good results in clinical patient management.[7,8] Thus, the complex clinical implication of breast tumors and heterogeneous medical data motivated us to apply ML techniques for breast cancer classification.

The effect of artificial intelligence in analyzing breast cancer with different image modalities was summarized in Shahid Shah et al.[9] Saber et al[10] designed a deep learning model in combination with transfer learning for mammographic image feature extraction. Transfer learning techniques proved to be a powerful tool for automatic breast cancer diagnosis in terms of overall performance accuracy. Deep learning model was utilized in reducing the variability of BC subtype predictors by embedding prior knowledge into the loss function.[11] A nature inspired algorithm namely EHSSA (enhanced Salp Swarm algorithm)[12] was employed in microscopic image segmentation of breast cancer and the results showed significant accuracy in assisting physician for patients' rehabilitation. A new differential evolution algorithm inspired by slime mold foraging behavior leading to the development of superior BC image segmentation model was proposed in Liu et al[13] to achieve high convergence accuracy avoiding local optimum. Huang et al[14] developed a computer aided breast cancer diagnostic system with fruit fly optimization algorithm and SVM based on improved levy flight strategy. The performance of the model was tested with several benchmark functions and yield good classification accuracy. However, only 14 key features were assessed for that study. Hybrid model centered on adaptive SVM framework RF-CSCA-SVM for predicting students' choice of entrepreneurial intention was reported in Tu et al.[15] The hybrid model was tested with 23 classic benchmark functions and the results were compared with other SVM centric models. Hybrid model of genetic algorithm (GA) and SVM was applied in voice analysis of Parkinson's disease patients.[16] The analysis was carried with 31 patients, 23 PD patients, and eight healthy ones and 14 features were extracted mainly dependent on four main voice factors. Another study[17] combines competitive adaptive reweighted sampling and GA for variable feature selection for classifying Tegillarca granosa into contamination and non-contamination samples. Breast cancer subtype prediction is highly associated with the identification of most significant miRNA biomarkers. A two-phase ML ensemble feature selection technique was suggested in Sarkar et al[18] for breast cancer subtype prediction with specific miRNA biomarker followed by survival analysis. An absolute predictor was built by applying ML algorithms with limited number of probes for triple negative breast cancer (TNBC) subtype classification.[19]

In this paper, categorization of breast cancer into two groups-triple negative and non-TNBCs was performed based upon the hybridization of genetic algorithm and support vector machine (GA-SVM) model with the clinicopathological features collected from two North Western African countries tertiary care hospitals. SVM provides good classification results by finding optimal hyperplane with maximal margin width. The feature selection ability of GA had been utilized in classification problems for selecting subset of features from the feature pool with better fitness score that can participate in model training. The classification results obtained from GA-SVM hybrid model were compared with state-of-the-art feature selection hybrid models like support vector machine-recursive feature elimination (SVM-RFE), LASSO-SVM, Grid-SVM, and linear SVM. It had been found that the classification accuracy obtained from GA-SVM hybrid model outperformed the other compared models. To validate the predictive model accuracy results, 10-fold cross-validation method was applied on all models with the same multicentered datasets. The model performance was evaluated with well-known metrics like mean squared error, logarithmic loss, F1-score, area under ROC curves, and the precision–recall curve. Thus, hybrid ML model was employed in classifying breast cancer subtypes that could assist the doctors in clinical decision-making and treatment outcome. The paper is organized as follows: Section 2 highlights the datasets utilized for analysis, classic ML feature selection methods, and the proposed model. Section 3 describes about the performance of the proposed GA-SVM model, comparison with classic feature selection methods, and the statistical analysis to show the dependency of clinicopathological parameters in categorizing TNBC/non-TNBC cases. Section 4 deals with the discussion and lastly Section 5 concludes the paper.

Methods

In 2018, the breast cancer prevalence rate among African women was 26.3 per 100,000 women and the occurrence rate of breast cancer in Caucasian females was found to be 22.8 per 100,000 women.[20] TNBC, the most destructive subtype occurs predominantly in Black and Africa–American women.[21–24] TNBC can be investigated as a combination of chemotherapy, surgery, and radiation therapy although there exist no approved targeted therapies by FDA.[25] These necessities the classification of TNBC with non-TNBC groups of breast cancer. The data were accumulated from two retrospective studies of African countries available at Biostudies. Biostudies database is an EMBL-EBI facility illustrating the biological studies and linking the data from these studies to another database at EMBL-EBI.[26] Moreover, the authors can submit any supplementary information and can link it with their publication in specific file format which can be accessible from Biostudies.

Data Sets

A retrospective study was included comprising of 905 patients treated with breast cancer. This study was conducted at National Institute of Oncology, Rahat, Morocco in 2009 and was followed up till 2014.[27] The authors have supplied anonymous patient dataset as a supplementary material in excel file format freely available at Biostudies. The data were gathered from each patient's medical record and information about their clinical, pathological, and therapeutic characteristics were reviewed. A total of 405 cases were debarred due to incomplete data, foreign and male patients. Left-over 500 breast cancer cases were partite into two molecular subtypes: 85 TNBC and 415 non-TNBC cases, respectively.[28] Further, the clinicopathological features, the pathological data of SBR grading, treatment, and prognosis of TNBC patients were investigated. Another study[29] of 251 breast cancer patients diagnosed at Lagos Teaching University & Hospital, Nigeria has been considered. Female patients above 18 years of age were investigated between July 2017 and July 2019. The study focused on clinical, pathological, and socio-demographical information. One-hundred and nineteen (47.4%) TNBC cases and 43.2% non-TNBC cases were evaluated based on statistical analysis. The patient's dataset is easily available on Biostudies as a supplementary Material.[30]

Classic ML Feature Selection Models

Feature selection process involves the identification of a subset of relevant features from the feature pool thereby reducing the complexity of the ML model and allowing the model to train faster irrespective of the choice of ML algorithms. Further, the removal of less important features that does not contribute to the prediction of targeted variables can also reduce the overfitting problem and improves the generalization skill of the model.

Recursive feature elimination (RFE),[31] a wrapper type feature selection technique is used to train the model with all possible combination of features in an iterative manner. At first, the algorithm works with the available features in the training set and assigns the ranking weights to all the features. The features with smaller weights are removed with backward elimination technique and the model accuracy is calculated with the latest set of features. This process iterates till the optimal combination of features is attained or when the model performance decreases. Thus, the model is created with the best possible subset of features that produces the highest classification performance. SVM-RFE is SVM-based feature selection methods which utilize the classification skill of SVM at the core of the model and RFE is wrapped around it to produce the best possible combination of features achieving the optimal model performance. LASSO (least absolute shrinkage and selection operator), a linear regression extension with the addition of L1 penalty in the loss function, shrinkage the coefficients of input variables that does not contribute in the prediction of targeted variables. This regularization technique eliminates the features whose coefficient values are shrunk to zero thereby providing a flavor of automatic feature selection. It is useful when fitted on a scaled dataset with high variance in training and test cases. With the emergence of SVM as a powerful breast cancer diagnostic classifier, popular Lasso-SVM hybrid strategy-based feature selection model was entailed in this study for comparison. The linear SVM was originally suggested by Vapnik in 1963.[32] This algorithm creates a decision boundary that can segregate the data into distinct classes such that the misclassification errors can be minimized. Choosing the optimal decision boundary or hyperplane involves maximizing the distance from all nearest data points of the partitioned classes. These nearest data points from the either side of optimal hyperplane are called support vectors and the hypothetical lines that pass through these support vectors are called margin. Thus, the optimal hyperplane can be obtained by maximizing the margin width. In case of non-linear data points, an optimal kernel function is being selected to transform the non-linear data into a high dimension to make the data linearly separable. The main advantage of SVM is its ability to handle linear as well as nonlinear data efficiently using different kernel functions. ML models like neural network and SVM have many parameters that do not get trained during the training stage but can control the behavior of the model. They have to be configured upfront before the model is being trained. Such types of parameters are called hyperparameters. Thus, finding the optimal values of hyperparameters is a challenging task. Grid search, an exhaustive search technique is generally applied to tune the values of hyperparameters before the training phase of such ML models. The model is created with the set of hyperparameter values and the classification accuracy is noted. Finally, the optimum set of hyperparameters values with highest model accuracy is considered for the training phase. Here, grid-SVM hybrid model has been employed for comparison with the proposed model.

The Proposed Model

Genetic algorithm,[33–36] a heuristic search and optimization technique depends upon the concept of natural evolution proposed by Darwin. This algorithm is capable for producing near optimal solution of an objective function for handling optimization problems in ML. The process of GA starts with encoding the parameters of an individual as strings called chromosomes. Collection of all such chromosomes is called population. At first, a random population is selected and each of the individual is assigned with a fitness score which indicates the degree of goodness of an individual in the selected population. The fitness function evaluates the fitness score of all individuals which denotes the probability of being chosen as the fittest individual for reproduction. Based on the natural selection phenomenon, few individuals with higher fitness value are selected for the mating pool. Bio inspired operators like crossover and mutation are being applied to these individuals to generate the next generation of off springs. The process of selection, crossover, and mutation iterates till a stopping criterion is achieved. The algorithm terminates with the execution of maximum iterations or when the population converges, i.e., the last population is unable to produce new off springs substantially different from the previous population. Thus, the feature selection ability of GA can be utilized in classification problems for selecting the features subset with better fitness score that can participate in model training. In this study, GA has been applied on SVM model for selecting potential features involved in model training. The details about SVM had been discussed in the previous section. This hybrid model was capable of classifying different variant of breast cancer improving the classification accuracy and other model performance metrics. But it is necessary to define chromosome, crossover rate, mutation rate, and number of iterations before applying GA to the proposed model. The entire implementation was performed in python version 3.7.4 and its associated statistical packages.

The genetic selection procedures are as follows:

1. estimator = SVM
2. cv = 10
3. verbose = 1
4. scoring = ”accuracy”
5. max_features = 5
6. n_population = 50
7. crossover_proba = 0.5
8. mutation_proba = 0.2
9. n_generations = 40
10. crossover_independent_proba = 0.5
11. mutation_independent_proba = 0.05
12. tournament_size = 3
13. n_gen_no_change = 10

where cv = cross validation, Scoring = “Accuracy” means that the score is associated with every individual of the initial population as the targeted metrics, verbose = data logging information, max_features = maximum size of each feature subset selected for the initial population, n_population = initial population generated randomly from the feature sets, crossover_proba = the probability of having crossover among the parents to form child in passing the genetic material from one generation to next generation, mutation_proba = the probability that the mutation will happen within the features randomly, n_generations = number of generation to repeat for genetic selection, crossover_independent_proba = the chance of a particular feature to crossover and selected as a child in the next generation, mutation_independent_proba = the chance of every feature in the feature set to mutate at each generation, tournament_size = the size of the fittest individuals selected for the tournament based on scoring metrics, n_gen_no_change = number of generations to iterate till the population converges.

The datasets were imported as Pandas’ data frame, an open-source python library. Pandas’ data frame was similar to feature matrix with rows representing the patient’s anonymous identity and the columns denotes the clinicopathological features of the respective patients. The class label was represented as the targeted array. Python has in-built ML library Scikit-learn,[37] also supports python scientific and numerical libraries for data analysis. Data pre-processing was performed with the identification of missing values with Simple Imputer function and replacing the same with the most frequent feature value. Data standardization function standard scaler rescaled the data with features mean value zero and unit variance. The datasets were divided into two subsets: training and test dataset. The training dataset was utilized to train the model with known examples and test dataset estimated the generalization of the model with unseen examples. The train_test_split function of sklearn package was used to randomly split the datasets into training and test dataset in the proportion of 7:3. Genetic algorithm was applied to mimic the natural selection procedure for finding the best possible value of radial basis SVM function. Radial Basis is preferable over other kernel functions as it can store support vectors only during the model training instead of the entire dataset. Ten-fold cross validation was performed to test the performance of hybrid model in classifying the unseen data and to reduce overfitting. The chromosomes represent the clinicopathological features encoded as strings. The maximum number of features selected for initial population was 5. The scoring parameter associated with each individual in the population with target metric was accuracy. The data log verbose was 1. Initial population started with 50 chromosomes. The probabilities of getting crossover and mutation were fixed with 0.5 and 0.2 values, respectively. The independent probability of crossover and mutation for each attribute was assigned with 0.5 and 0.05

values. The tournament size was fixed with value 3 which means that size of the fittest individuals selected for the tournament was 3 and will be passed to next generation for mating. Number of generations for termination was considered as 10 after the last population was unable to produce better off springs than the previous populations.

The steps of the proposed GA-SVM model implemented in python are:

Step 1: Load the dataset as data frame with m = patients' identity and n = clinical and pathological parameters ($m \times n$ feature matrix).
Step 2: Missing value identification and data standardization with the Simple Imputer and Standard Scaler functions.
Step 3: Class labels as ($m \times 1$) targeted array.
Step 4: train_test_split () function for training and test datasets in the ratio of 7:3.
Step 5: Choose the best kernel =: ["rbf," "sigmoid," "linear"], C and gamma values.
Step 6: Genetic selection on estimator = SVC.
Step 7: Display feature selector support_.
Step 8: Calculate scoring = "accuracy."

Ethical Consideration

As stated earlier, the original datasets were available freely from Biostudies as a Supplementary Material. The informed consent had been taken from the corresponding authors of reference [27,29] for performing this secondary analysis. Since this study does not involve human subject participation directly as a result the ethical clearance from institutional review board was not imperative.

Results

Hospital-based datasets of patients consisting of clinicopathological features for breast cancer investigation of two North West African countries, Morocco and Nigeria, were analyzed in this article. A total of 905 patients were admitted at National Institute of Oncology, Rabat, Morocco for breast cancer treatment. Incomplete medical records, foreign and male patients were left-out and finally 500 cases were considered for analysis. Another dataset of 251 cases with breast cancer diagnosed at Lagos Teaching University, Nigeria was examined for classification.

Performance Evaluation of GA-SVM Model

The performance of GA-SVM model was assessed with several performance evaluation metrics like confusion matrix, classification accuracy, area under the ROC curve, mean square error, logarithmic error, precision–recall curve, and learning curve. Confusion matrix is a N × N matrix representing the tabular summary of the actual outcome versus the predicted outcome made by the classifier. N represents two classification classes: TNBC and non-TNBC. For simplicity, TNBC is denoted by 1 and non-TNBC by 0. The Lagos University, Nigeria breast cancer dataset and National Institute of oncology, Rabat, Morocco breast cancer dataset were abbreviated as dataset 1 and dataset 2, respectively. The classification report of GA-SVM model on dataset 1 and dataset 2 was shown in ►Fig. 1. It exhibits the values of precision, recall, F1-score, support and accuracy of dataset 1 and 2 subsequently. Higher values of these performance metrics on both the datasets justify that the hybrid model classifies TNBC patients and non-TNBC patients almost accurately.

Area under the ROC curve (AUC) estimates the capability of a classifier to distinguish between the classes at varying probabilistic threshold settings. AUC is plotted graphically with false positive rate in the x-axis and true positive rate in the y-axis. When the value of AUC is 1, the classifier is able to discriminate all the classes correctly and with AUC = 0, the classify will assign a random or a particular class in every case. When this condition arises, the classifier is called as no-skill classifier and is represented at (0.5, 0.5) in the AUC curve. No skill models are plotted with diagonal line from bottom left to top of the right for every threshold. Model is said to be perfect skill when it lies between (0, 1) and plotted with a line from bottom left to the top left through the top of the top right of the curve.[38] ►Fig. 2 depicted the AUC of the hybrid model on two North West African datasets. The AUC of dataset1 attain the sensitivity = 1 with 0.1 value of false-positive rate and covers the largest area before coinciding with no-skill line. The curve of dataset 2 increase steadily and finally attained sensitivity = 1 to reach (1, 1).

Mean square error is the popular loss function which calculates the sum of squared difference between the models predicted value and the actual value divided by the total number of patients considered as test cases in the dataset.

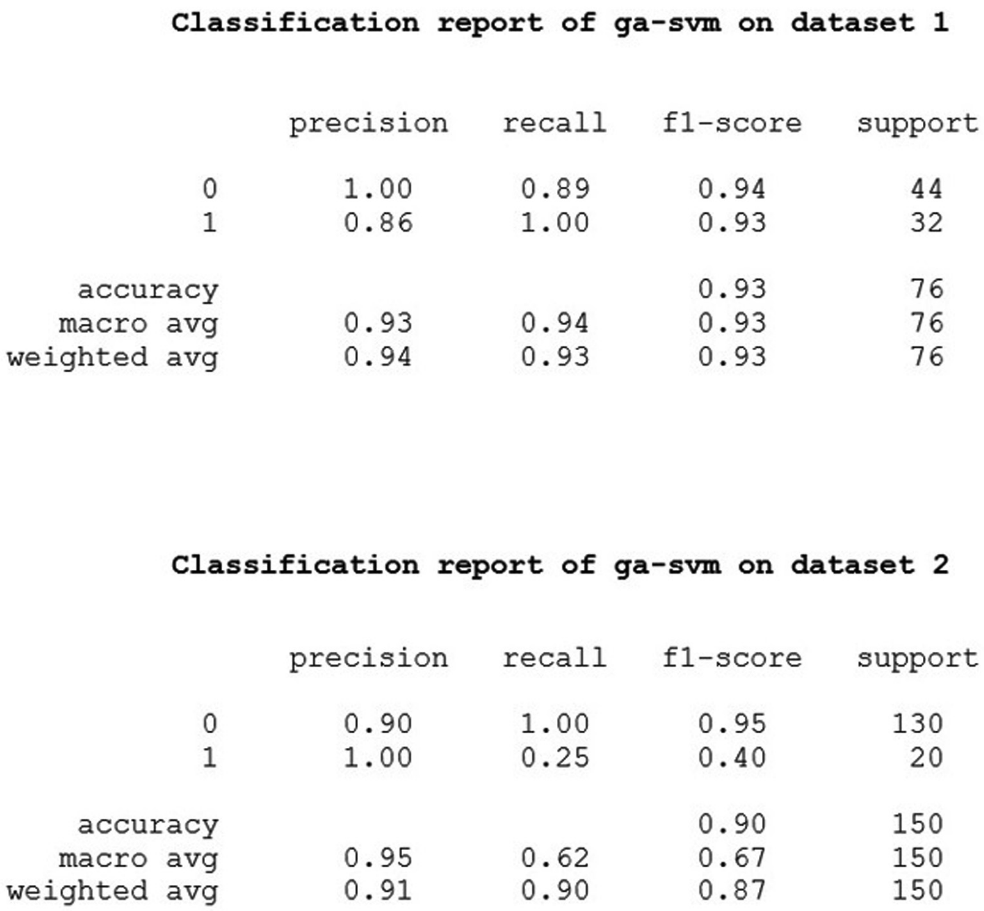
$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \quad (1)$$

where y_i stands for the model predicted value, $\hat{y}_i$ represents the actual value and N denotes the total number of patients as test cases in the respective North West African datasets. Lower the value of MSE, the closest is the predicted value to the actual value. MSE cannot be negative due to the error squares. The mean square error (MSE) values of dataset 1 and dataset 2 in GA-SVM model were calculated as 0.06 and 0.1 indicating the lower MSE values and higher classification strength of the classifier.

The logarithmic loss (Log loss), a classification metric applied in the prediction process of ML depends on the concept of probability. Lower value of log loss denotes the proximity of the prediction probability with respect to the actual value. Higher value indicates more deviation of predicted probability from the true or actual value. Log loss is calculated as the negative average of the logarithmic corrected predicted probabilities of each patient.

$$Logloss_i = -[y_i \log p_i + (1-y_i)\log(1-p_i)] \quad (2)$$

where i denotes a particular patient, y_i stands for actual value, p_i is the predicted probability, and *log* denotes the logarithmic value of a number. Smaller value of Log loss indicates better predictive results. The log loss values of the hybrid model on dataset 1 and dataset 2 were 0.55 and 0.31, respectively.

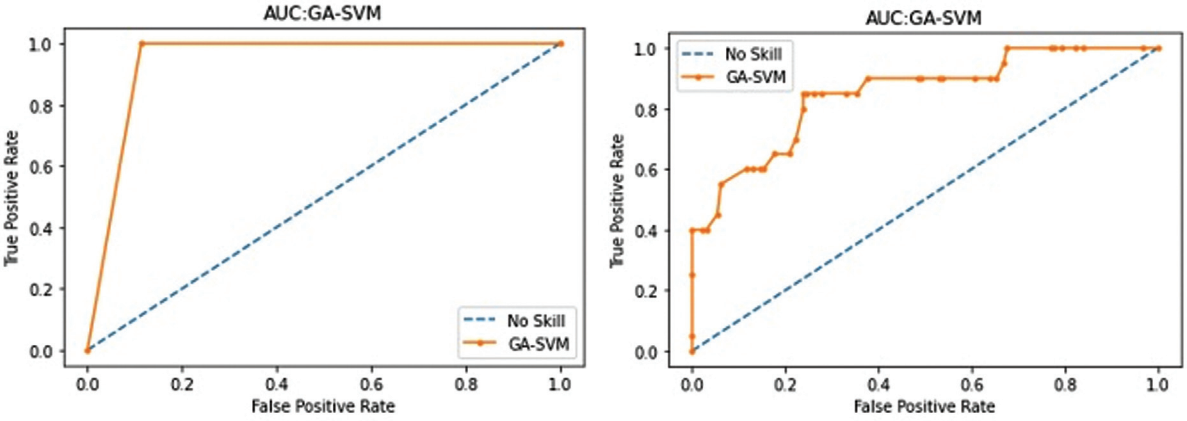


```
Classification report of ga-svm on dataset 1

              precision    recall  f1-score   support

           0       1.00      0.89      0.94        44
           1       0.86      1.00      0.93        32

    accuracy                           0.93        76
   macro avg       0.93      0.94      0.93        76
weighted avg       0.94      0.93      0.93        76
```

```
Classification report of ga-svm on dataset 2

              precision    recall  f1-score   support

           0       0.90      1.00      0.95       130
           1       1.00      0.25      0.40        20

    accuracy                           0.90       150
   macro avg       0.95      0.62      0.67       150
weighted avg       0.91      0.90      0.87       150
```

Fig. 1 Classification report of GA-SVM proposed model on both the datasets. 0 stands for non-TNBC cases and 1 represents TNBC cases. TNBC, triple negative breast cancer.

Precision–recall curve assess the adjustment between the true positive rate (recall) and positive predictive value (precision) at varying probabilistic threshold values. Precision recall curves are more informative and best suited for imbalance datasets as compared with ROC curves which are appropriate for balance datasets. Precision recall curve is plotted with recall in the x-axis and precision in the y-axis at every threshold points. It often follows a zigzag path moving up and down when plotted. Usually, precision recall curve of no overlapping results signifies better performance level as compared with the one near the baseline. To assess the efficiency of GA-SVM model, two heterogeneous datasets of North West



Dataset 1

Dataset 2

Fig. 2 Area under the curve (AUC) of GA-SVM hybrid model on both datasets.

Africa predominantly treated with breast cancer precision recall curve are plotted in ►Fig. 3. The precision recall curve of dataset 1 shows no overlapping region above the baseline and it implies better performance and dataset 2 curve follows zigzag path ultimately reaching near the baseline.

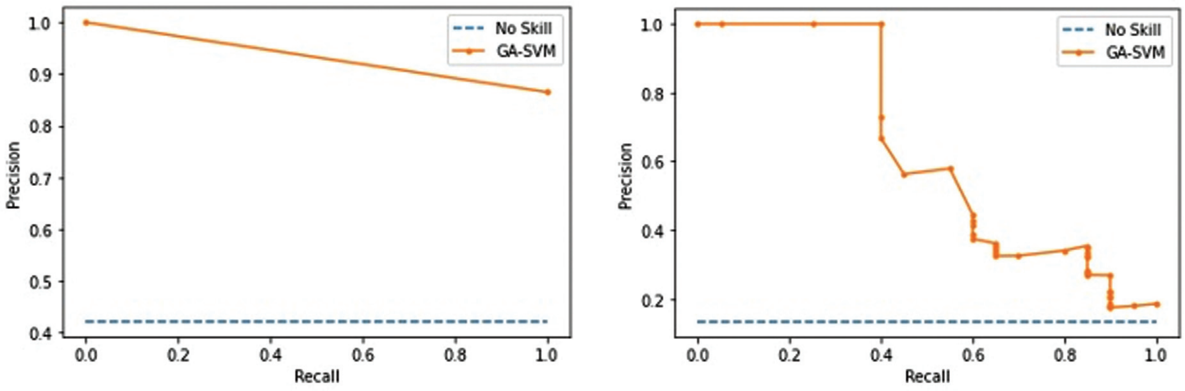
Convergence of ML algorithms is measured empirically using the learning curve. It usually refers to a stable point attained by the algorithm at the end of optimization beyond which further improvement or changes cannot be expected. The learning curve plots the learning performance of the hybrid model versus experience or time. The learning curves for two datasets are available in ►Figs. 4 and 5 with training set size in the x-axis and accuracy score in the y-axis, respectively. It shows how the training score and cross validation score changes on incremental addition of training dataset. It provides a conception of how efficient the model is in learning and generalizing the unseen data. Learning curves are primarily used to diagnose overfit, underfit, and well-fit models and evaluate the exact amount of training data best suited for the model with variance-bias trade-off. For dataset 1, the training score was high initially but as the training sample size increases, it decreases. The cross-validation score was low at first but increases with the addition of sample size. However, the training score for dataset 2 decreases at the beginning but improves steadily with training size above 350 while the cross-validation score remains almost stable around 0.83 accuracy score. The scalability of the model was assessed with the time required by the model to fit the estimator with the training dataset. The scalability is plotted with training dataset on the x-axis and the fit_times on the y-axis. fit_times is the time taken by the model to fit the estimator with the training set for every cross validation. The curve on both the datasets increases as the training samples are added subsequently and reach peak with fit_times = 0.35. The model performance was also analyzed with fit_times versus the test score. The model performance on dataset 1 attained the stability with test score above 0.50 and fit_times = 0.35 whereas the performance on dataset 2 remained uniform with the test score around 0.83 and the same fit_times.

Comparison with Other Models

For comparative analysis, the developed GA-SVM hybrid model performance was compared with the existing state of the art feature selection techniques for breast cancer diagnosis and classification like SVM-RFE,[39] LASSO-SVM,[40,41] and Grid-SVM.[42] Recently, SVM[43,44] evolved as a superior model for breast cancer classification and had also been entailed for comparisons. Basic details about these models had been discussed in the earlier section. The classification accuracy results of the classic models with GA-SVM hybrid model on two North West African datasets were depicted in ►Table 1. The classification accuracies of dataset 1 and dataset 2 were 93.4 and 90% which outperformed the classification accuracy of all other classic models. The well-known evaluation metrics like MSE, Log loss, AUC, F1-score were evaluated for the respective models on two datasets and the compared results are demonstrated in ►Table 2. Higher values of AUC and F1-score and smaller values of MSE and Log loss of both the datasets reveal substantial predictive power of GA-SVM model for classification. AUCs of SVM-RFE, Lasso-SVM, Grid-SVM, Linear SVM, and GA-SVM on dataset 1 and dataset 2 are laid-out in ►Figs. 6 and 7 consecutively. The AUC value 0.94 on dataset 1 and 0.84 on dataset 2 of GA-SVM model indicates that the classifier is able to discriminate all the TNBC and non-TNBC cases almost accurately. Precision recall curves of SVM-RFE, Lasso-SVM, Grid-SVM, Linear SVM, and GA-SVM on dataset 1 and dataset 2 are delineated in ►Figs. 8 and 9 successively. It was noted that GA-SVM model with higher AUC value in ROC curve generates good precision–recall curve as well.

Statistical Analysis

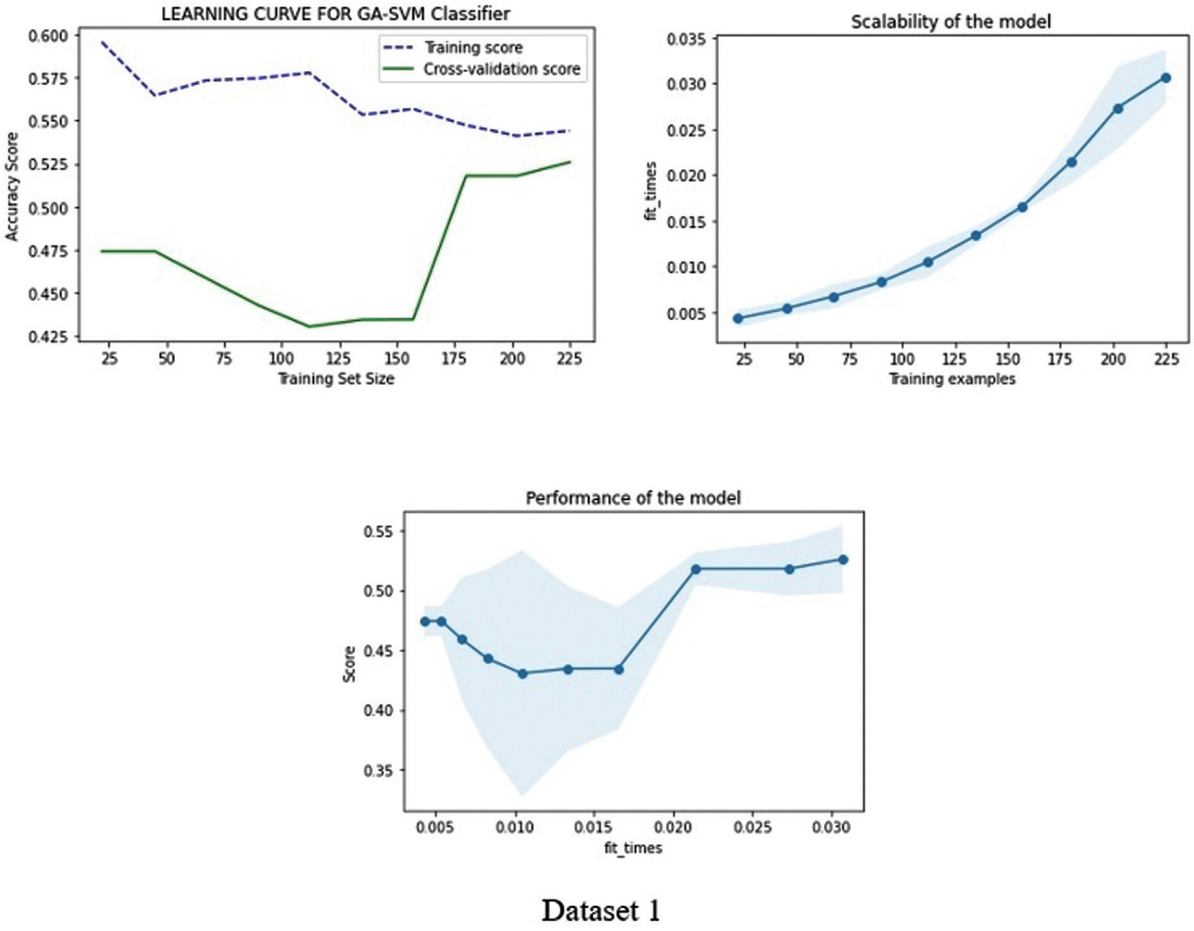
To understand the significance of clinicopathological parameters related to breast cancer classification, the correlation among the categorical clinical and pathological attributes was figured out by means of heatmap. Heatmap, a two-dimensional graphical representation of correlation matrix was used for identifying the linear relationship between the involved clinicopathological parameters. The positive–negative correlation between the clinicopathological attributes in



Dataset 1

Dataset 2

Fig. 3 Precision–recall curves of GA-SVM hybrid model on both datasets.



Dataset 1

Fig. 4 Learning curve, scalability, and performance of GA-SVM model on dataset 1.

the heatmap was highlighted with blue and red color, respectively. The higher correlation value among the clinicopathological parameters was indicated with the stronger color shades. The dark blue color heatmap diagonal signifies the correlation of the same variable with itself. The correlation heatmap of dataset 1 and dataset 2 are demonstrated in ►Figs. 10 and 11, respectively. ►Fig. 10 reveals the positive correlation of age with menopausal status, comorbidity, and hypertension. Strong correlation exists between histology type and node status, disease stage, and metastasis. Poor nutritional status among the older African[45] was justified with positive correlation of age and nutritional status. There also exists correlation between age and menopause, number of full-time pregnancies, nulliparity, and familial history of breast cancer on heatmap (►Fig. 11) of dataset 2. Further, tumor size bears correlation with surgery type, adjuvant chemotherapy, and radiotherapy.

Table 1 Classification accuracy of GA-SVM and other compared models on two datasets

Models	Classification accuracy	
	Dataset 1	Dataset 2
GA-SVM	93.4	90
SVM-RFE	90.4	84
LASSO-SVM	89.4	86.6
Grid-SVM	90.3	82.3
SVM	90.4	84

Abbreviations: Dataset 1, Lagos university, Nigeria breast cancer dataset; Dataset 2, National Institute of Oncology, Rabat, Morocco breast cancer dataset.

To test whether TNBC/non-TNBC is dependent or independent of the categorical clinicopathological parameters, Pearson's Chi-square test was conducted on two datasets. Chi square was formulated with the summation square of the observed value from the expected value of features divided by the respective expected feature value. Chi-square statistics adjust the degree of freedom of the feature level with the number of class level.

$$\tau^2 = \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \cdots + \frac{(O_{mn}-E_{mn})^2}{E_{mn}}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (3)$$

Here, τ represents Chi-square value, O_{ij} = observed value and E_{ij} = expected value of the features.

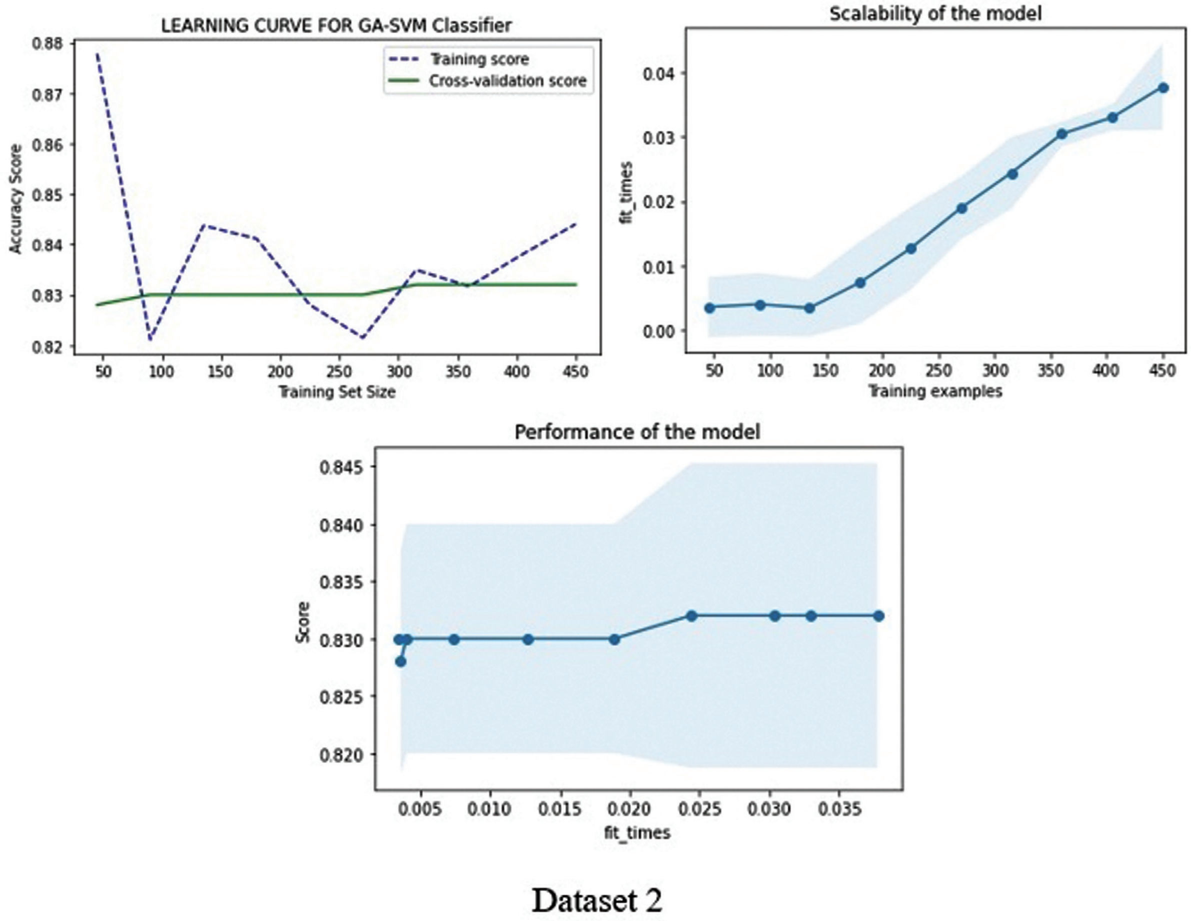


Fig. 5 Learning curve, scalability, and performance of GA-SVM model on dataset 2.

Chi-square statistics was implemented in python version 3.7.2 with output as features Chi-square score, Chi-square *p*-values, F-score, F-score *p*-values, and mutual information among the clinicopathological features with respect to the class level TNBC/non-TNBC. The results shows that patients height, body mass index, family history of breast cancer, and hormone receptors status are statistically significant ($p < 0.05$) clinicopathological parameters in dataset 1 for categorizing TNBC versus non-TNBC cases. In dataset 2, clinicopathological features hormone therapy and progression (metastasis/relapse) were found to be statistically significant ($p < 0.05$) in identifying TNBC/non-TNBC cases. Thus, the aggressiveness of breast cancer was justified with hormone receptor status, hormone therapy, distant metastasis, and early development of recurrence after surgery. The detailed statistical analysis with mean, standard deviation of clinical, pathological, and demographic parameters and their prognostic significance was investigated in the original studies.[27,29]

Discussion

The GA-SVM hybrid model performance was validated with several evaluation metrics, AUC, precision–recall curve,

Table 2 Several evaluation metrics comparative analyses of all models on dataset 1 and dataset 2

Models	Dataset 1				Dataset 2			
	Mean square error (MSE)	Log loss	AUC score	F1-score	Mean square error (MSE)	Log loss	AUC score	F1-score
GA-SVM	0.06	0.70	0.94	0.93	0.10	0.31	0.84	0.87
SVM-RFE	0.05	0.83	0.96	0.95	1.33	0.39	0.74	0.87
LASSO-SVM	0.10	0.64	0.93	0.90	1.33	0.32	0.87	0.87
Grid-SVM	0.06	0.19	0.96	0.93	0.12	0.29	0.91	0.87
SVM	0.10	0.66	0.95	0.89	0.13	0.33	0.89	0.87

Abbreviations: Dataset 1, Lagos university, Nigeria breast cancer dataset; Dataset 2, National Institute of Oncology, Rabat, Morocco breast cancer dataset.

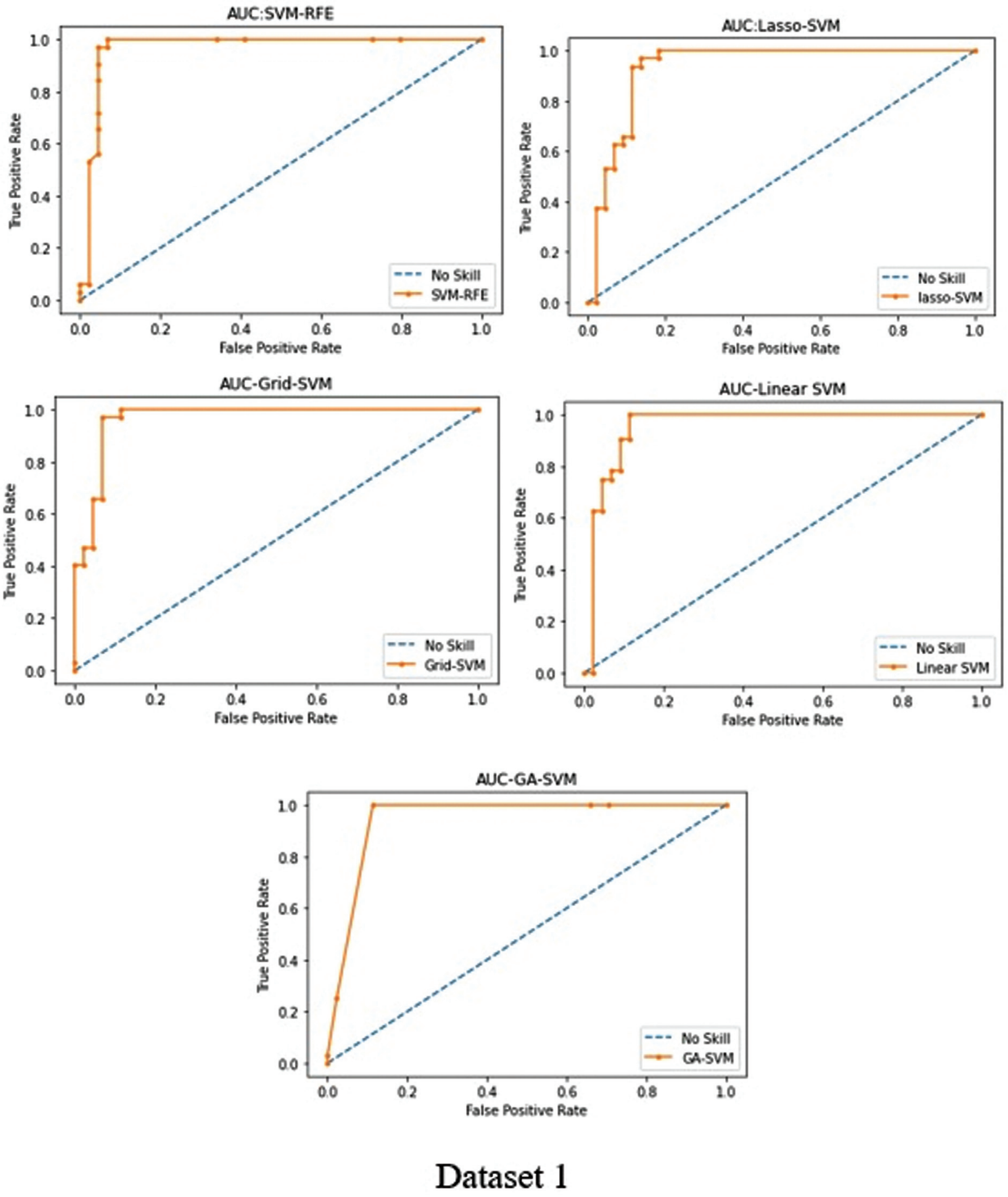


Fig. 6 Area under the curve (AUC) of SVM-RFE, Lasso-SVM, Grid-SVM, Linear SVM, and GA-SVM on dataset 1.

learning curves, statistical analysis and also obtained better classification accuracy as compared with classic feature selection SVM hybrid model which implies effective classification of TNBC versus non-TNBC variants of breast cancer patients. This study is consistent with Huang et al[46] where the predictive performance of SVM and SVM ensembles was assessed on small- and large-scale datasets for breast cancer prediction. Similar types of hybrid models have been applied in Alba et al, Moteghaed et al, and Zu et al[47–50] for breast cancer diagnosis, prediction, and classification. Recently, Xu et al[51] investigated the performance of supervised learning models like logistic regression, decision tree, random forest, gradient boosting, and light GBM with clinicopathological parameters for predicting 5-year survival analysis of TNBC patients at Sun Yat-sen Memorial Hospital, China. Hybrid models combine with different heterogeneous ML

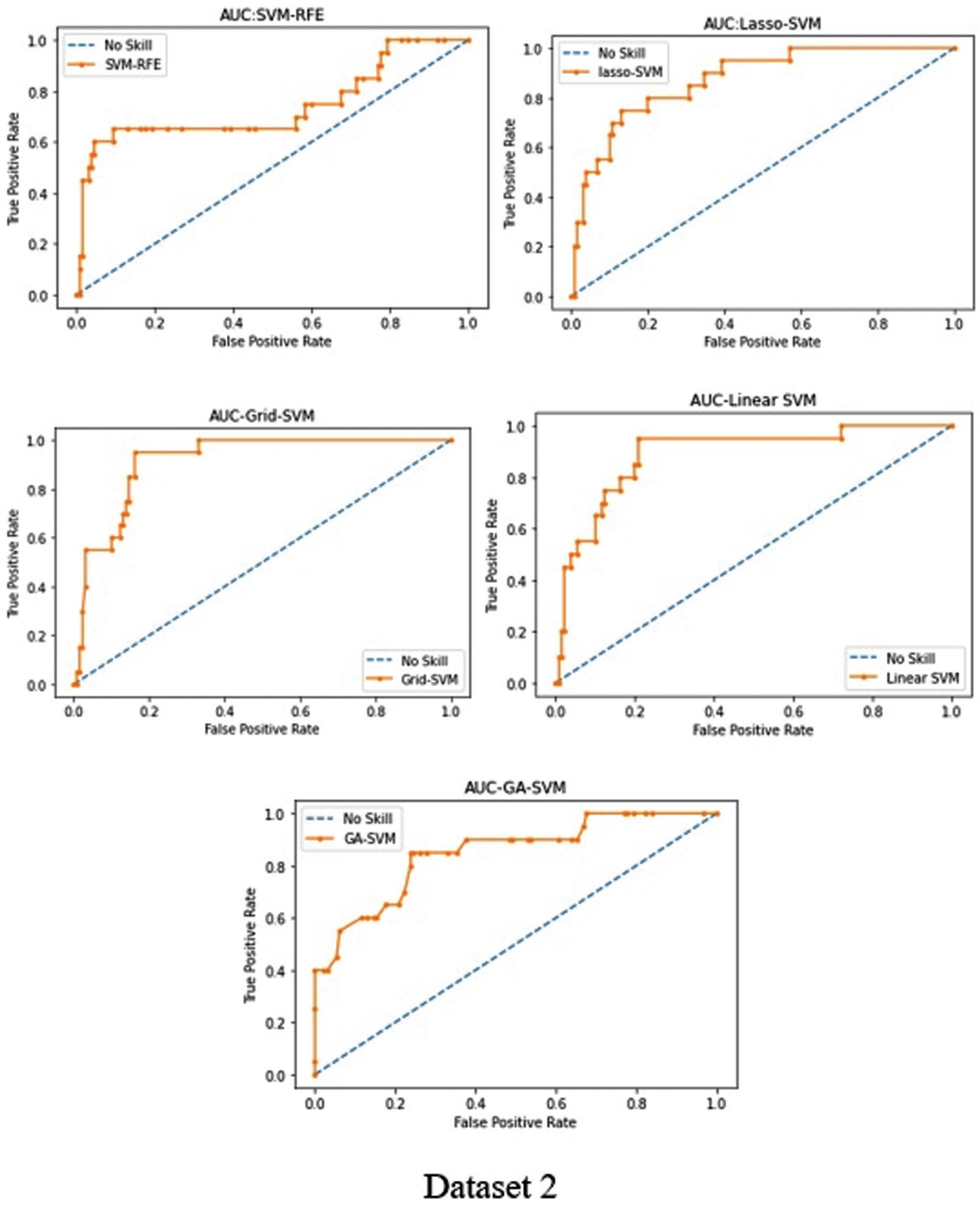


Fig. 7 Area under the curve (AUC) of SVM-RFE, Lasso-SVM, Grid-SVM, Linear SVM, and GA-SVM on dataset 2.

techniques and take the advantage of overcoming the weakness of individual models by integrating the complementary features of all the models involved.[52] As a result, hybrid models become more effective and robust as compared with individual ML classifiers.

Microarray-gene expression profiling has been largely studied for breast cancer classification, prediction, and prognosis. Recently, several studies encompassing the breast cancer classification on intrinsic subtypes mostly on gene expression data with ML approaches have been reported in the literature. Somatic mutation in genome exists predominantly in almost all cancers due to which identification of breast cancer on somatic mutation profile data emerged as an effective tool in clinical decision-making for personalized

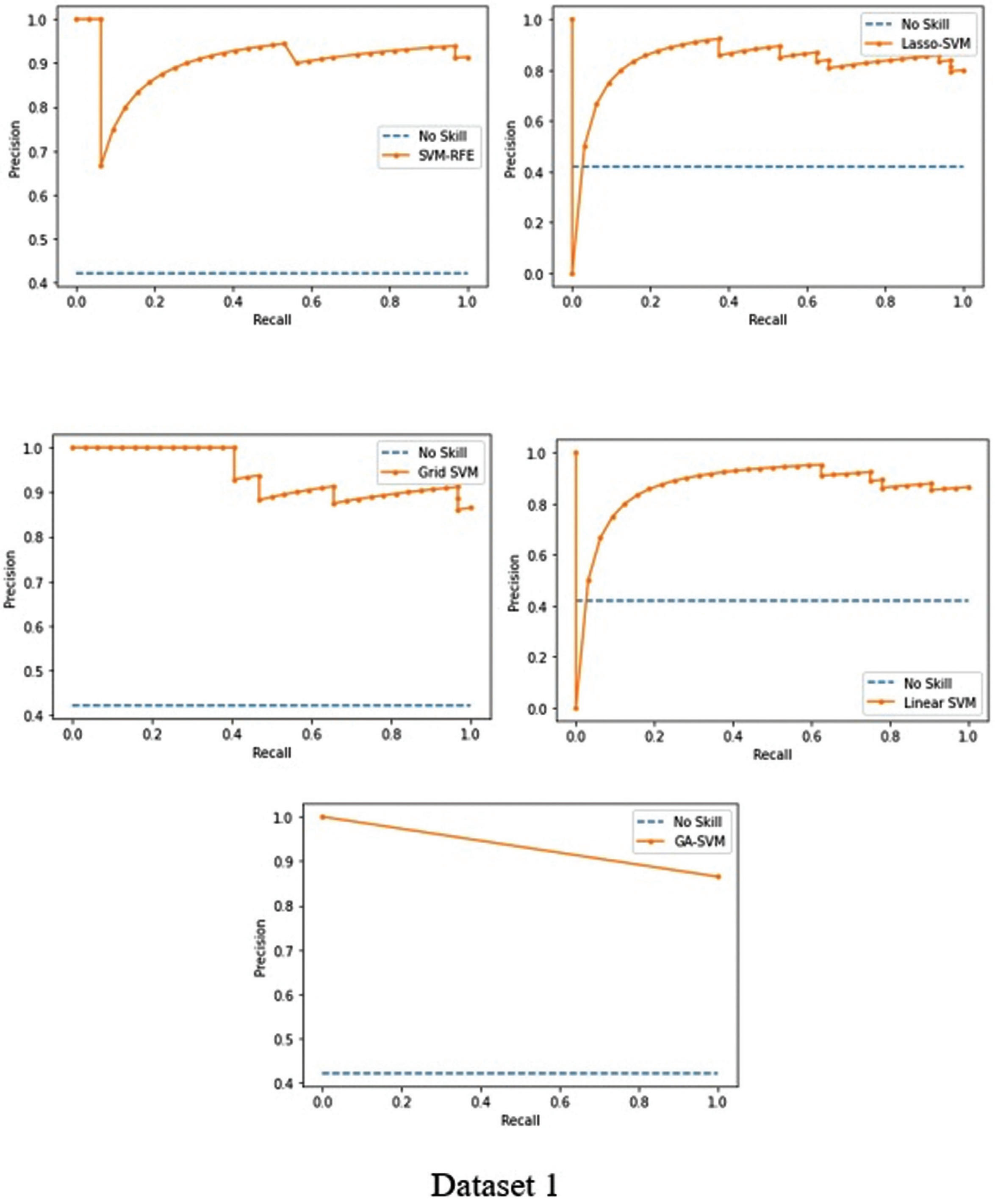


Fig. 8 Precision-recall curve of SVM-RFE, Lasso-SVM, Grid-SVM, Linear SVM, and GA-SVM on dataset 1.

patient care.[53] This study was the first to use somatic mutation profile data for categorizing breast cancer into subtypes with unsupervised ML methods. Beykikhoshk et al[54] applied deep learning architecture to classify gene expression signature of breast cancer subtypes namely luminal A and luminal B and calculated the individual patient biomarkers scores. A single sample platform independent subtype classifier with minimal number of genes that yield high classification accuracy by applying random forest classification algorithm was reported in Seo et al.[55] ML models were also employed to explore the interaction mechanisms of genes in identifying the five inherent categories of breast cancer with RNA-Seq data.[56] Xie et al[57] investigated the performance of MR multiparametric radiomics in differentiating the breast cancer subtypes with several ML models. Further, the radiomics model was also examined in

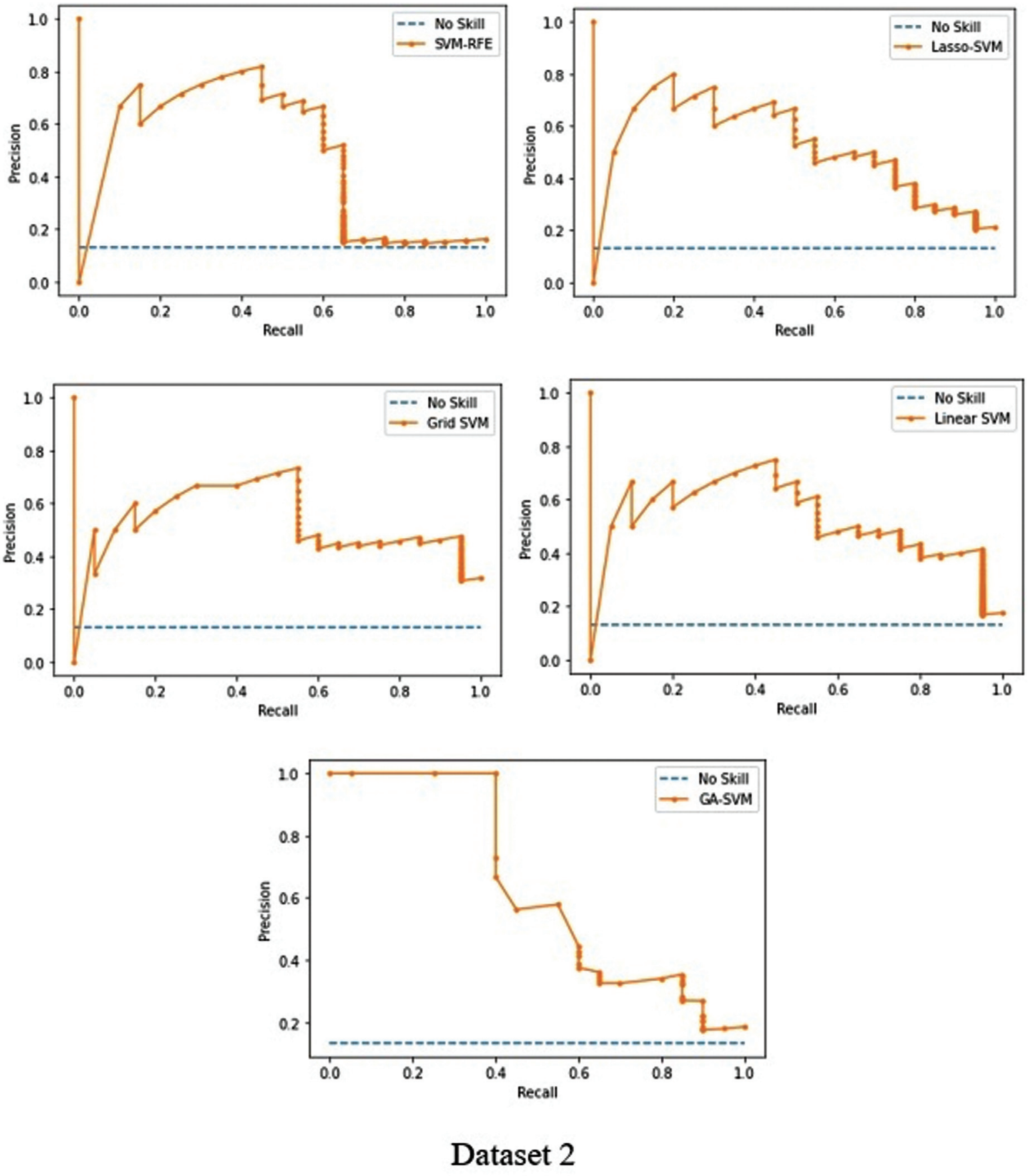


Fig. 9 Precision-recall curve of SVM-RFE, Lasso-SVM, Grid-SVM, Linear SVM, and GA-SVM on dataset 2.

identifying the aggressive TNBC from other inherent subtypes of breast cancer. Another study[58] also reveals the potential application of ML in classifying patient populations and it has been proved that SVM is capable of producing more accurate results with less misclassification errors. Ma et al[59] demonstrated the feature extraction of radiomics images from digital mammography exploring the strength of ML techniques to justify the association of breast cancer intrinsic subtypes in a Chinese population. One of the drawbacks of digital mammography is its incompetency to characterize certain biological and physiological properties of breast tissue. Moreover, the younger premenopausal women suffering from TNBC are recommended to avoid radiation hazards of mammography. Microarray is labeled as the golden procedure for breast cancer classification[60] but dynamic nature of genes in an individual may lead to misclassification errors. Hence, the inability to assign the molecular subtypes consistently is found to be a constraining factor. Owing to this, the potentiality of ML to uncover underlying patterns and simultaneously providing the predictive power to discriminate the various breast cancer types has been utilized.

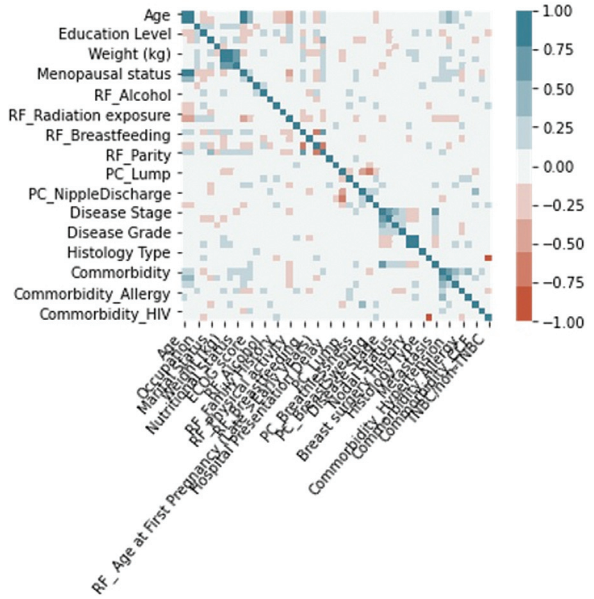


Fig. 10 The correlation heat map of dataset 1.

With the rapid advancement of ML techniques hybrid models have started to come out in the recent literature. Hybridization with one or more ML algorithms will increase the efficiency of the predicted models related to computation and accuracy. Many ML hybrid models have been applied in cancer diagnosis and classification.[61,62] Resmini et al[63] combines GA and SVM for investigating breast cancer with infrared thermography. However, most of the studies reported hybrid ML methods in ultrasound images, mammogram images, or digitized images of breast mass.[64,65] It is imperative to retrieve features from the medical image data before processing as it cannot be utilized as input directly. In medical sciences, classification of breast tumors depends on the expression profile of IHC markers and their relationship with clinical and pathological attributes. Due to heterogeneous behavior of breast cancer, evaluation of clinical and pathological parameters is also pivotal in distinguishing breast tumors for precise treatment and prognostic analysis. ML has been adopted for developing knowledge-based diagnostic system due to its capability to detect, identify, and distinguish breast tumors effectively. ML models with clinicopathological features can assist the doctors in clinical patient evaluation, determine surgical procedure, adjuvant therapies and develop precise treatment outcome. This motivated us to develop a classification model based on hybridization of ML techniques to identify TNBC and non-TNBC patients with clinicopathological parameters collected from multiple tertiary care hospital/centers.

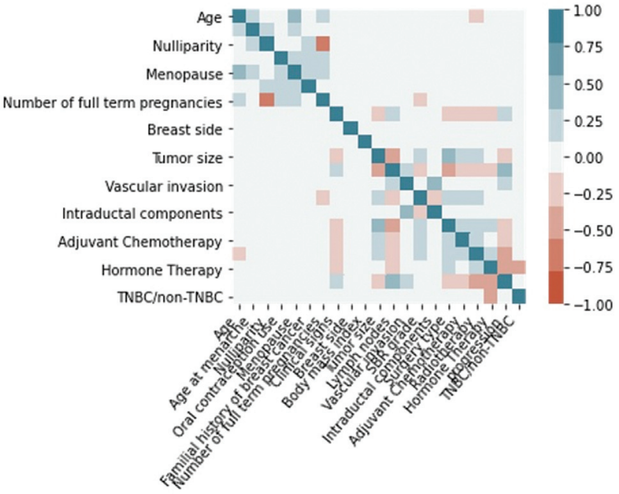


Fig. 11 The correlation heat map of dataset 2.

Several literature studies[66–70] that categorized TNBC versus non-TNBC group of patients based on the IHC staining with clinicopathological features of patient's data performed statistical analysis with SPSS or other well-known software rather than applying ML approaches for automatic breast cancer prognostic classification and treatment regimen. Our study of classifying the TNBC and non-TNBC groups of breast cancer from hospital collected patient datasets with clinicopathological parameters using hybrid ML model was rarely reported in literature. Further, our study has been validated with multicentered prospective patients' datasets concealing the privacy issues related to electronic medical record data which lack in earlier reported studies. The present study is perhaps the first to classify TNBC and non-TNBC subtypes of multicentered North West African breast cancer patients with clinicopathological features using hybrid ML model. The application of hybrid ML techniques in cancer classification leads to the emergence of knowledge-based system that would enable the doctors to take clinical decision more meticulously and within a short duration.[71] This present study reveals that GA-SVM hybrid model not only improves the prediction accuracy but also helps in identifying aggressive variant of breast cancer TNBC that is characterized with poor prognosis, distant metastasis, and early recurrence. Moreover, this knowledge-based system could assist the clinicians in providing appropriate treatment lay-out, prognostic prediction, and precision medicine due to the consideration of pathological and clinical features in the patient datasets.

Our ML hybrid model was compared and evaluated with the performance of three classic feature selection ML hybrid approaches but in reality, there are a lot of ML techniques that have not been considered in this study. Apart from TNBC, there exist multiple inherent variants of breast cancer that have not been investigated in this study. The smaller data size of the North West African datasets with unbalanced design of TNBC samples might be a constraint of this present study. However, the Lagos University, Nigerian dataset constitute 47.4% of TNBC samples which signifies almost balanced cases and high prevalence of TNBC in African subcontinent. In this paper, feature selection potency of GA has been applied on the SVM estimator but the capability of GA for tuning hyperparameters of SVM[72] has not been inspected in this study. Thus, the possible weakness of the study has been highlighted.

Conclusion

The present study reveals that GA-SVM ML hybrid model was suited to classify TNBC and non-TNBC variants of breast cancer accurately. However, more effective and accurate hybrid models are to be worked on for detecting precision medicine to tackle the aggressiveness of TNBC which lacks specific targeted therapy. Future research is suggested to investigate the multiple subtypes of TNBC with ML approaches and to identify the crucial clinical and pathological parameters for precise clinical outcomes.

Funding
None.

Conflict of Interest
None declared.

Acknowledgments
None.

References

1 Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71(03):209–249

2 World Health Organization. Breast Cancer. Accessed March 26, 2021 at:https://www.who.int/news-room/fact-sheets/detail/breast-cancer

3 Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast cancer prognosis using a machine learning approach. Cancers (Basel) 2019;11(03):328

4 Kim W, Kim KS, Lee JE, et al. Development of novel breast cancer recurrence prediction model using support vector machine. J Breast Cancer 2012;15(02):230–238

5 Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2014;13:8–17

6 Tao M, Song T, Du W, et al. Classifying breast cancer subtypes using multiple kernel learning based on omics data. Genes (Basel) 2019;10(03):200

7 Zolbanin HM, Delen D, Zadeh AH. Predicting overall survivability in comorbidity of cancers: a data mining approach. Decis Support Syst 2015;74:150–161

8 Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X. Developing prognostic systems of cancer patients by ensemble clustering. J Biomed Biotechnol 2009;2009:632786

9 Shah SM, Khan RA, Arif S, Sajid U. Artificial intelligence for breast cancer analysis: trends & directions. Comput Biol Med 2022; 142:105221

10 Saber A, Sakr M, Abo-Seida OM, et al. A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique. IEEE Access 2021; 9:71194–71209

11 Anderson P, Gadgil R, Johnson WA, Schwab E, Davidson JM. Reducing variability of breast cancer subtype predictors by grounding deep learning models in prior knowledge. Comput Biol Med 2021;138:104850

12 Zhao S, Wang P, Heidari AA, Chen H, He W, Xu S. Performance optimization of salp swarm algorithm for multi-threshold image segmentation: comprehensive study of breast cancer microscopy. Comput Biol Med 2021;139:105015

13 Liu L, Zhao D, Yu F, et al. Performance optimization of differential evolution with slime mould algorithm for multilevel breast cancer image segmentation. Comput Biol Med 2021; 138:104910

14 Huang H, Feng X, Zhou S, et al. A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features. BMC Bioinformatics 2019;20(08, Suppl 8):290

15 Tu J, Lin A, Chen H, et al. Predict the entrepreneurial intention of fresh graduate students based on an adaptive support vector machine framework. Math Probl Eng 2019;2019:1–16

16 Shahbakhi M, Far DT, Tahami E. Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine. J Biomed Sci Eng 2014;7(04):147–156

17 Chen X, Liu K, Cai J, et al. Identification of heavy metal-contaminated Tegillarca granosa using infrared spectroscopy. Anal Methods 2015;7(05):2172–2181

18 Sarkar JP, Saha I, Sarkar A, Maulik U. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. Comput Biol Med 2021;131:104244

19 Ben Azzouz F, Michel B, Lasla H, et al. Development of an absolute assignment predictor for triple-negative breast cancer subtyping using machine learning approaches. Comput Biol Med 2021; 129:104171

20 Howlader N, Noone AM, Krapcho M, et al. SEER*Explorer. Breast: Recent Trends in SEER Age-Adjusted Incidence Rates, 2000–2018, by Race/Ethnicity, Delay-Adjusted SEER Incidence Rate, Female, Ages 15–39, All Stages. Bethesda, MD: National Cancer Institute; 2021

21 Trivers KF, Lund MJ, Porter PL, et al. The epidemiology of triple-negative breast cancer, including race. Cancer Causes Control 2009;20(07):1071–1082

22 Amirikia KC, Mills P, Bush J, Newman LA. Higher population-based incidence rates of triple-negative breast cancer among young African-American women: implications for breast cancer screening recommendations. Cancer 2011;117(12): 2747–2753

23 Stead LA, Lash TL, Sobieraj JE, et al. Triple-negative breast cancers are increased in black women regardless of age or body mass index. Breast Cancer Res 2009;11(02):R18

24 Stark A, Kleer CG, Martin I, et al. African ancestry and higher prevalence of triple-negative breast cancer: findings from an international study. Cancer 2010;116(21):4926–4932

25 Nedeljković M, Damjanović A Mechanisms of chemotherapy resistance in triple-negative breast cancer-how we can rise to the challenge. Cells 2019;8(09):957

26 Biostudies. BioStudies – one package for all the data supporting a study. Available at: https://www.ebi.ac.uk/biostudies/

27 Mouh FZ, Slaoui M, Razine R, El Mzibri M, Amrani M. Clinicopathological, treatment and event-free survival characteristics in a Moroccan population of triple-negative breast cancer. Breast Cancer (Auckl) 2020;14:1178223420906428

28 Biostudies. Clinicopathological, treatment and event-free survival characteristics in a moroccan population of triple-negative breast cancer. Available at: https://www.ebi.ac.uk/biostudies/studies/S-EPMC7218339?query=Clinicopathological%2C%20Treatment%20and%20Event-Free%20Survival%20Characteristics%20in%20a%20Moroccan%20Population%20of%20Triple-Negative%20Breast%20Cancer%20Fatima%20Zahra%20Mouh

29 Adeniji AA, Dawodu OO, Habeebu MY, et al. Distribution of breast cancer subtypes among Nigerian women and correlation to the risk factors and clinicopathological characteristics. World J Oncol 2020;11(04):165–172

30 Biostudies. Distribution of breast cancer subtypes among Nigerian women and correlation to the risk factors and clinicopathological characteristics. Available at:https://www.ebi.ac.uk/biostudies/studies/S-EPMC7430856?query=distribution%20of%20breast%20cancer%20subtype%20among%20nigerian%20women

31 Zeng X, Chen YW, Tao C. Feature Selection Using Recursive Feature Elimination for Handwritten Digit Recognition. 2009. Paper presented at: Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing; 12–14 September 2009; Japan

32 Vapnik V, Lerner A. Pattern recognition using generalized portrait method. Autom Remote Control 1963;24:774–780

33 Goldberg DE. Genetic Algorithms in Search, Optimization and Machine Learning. New York: Addison-Wesley; 1989

34 Davis L. Handbook of Genetic Algorithms. Edition. New York: Van Nostrand Reinhold; 1991

35 Michalewicz Z. Genetic Algorithms+Data Structures, Evolution Programs. New York: Springer; 1992

36 Filho JLR, Treleaven PC, Alippi C. Genetic algorithm programming environments. IEEE Computer 1994;27:28–43

37 Scikit-learn. Machine Learning in Python. Available at: https://scikit-learn.org/stable/

38 Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5(09):1315–1316

39 Huang ML, Hung YH, Lee WM, Li RK, Jiang BR. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. ScientificWorldJournal 2014;2014:795624

40 Algamal ZY, Lee MH. Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Syst Appl 2015;42(23):9326–9332

41 Nursabillilah A, Nor A, Rosli B. Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection. Indonesian J Electrical Engineering Comp Sci 2020;20(02):712–719

42 Huang CL, Liao HC, Chen MC. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Syst Appl 2008;34(01):578–587

43 Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 2009;36(02):3240–3247

44 Asri H, Mousannif H, Moatassime AH, et al. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 2016;83:1064–1069

45 Charlton KE, Rose D. Nutrition among older adults in Africa: the situation at the beginning of the millennium. J Nutr 2001;131(09):2424S–2428S

46 Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF. SVM and SVM ensembles in breast cancer prediction. PLoS One 2017;12(01):e0161501

47 Alba E, Garcia-Nieto J, Jourdan L, et al. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. Paper presented at: 2007 IEEE Congress on Evolutionary Computation conference proceedings; September 25–28, 2007; Singapore

48 Moteghaed NY, Maghooli K, Garshasbi M. Improving classification of cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine. J Med Signals Sens 2018;8(01):1–11

49 Xu H, Chen T, Lv J, et al. A combined parallel genetic algorithm and support vector machine model for breast cancer detection. J Comp Methods Sci Engineering 2016;16(04):773–785

50 Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput Appl 2017;28:753–763

51 Xu Y, Ju L, Tong J, Zhou C, Yang J. supervised machine learning predictive analytics for triple-negative breast cancer death outcomes. OncoTargets Ther 2019;12:9059–9067

52 Castillo W, Melin O, Pedrycz P. Hybrid Intelligent Systems: Analysis and Design Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer; 2007:55–64

53 Vural S, Wang X, Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC Syst Biol 2016;10(Suppl 3):62

54 Beykikhoshk A, Quinn TP, Lee SC, Tran T, Venkatesh S. DeepTRIAGE: interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer subtypes. BMC Med Genomics 2020;13(Suppl 3):20

55 Seo MK, Paik S, Kim S. An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier. Cancers (Basel) 2020;12(12):3506

56 Yu Z, Wang Z, Yu X, Zhang Z. RNA-Seq-based breast cancer subtypes classification using machine learning approaches. Comput Intell Neurosci 2020;2020:4737969

57 Xie T, Wang Z, Zhao Q, et al. Machine learning-based analysis of MR multiparametric radiomics for the subtype classification of breast cancer. Front Oncol 2019;9:505

58 Wu J, Hicks C. Breast cancer type classification using machine learning. J Pers Med 2021;11(02):61

59 Ma W, Zhao Y, Ji Y, et al. Breast cancer molecular subtype prediction by mammographic radiomic features. Acad Radiol 2019;26(02):196–201

60 Peppercorn J, Perou CM, Carey LA. Molecular subtypes in breast cancer evaluation and management: divide and conquer. Cancer Invest 2008;26(01):1–10

61 Huerta EB, Duval B, Hao JK. A Hybrid GA/SVM approach for gene selection and classification of microarray data. evo workshops 2006. LNCS 2006;3907:34–44

62 Ngadi M, Nassih B, Hachimi H, et al. Genetic algorithms combined with support vector machine for breast cancer diagnosis. Paper presented at: International Workshop in Optimization and Applications Woa; 2016;17th-19th May 2016

63 Resmini R, Silva L, Araujo AS, Medeiros P, Muchaluat-Saade D, Conci A. Combining genetic algorithms and SVM for breast cancer diagnosis using infrared thermography. Sensors (Basel) 2021;21(14):4802

64 Wu T, Sultan LR, Tian J, Cary TW, Sehgal CM. Machine learning for diagnostic ultrasound of triple-negative breast cancer. Breast Cancer Res Treat 2019;173(02):365–373

65 Turkki R, Byckhov D, Lundin M, et al. Breast cancer outcome prediction with tumour tissue images and machine learning. Breast Cancer Res Treat 2019;177(01):41–52

66 Parshad R, Kazi M, Seenu V, Mathur S, Dattagupta S, Haresh KPSuhani. Triple-negative breast cancers: are they always different from nontriple-negative breast cancers? An experience from a tertiary center in India. Indian J Cancer 2017;54(04):658–663

67 Gogia A, Raina V, Deo SVS, Shukla NK, Mohanti BK. Triple-negative breast cancer: an institutional analysis. Indian J Cancer 2014;51(02):163–166

68 Sharma D, Singh G. An institutional analysis of clinicopathological features of triple negative breast cancer. Indian J Cancer 2016;53(04):566–568

69 Doval DC, Sharma A, Sinha R, et al. Immunohistochemical profile of breast cancer patients at a tertiary care hospital in New Delhi, India. Asian Pac J Cancer Prev 2015;16(12):4959–4964

70 Sharma M, Sharma JD, Sarma A, et al. Triple negative breast cancer in people of North East India: critical insights gained at a regional cancer centre. Asian Pac J Cancer Prev 2014;15(11):4507–4511

71 Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 2004;3(02):179–196

72 Kuang F, Xu W, Zhang S. A novel hybrid KPCA and SVM with GA model for intrusion detection. Appl Soft Comput 2014;18:178–184