Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–20

DOI: 10.1177/20552076231207203
journals.sagepub.com/home/dhj

Sage

Firefly-SVM predictive model for breast cancer subgroup classification with clinicopathological parameters

Suvobrata Sarkar[1] and Kalyani Mali[2]

Abstract

Background: Breast cancer is a highly predominant destructive disease among women characterised with varied tumour biology, molecular subgroups and diverse clinicopathological specifications. The potentiality of machine learning to transform complex medical data into meaningful knowledge has led to its application in breast cancer detection and prognostic evaluation.

Objective: The emergence of data-driven diagnostic model for assisting clinicians in diagnostic decision making has gained an increasing curiosity in breast cancer identification and analysis. This motivated us to develop a breast cancer data-driven model for subtype classification more accurately.

Method: In this article, we proposed a firefly-support vector machine (SVM) breast cancer predictive model that uses clinicopathological and demographic data gathered from various tertiary care cancer hospitals or oncological centres to distinguish between patients with triple-negative breast cancer (TNBC) and non-triple-negative breast cancer (non-TNBC).

Results: The results of the firefly-support vector machine (firefly-SVM) predictive model were distinguished from the traditional grid search-support vector machine (Grid-SVM) model, particle swarm optimisation-support vector machine (PSO-SVM) and genetic algorithm-support vector machine (GA-SVM) hybrid models through hyperparameter tuning. The findings show that the recommended firefly-SVM classification model outperformed other existing models in terms of prediction accuracy (93.4%, 86.6%, 69.6%) for automated SVM parameter selection. The effectiveness of the prediction model was also evaluated using well-known metrics, such as the F1-score, mean square error, area under the ROC curve, logarithmic loss and precision–recall curve.

Conclusion: Firefly-SVM predictive model may be treated as an alternate tool for breast cancer subgroup classification that would benefit the clinicians for managing the patient with proper treatment and diagnostic outcome.

Keywords

Firefly algorithm, support vector machine, predictive model, classification, clinicopathological parameters, triple-negative breast cancer

Submission date: 18 June 2023; Acceptance date: 26 September 2023

Introduction

Breast cancer, one of the prevailing cancers among the females, occurs due to proliferation of breast cells leading to the development of a breast mass. According to a 2022 report by the American Cancer Society,[1] there are around 287,850 new cases of invasive breast cancer among women in the United States, while there are 51,400 new

[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

Corresponding author:
Suvobrata Sarkar, Department of Computer Science and Engineering, Dr. B.C. Roy Engineering College, Durgapur, West Bengal 713206, India.
Email: suvobrata.sarkar@gmail.com

cases of non-invasive breast cancer. In total, 43,250 women have died from the disease. Breast cancer is believed to be a diversified disease with respect to its biological behaviour, treatment response and prognosis. Clinicopathological characteristics (such as tumour grade, size, lymph node status) and molecular parameters (like molecular subtypes, HER-2 and hormone receptor) are some of the variables that affect the prognosis of breast cancer. Based on the presence or absence of immunohistochemical markers that include oestrogen receptor, progesterone receptor, and human epidermal growth factors receptor-2 (HER-2), breast cancer can be categorised into four subtypes: luminal A, luminal B, HER-2 positive and triple-negative breast cancer (TNBC). Endocrine and targeted therapy have higher prognosis and survival rates for the majority of breast cancer subtypes.[2] In contrast to other breast cancer subtypes, TNBC[3] is distinguished by the absence of three receptors: oestrogen, progesterone and HER-2. This breast cancer subtype is also characterised by different tumour biology, an early recurrence rate, and a poor prognosis. Furthermore, there are no approved targeted therapies for this particular form of breast tumour. Cytotoxic chemotherapy is considered as the main therapeutic modality although resistance develop in some patients, while others suffer from various side effect.[4] This necessitates the categorisation of breast cancer into distinct subgroup for appropriate treatment planning and precise therapeutic options.

In order to handle heterogeneous medical dataset and to find complex relationship between them, support vector machine (SVM) acts as a powerful ML algorithm and is ideal for categorisation tasks. The idea of linear SVM was first investigated by Vapnik in 1963.[5] Finding a decision boundary that divides the data points into two components and minimises misclassification error is the core idea behind SVMs. It is necessary to select a decision boundary or hyperplane that maximises the distance between all nearby data points on both sides of the different classes. Support vectors are the closest data points on either side of the decision boundary, and the hypothetical line separating these closest points are known as margins. Thus, maximising the margin width will result in the best hyperplane. The main advantage of SVM is its capability to handle high-dimensional data as well as to perform well on small datasets. Non-linear data points are best handled by different kernel functions. However, the choice of kernel functions is pivotal as SVM is very sensitive, and for larger datasets, they are computationally expensive. The kernel SVM transforms the non-linear data points in low dimension to linear separable points into one dimension higher such that the data points in distinct classes are mapped in different dimension. SVM works well when there are more dimensions than samples. Many literatures have reported successful application of SVM in breast cancer type classification, cancer genomics, developing prognostic model, recurrence predictive model and survival analysis. Wu et al.[6] stated that SVM is capable of distinguishing TNBC and non-triple-negative breast cancer (non-TNBC) more accurately using RNA-sequence data of two patient populations. Huang et al.[7] investigated the classification power of SVMs on cancer genomics as a result of which new biomarkers, targeted medications and important knowledge about cancer-driver genes have evolved. By extracting prognostic information from clinical, demographic and biochemical data, a prediction model was developed in combination of SVM and random optimiser (RO) and has been reported in Ferroni et al.[8] Kim et al.[9] focussed on constructing breast cancer recurrence model that could predict 5-year recurrence rate following breast surgery in a Korean population, and the model prediction performance was also compared with other existing models. Bai et al.[10] explored the effect of peripheral lymphocytes in identifying prognostic markers among breast cancer patients, and SVM has been used in developing prognostic classifier. Mihaylov et al.[11] predicted the survival time of breast cancer patients originally generated from tumour-oriented clinical parameters like age of diagnosis, tumour stage and tumour size. The results showed the advantages of Linear SVM and other models in survival analysis. SVM has certain parameters that does not participate in the training phase of machine learning (ML) model but controls the behaviour of the model. These parameters are called hyperparameters and need to be adjusted upfront before the training phase. Hyperparameters play a crucial role in creating robust and precise model. They also contribute bias-variance trade-offs, thereby preventing underfitting or overfitting. As the real-world data are often noisy, linear separability into distinct classes may not be possible for linear SVM. A standard SVM attempts to segregate positive and negative classes almost accurately such that misclassification is minimised and finally leads to overfitting. To tackle this concept, the idea of 'soft margin' was introduced. Soft margin allows some data points to be misclassified at the expense of better generalisation. Soft margin determines the decision boundary in an optimisation problem by increasing the distance of decision boundary from the support vectors and also maximise the correctly classified data points in the training phase. This trade-off is handled by C parameter. The C parameter assigns a penalty for every misclassified point. For a smaller C value, decision boundary with larger margin is chosen as the misclassification penalty is low and larger C value results in decision boundary with smaller margin due to high penalty value. In case of radial basis function (RBF) SVM, gamma parameter determines the far-off influence of a particular training point. Low gamma value denotes larger similarity radius such that the margin between the classes is more generalised. For larger gamma value, the data points are very close to each other to lie within a class, resulting in overfitting. Hence, finding the optimum value for the hyperparameters is still challenging.

Over two decades, scientists and biologists have been studying the cooperative behaviour of social insects like

ants and bees' colonies, school of fish, birds flocking, worms and termites for building their nest, foraging, locating their prey and mating, which led to the evolution of swarm intelligence. In cellular robotics systems, Gerardo Beni and Jing Wang[12] originally put forth the idea of swarm intelligence, which were capable of producing intelligent behaviour. It focuses on the collective behaviour of individuals that coordinates in a decentralised manner and can self-organise within themselves. The individuals interact with each other on the basis of simple rules and utilises the local information exchange between the individuals during interaction or with their environment. The main characteristic of swarm intelligence is its capability to show group behaviour in spite of the absence of coordinator/in-charge of the group. All individuals in the swarm are independent of each other and have their own contribution in the group despite of other members' activity. Each individual behaves in a stochastic manner, which depends on the local perception of its neighbouring individuals. The collective and social behaviour of insects have been employed in finding complex optimisation problems efficiently such as the choice of shortest path from nest to food source by the colonies of ants. The chemical pheromones are left behind by the ants on the earth during their movement from nest to food source back and forth, while every ant takes a probabilistic decision of finding the shortest route based on the perceived pheromone intensity. Particle swarm optimisation (PSO),[13] ant colony optimisation,[14,15] artificial bee colony,[16,17] cuckoo search,[18] bat algorithm,[19] firefly algorithm,[20] krill herd method[21] and clustering algorithms[22,23] are few of the well-known swarm intelligence algorithms. Recently, some of the new promising nature-inspired metaheuristic algorithms like owl search algorithm,[24] monkey-king evolutionary algorithm[25] and Harris Hawks Optimisation[26] have also emerged. The search process for metaheuristic algorithms depends on a delicate balance between two factors: exploration and exploitation.[27] Exploration is the capability of an algorithm to locate diverse solution within the search space, and exploitation refers to the process of looking for the best solution nearby while making use of the knowledge already known. A nature-inspired, stochastic method called the 'Firefly Algorithm'[20] is based on the flashing patterns of tropical fireflies. The purpose of these flashing lights is to draw in potential mates and warn off potential predators. The fireflies' lantern, a light-emitting organ, is what creates the flashing light. The adult male fireflies have the ability to produce high and discrete flashes that attract females on the ground. The females begin generating continuous or blinking lights in specific patterns in response to this courtship signal. The female fireflies always prefer to have brighter male partners. However, there are rare instances in which the female fireflies cannot tell apart between distant flashes produced by strong light sources and nearby flashes produced by weaker light sources. These movement of fireflies based on the flashing patterns is being employed to develop a mathematical model for solving optimisation problems. The primary benefit of the Firefly algorithm is its speedy, simultaneous discovery of both global and local optimal solutions[28] when compared to other popular metaheuristic algorithms viz. genetic algorithm (GA) and PSO. The firefly algorithm has been used in this study to obtain the ideal SVM hyperparameter value.

Utilising clinical, pathological and demographic data gathered from three tertiary care cancer hospitals/oncological centres, a breast cancer classification model has been developed in this article based on the hybridisation of the firefly algorithm and SVM. This model can distinguish between patients with TNBC and non-TNBC. The social behaviour and the bioluminescent communication of tropical fireflies have been used for determining the optimal hyperparameter values of SVM. In firefly algorithm, the local attraction is always greater than distance attraction, which results in automatic division of population into sub-classes. Additionally, the firefly method is effective at handling non-linear and multimodal optimisation issues. RBF has been used as the basic kernel for the majority of literature presenting firefly approach for tweaking SVM hyperparameters over the years because of its dependability and flexibility in parameter management. Different kernel functions have been provided as alternatives and implemented in Python in this paper in order to achieve the best hyperparameter combination of SVM, rather than considering the RBF as the only option. This will allow the best kernel function to be evolved as an automated SVM parameter. The optimised SVM parameters are then applied in the training phase of ML model, which is capable of segregating the breast cancer patients into two distinct classes. The classification outcomes of the firefly-support vector machine (firefly-SVM) model have been assessed with other hyperparameter tuning models of traditional grid search-SVM (Grid-SVM) model, particle swarm optimisation (PSO-SVM) and the genetic algorithm-SVM (GA-SVM). The same multicentric datasets were used for all models, and the five-fold cross-validation approach was applied. The effectiveness of the prediction model was also evaluated using well-known metrics, such as the F1-score, mean square error (MSE), area under the ROC curve (AUROC), logarithmic loss and precision–recall curve.

The remaining part of the paper is structured as follows: Methods section emphasises on the datasets taken into consideration for study as well as the proposed hybrid model and existing hyperparameter tuning ML models. The suggested firefly-SVM model's classification performance is discussed in the Results section, along with comparisons to other existing models and statistical analysis demonstrating the dependence of clinicopathological characteristics in separating patients into the TNBC and non-TNBC classes. The discussion is highlighted in the next section, and finally, the paper ends with the Conclusion.

Methods

Datasets

Biostudies databases serve as descriptions of biological studies and connect their data to other databases, both

inside and outside of EMBL-EBI, that do not match the definition of structured archives.[29] In addition, the authors can provide supplemental data that is related to the publication. From Biostudies, two retrospective study datasets for African nations were obtained, while the third dataset was a longitudinal one from a tertiary care hospital in central India. A retrospective analysis was conducted at the National Institute of Oncology, Rabat, Morocco, consisting of 905 breast cancer patients treated in 2009 and was followed up to 2014.[30] The medical record of every patient was investigated carefully to obtain clinical, pathological and therapeutic implications. A total of 405 cases were taken out of the analysis because of missing data, overseas patients and male patients. The remaining 500 instances of female breast cancer were divided into two categories: 415 cases were non-TNBC, leaving only 85 cases with TNBC.[31] Another non-interventional study[32] was conducted among the participants at Lagos University Teaching Hospital, Nigeria. In this study, 251 patients who underwent initial outpatient clinic visits from July 2017 to July 2019 and were histologically determined to have breast cancer were taken into account. Only female participants above the age of 18 were included. A structured pro forma was prepared by interviewing the individual patients about their sociodemographic and disease attributes. Based on the molecular subtypes, the patients were evaluated into two groups: 119 (47.4%) cases as TNBC, while the remaining 43.2% were non-TNBC cases. The patient's datasets are readily available at Biostudies.[33] It should be noted that the authors of the original research[30,32] have provided an anonymised patient dataset in an excel file format, which is publicly available at Biostudies and has been utilised for performing this secondary analysis. The third study was a longitudinal one investigated at NKP Salve Institute of Medical Sciences and Research Centre, Nagpur, India. A total of 85 patients with breast cancer who had both histological and cytological confirmation of their diagnosis were recruited during the period 2012 and 2014. Case sheets for individual patient consisting of demographic information, clinical profile, associated risk factor and disease staging were recorded. A total of 37 (43.7%) cases were categorised as TNBC, and the remaining 48 patients are of non-TNBC. Moreover, TNBC tumour has aggressive histology of grade 3 when compared to non-TNBC group. The original study[34] has been conducted retrospectively and published in 2015, and as a result, ethical clearance for performing this secondary analysis with the same retrospective dataset has been waived.

Existing hyperparameter tuning ML models. To control the behaviour of ML models, hyperparameter tuning is a crucial component. If the hyperparameter is not tuned properly, the model parameters will estimate suboptimal results and unable to minimise the loss function. As a result, the model will have high misclassification errors leading to decrease in classification accuracy. There lies a distinction between hyperparameter and parameter in ML. The learning algorithm learns during the training phase, estimates the values of the model parameters and continues to update its values until the learning is completed. After the training phase, these model parameters become a part of the model, for example, the weight and bias of neural network. On the other hand, model parameters are computed by the hyperparameters, which are algorithm specific. Thus, hyperparameter tuning is essential and deals with manipulating optimal set of hyperparameter values for any learning algorithm.

Many research articles of nature-inspired optimisation algorithms viz. GA[35] and PSO[36] to train the hyperparameters of SVM and feature selection are available over a decade. Recently, Korovkinas et al.[37] applied PSO to tune the cost(penalty) parameter of linear SVM, thereby improving the accuracy of classifying textual data. Additionally, the majority voting ensemble technique is used to boost the model's effectiveness. GA suggested by John Holland[38] depends on the biological evolutionary process of natural selection. To evaluate optimisation problems, a GA seeks for a nearly optimal solution from an objective function. Initially, a random population is formed from a set of chromosomes that encodes the parameter of search space as strings. The fitness value, which represents the quality of the solution in the search space, is used to rank each member of the population. Based on the theory of natural evolution, a new population is created from the chromosomes with better fitness values. Then, processes of evolution like crossover and mutation are used to evolve subsequent generations of offspring. The process of natural selection, crossover and mutation continues for finite number of generations or until the stopping criteria is attained. In the context of SVM hyperparameter setting, every chromosome consists of two genes: cost C and gamma. On the training dataset, a five-fold cross-validation error is used to fit the fitness function. The generalisation capability of SVM strongly depends on the optimised setting of hyperparametric values of C and gamma. However, the basic hyperparameter of SVM is kernel for mapping non-linear data. Eberhart and Kennedy devised the population-based stochastic technique called PSO[39] after being inspired by the flocks of birds and their social behaviours. Every particle is considered as the potential solution of any minimisation function. There exists a global minimum in the search space. None of the particle knew the actual location of global minimum, but all the particles are assigned with a fitness score by the fitness function. The magnitude and velocity of the particles determine how they move in the subsequent iteration. Every particle can interact with other particles and share location and velocity. For a particular iteration, velocity is calculated by the inertia of the particle and two best positions: personal best and global best. Personal best refers

to a particle's optimal fitness value position, whereas global best refers to the optimal outcome attained by the entire swarm. The particle updates its position, and the fitness function is recalculated depending upon the new information available. On the training dataset, the fitness function changes out for a five-fold cross-validation error in order to adjust the hyperparameter values. The process is repeated until the population converges. The particle final position is the optimised solution attained. One of the most traditional methods for hyperparameter tuning is grid search.[40] In grid search, the hyperparameter domains are divided into discrete grids. Every combination of grid values is evaluated with cross-validation metrics. The grid point that maximises the cross-validation average value is considered as the best combination of hyperparametric values. Thus, the grid search is utilised for obtaining the optimised hyperparameters of a model to make accurate prediction. But the drawback of grid search is its computational complexity in calculating every possible combination of all hyperparameters.

The proposed model

The luminescence of the tropical fireflies acts as the inspiration for the firefly algorithm, a swarm intelligence, metaheuristic optimisation technique.[41–43] Firefly algorithm depends on three idealised rules: (1) Since fireflies are all unisex, they can be attracted to one another regardless of sex. (2) The attractiveness reduces with increasing distance between the fireflies and is closely correlated to the intensity of the light of the fireflies. The firefly with less luminescence will migrate in the direction of the firefly with more luminescence. If there exist no brighter fireflies than the current one, then it will travel at random inside the search area. (3) A firefly's brightness is related to the landscape of objective function. Thus, the firefly deals with two vital issues: the formulation of luminous intensity and the second one is the variation of attractiveness. Due to the fact that light intensity varies with distance and some light is absorbed by the medium, therefore:

$$I = I_0 e^{-\gamma r^2} \tag{1}$$

Where I denotes light intensity, I_0 = light intensity at the original source, distance = r and γ represents light absorption coefficient.[44] As observed by the nearby firefly, the attractiveness is directly proportional to the light intensity, hence the fluctuation in attractiveness β with distance r is given by the equation:

$$\beta = \beta_0 e^{-\gamma r^2} \tag{2}$$

Where β_0 is the attractiveness at the source $r = 0$. The movement of firefly i to another brighter firefly j is defined as:

$$x_i = x_i + \beta_0 e^{-\gamma r^2}(x_j - x_i) + \alpha \varepsilon_i \tag{3}$$

Where x_i denotes the current location of firefly i, the second term refers to how attracted the fireflies are to one another, α signifies the randomised scaling parameter for regulating the step size, and ε_i is the random vector selected at random from a variety of distributions. This attractiveness behaviour of fireflies has been applied in handling exploitation and exploration efficiently and thereby making it more robust when compared to other evolutionary algorithms. Moreover, firefly algorithm is good at finding global optima and local optimal values simultaneously. This urged us to apply firefly algorithm in obtaining the optimal set of hyperparameter values of SVM. The detailed discussion about SVM has been provided in the previous section. The firefly algorithm performs several runs to optimise certain population size for fixed number of generations. There are certain parameter combinations that are to be cached to get benefited for more optimised runs. These necessitate to define size of the population, maximum number of generations, number of independent runs, maximum number of stagnating generation, scoring and random state before applying firefly algorithm. Python version 3.11.2 and its related statistical packages have been utilised for entire implementation purpose.

The firefly algorithm search procedures are as follows:

estimator = svclassifier
param_grid = {"kernel": ["rbf", "sigmoid", "linear"], "C": np.logspace(−1, 1, num = 25, base = 10), "gamma": np.logspace(−1, 1, num = 25, base = 10)}
cv = 5
verbose = 1
population_size = 25
max_n_gen = 100
max_stagnating_gen = 10
runs = 3
scoring='f1_macro'
random_state = 42

where the estimator means the model instances that are to be passed to check the hyperparameters; param_grid represents the dictionary with parameter names along with the list of parameter settings that we want to find out; cv = 5 stands for five-fold cross-validation; verbose denotes the logging information and the possible values of level can be 0, 1, 2; population_size is represented by the number of population's trained estimators. The search process will move more quickly in areas with lower populations, but there exists a possibility of getting struck to local optimum while larger population size may slow down the search process. On the other hand, the search method turned out to be a traditional grid search if the population size is equal to or close to the number of parameter combinations. The max_n_gen specifies the total number of generations after which the search will be stopped. The max_stagnating_gen = the most generations up to which the classification score will remain constant before the optimisation for a specific run is terminated. runs = the

number of independent iterations required for optimisation. The optimised results obtained from parameter combination are temporarily stored between the runs that make every subsequent runs faster than the previous ones. Scoring means the passing of valid string/object of an evaluation metric, and random_state with an integer value signifies the seeding of the random generator.

Pandas, an open-source library of python, has been utilised for dealing with relational and labelled data. Pandas' data frame has been created by importing the datasets from existing excel files. Pandas' data frame is comparable to a feature matrix, where columns reflect the relevant patient's clinicopathological parameters, and rows represent the patient's anonymous identity. Details about the sociodemographic, clinical and pathological features utilised for model creation have been presented in the form of a Supplemental table. Scikit-learn,[45] a python in-built ML library, has been applied to support various data analysis functionality. Another python open-source library, NumPy, was imported to work with matrix data structure and multidimensional arrays. Data preprocessing in python was carried out while taking care of missing value, categorical features, normalisation of dataset and finally splitting the dataset into training and validation sets. The SimpleImputer function has been applied to replace the missing values with imputation strategies like mean, median, most_frequent and constant. StandardScaler, a data standardisation function, was used to rescale the data such that the distribution of values has zero mean and unit variance. The train_test_split technique was used to split the dataset into training and test sets. The model was fitted using the training data, which was obtained from the observed data. The test set was utilised for model evaluation with unseen data. The validation dataset, a subset of training set, was used for estimating model performance, adjusting the model hyperparameters. For the proposed hybrid model, the train_test_split method splits the datasets randomly in the proportion 7:3, that is, 70% of the dataset was used as a training set and the remaining 30% as test set. The estimator object that learns from the data was svclassifier. The sequence of dictionary with parameter name kernel, C and gamma along with the discrete values of each parameter viz. 'rbf', 'sigmoid', 'linear' for kernel constitute the parameter grid. The initial population size of firefly was 25. The maximum number of generations for optimisation and maximum stagnating of generation was fixed at 100 and 10, respectively. The data logging information verbose was set to 1. Scoring = f1_macro denotes the arithmetic mean of all per-class F1 scores, and the number of independent runs was assigned to 3. Table 1 exhibits the development process of firefly-SVM predictive model as a flowchart.

The necessary algorithms for the suggested firefly-SVM predictive model implemented in python are as follows:

Step 1: Import the dataset as pandas' data frame with m = patients' identity and n = clinicopathological features.
Step 2: Handling the missing values and data standardization technique with SimpleImputer and StandardScaler function respectively.
Step 3: Class labels as (m × 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 with FA algorithm.
Step 6: Display the best parameter combination with best_params_
Step 7: Train the model with fit method: svclassifier.fit
Step 8: Perform prediction using predict method: svclassifier. predict
Step 9: Generate classification_report.

Research ethics and patient consent. The original datasets of African countries were publicly available from Biostudies as a Supplemental material. For the purpose of conducting this secondary analysis, the corresponding authors of the original research have been communicated regarding the ethical permission. The ethical permission has been waived as the original research has already been conducted and published in the year 2020. Multiple times ethical approval was not required for the same data to be investigated, which are publicly available. Moreover, this secondary analysis does not involve direct human participation. As a result, informed consent from the patients was not necessary for carrying out the present study. The datasets collected were scrutinised and validated with the clinical collaborators for performing the current study. I hereby also declare that the patient's dataset utilised in this paper is purely for research work, maintaining the anonymous patient's identity that is neither disclosed nor shared publicly at any time.

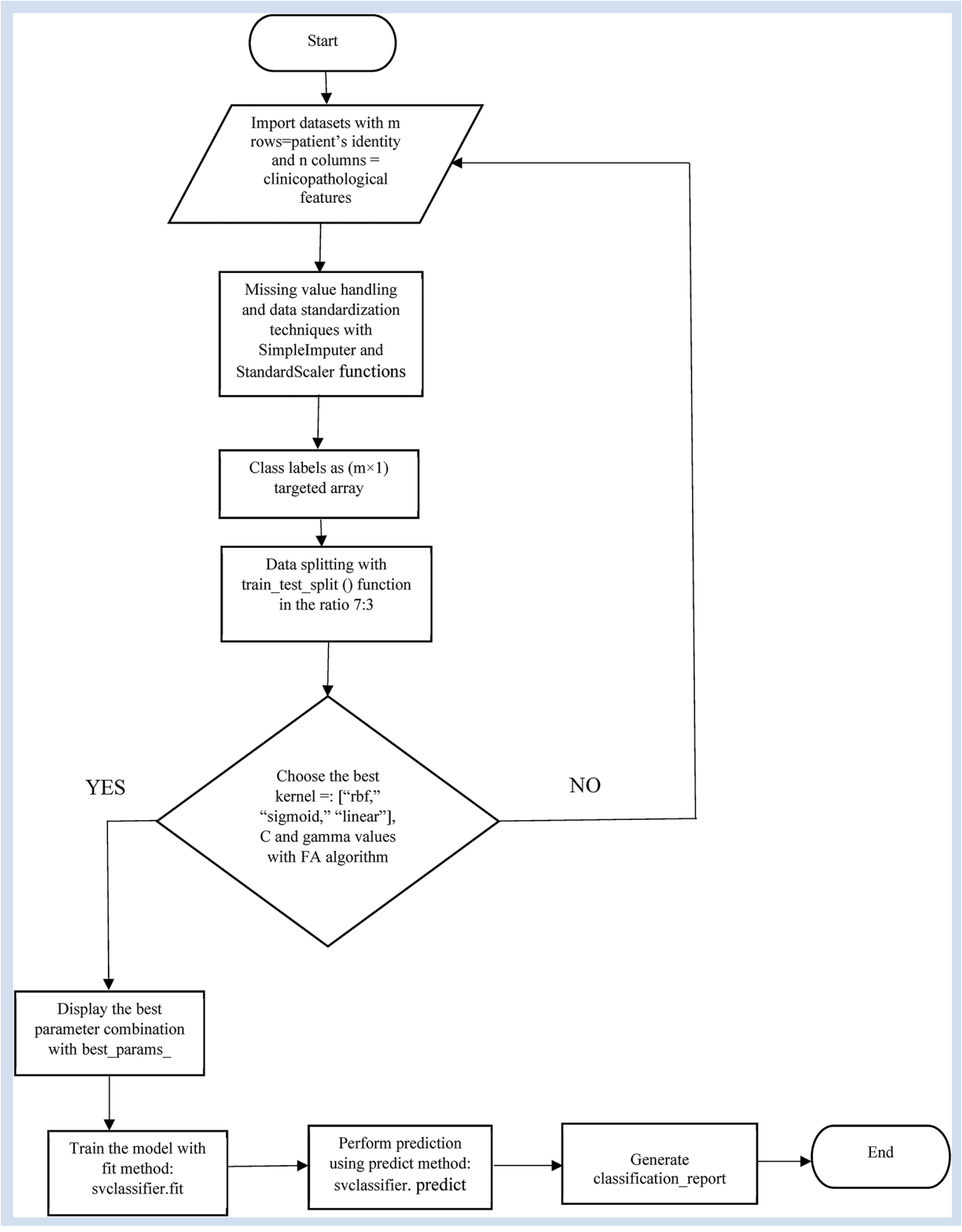
Results

This paper analyses three hospital datasets of patients with clinicopathological features and are diagnosed with breast cancer in different tertiary care hospitals or oncological centres. At Morocco's National Institute of Oncology, the study's first dataset contained 905 individuals who had received breast cancer treatment. Finally, 500 cases were taken into consideration due to incomplete medical records and male participants. In total, 251 breast cancer patients recruited at the Lagos University Teaching Hospital in Nigeria were assessed in the second dataset. The third dataset was investigated at NKP Salve Institute of Medical Sciences and Research Centre, Nagpur, India, with 85 enrolled patients with histopathological and cytologically confirmed cases of breast cancer.

Performance evaluation of firefly-SVM hybrid model

Some popular assessment metrics, including the confusion matrix, area under the receiver operator curve (ROC curve), MSE, logarithmic loss, precision–recall curve, and learning curve, were used to assess the performance of the

Table 1. Flowchart of the proposed firefly-SVM predictive model.



firefly-SVM model. A confusion matrix is a tabular representation that depicts the number of accurate and inaccurate predictions made by the classifier. It assesses the performance of the hybrid model through the evaluation of widely known metrics viz. accuracy, precision, recall, F1-score and support. For convenience, Lagos University

Hospital, Nigeria dataset, National Institute of Oncology, Morocco dataset and NKP Salve Institute of Medical Sciences, Nagpur dataset were designated as dataset 1, dataset 2 and dataset 3, respectively. Further, TNBC and non-TNBC cases were denoted as 1 and 0. Figure 1 displays the classification report for the firefly-SVM hybrid model on datasets 1, 2 and 3. Higher precision, recall and F1-score values suggest that the hybrid model almost accurately distinguished TNBC patients from non-TNBC instances.

The AUROC is a diagnostic tool for predicting probability under different threshold values on two-class classification problems. It is laid out graphically by plotting false-positive rate along the *x*-axis and true positive rate along the *y*-axis for different threshold values lying between 0.0 and 1. Smaller values on the *x*-axis suggest low false-positive and high true negative cases, while the larger values on the *y*-axis indicate higher true positive cases and low false negatives. On an average, randomly chosen positive instances are assigned with higher probability by a skillful model when compared to negative instances. Skillful models are illustrated with curves that move upwards from the top left. A no-skill model cannot differentiate between the classes and in most cases predict a random or any constant class. For all threshold values, it is drawn with a diagonal line from the bottom left to the top right of the plot, appearing at point (0.5,0.5) with an ROC value of 0.5. When a model's values lie between (0,1) and a line is drawn from the bottom left of the plot to the top left and then upwards, the model is said to have perfect skill.[46] Points lying above the diagonal line denote better classification results, while points below the line signify bad classification. The ROC of the predictive model was shown on three different datasets in Figure 2. The ROC of dataset 1 attained true positive rate = 1 with a corresponding fall-out value of 0.1 and moved across to

Classification report of firefly-SVM model on dataset 1

	precision	recall	f1-score	support
0	1.0000	0.8864	0.9398	44
1	0.8649	1.0000	0.9275	32
accuracy			0.9342	76
macro avg	0.9324	0.9432	0.9336	76
weighted avg	0.9431	0.9342	0.9346	76

Classification report of firefly-SVM model on dataset 2

	precision	recall	f1-score	support
0	0.9435	0.9000	0.9213	130
1	0.5000	0.6500	0.5652	20
accuracy			0.8667	150
macro avg	0.7218	0.7750	0.7432	150
weighted avg	0.8844	0.8667	0.8738	150

Classification report of firefly-SVM model on dataset 3

	precision	recall	f1-score	support
0	0.5238	0.8462	0.6471	13
1	0.6000	0.2308	0.3333	13
accuracy			0.5385	26
macro avg	0.5619	0.5385	0.4902	26
weighted avg	0.5619	0.5385	0.4902	26

Figure 1. Classification report of firefly-SVM predictive model on three datasets. 0 stands for non-TNBC cases and 1 represents TNBC cases.

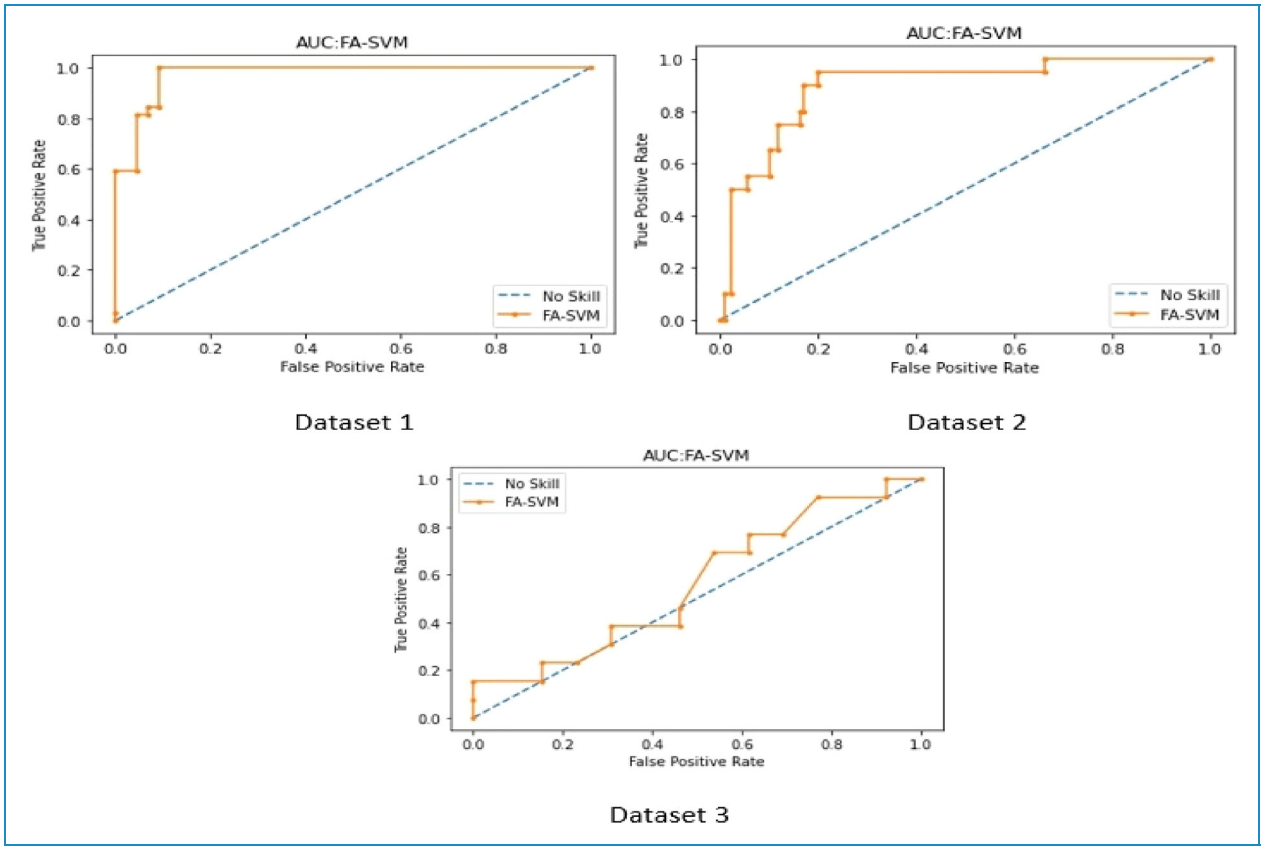


Figure 2. Area under the curve (AUC) of firefly-SVM predictive model on three datasets. AUC is plotted graphically with false-positive rate in the *x*-axis and true positive rate in the *y*-axis. The blue dashed line denotes the no-skill line. The orange colour line represents the model skill before reaching (1,1).

coincide with no-skill line. Dataset 2 curve bowed up steadily and achieved sensitivity 1 at 0.7 false-positive value before reaching (1,1). The dataset 3 ROC curve rises up from bottom left with some threshold points touching the diagonal line and finally attain (1,1).

A well-known loss function called mean square error calculates the average of the square variations between the value predicted by the model and the actual value.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \tag{4}$$

where y_i denotes the model predicted value, $\widehat{y_i}$ stands for the actual value and N signifies the total number of patients in two datasets from north-western Africa and one from India that were used as test cases. A lower MSE value implies that the model classified the data almost precisely and the estimated value is close to the actual value. MSE is never negative because of the squaring of the error. On datasets 1, 2 and 3, the firefly-SVM model's respective MSE values were 0.06, 0.13 and 0.46, indicating low MSE values and a good degree of classification power for the hybrid model.

To investigate the effectiveness of the classification model based on the idea of probability, another Loss function called Logarithmic Loss or Log Loss is employed as an evaluation metric. Log loss value deals with the measure of uncertainty of the model predicted value and its variation with the actual class label. The smaller is log loss value, the lesser is the deviation of predicted probability from the actual class label. For a model to be perfect, log loss value is equal to zero. The negative average multiplied by the sum of each patient's logarithmic predicted probability results in the computation of log loss.

$$Logloss_i = -[y_i \ln p_i + (1 - y_i)\ln(1 - pi)] \tag{5}$$

where i stands for the related patient, y_i for the actual value, p_i for the expected probability and log for the number's logarithmic value. The log loss values for datasets 1, 2 and 3 using the hybrid model were 0.19, 0.32 and 0.74, respectively.

Under different probabilistic thresholds, the precision–recall curve determines the balance that exists between the true positive rate (recall) and positive predictive value (precision). It is plotted with multiple threshold settings with recall on the *x*-axis and precision on the *y*-axis. It illustrates how well the model predicts the positive class and is suitable for datasets with imbalances. Precision–recall

curve with non-overlapping regions indicates better classification results than the one close to the baseline. Precision–recall curve of hybrid model on the three datasets are demonstrated in Figure 3. Dataset 1's precision–recall curve shows no overlapping areas and is significantly higher than the baseline. The dataset 2 curve moves in a zigzag manner before reaching near the baseline, while dataset 3 finally touches the baseline.

Convergence of optimised ML algorithms can be recognised as stopping condition on attaining a stable point beyond which further iteration of the algorithm does not produce any further improvement or changes. It is measured and explored empirically with the help of learning curve. Learning curve is a diagnostic tool for ML algorithms that learn from the training datasets incrementally. The model's learning and generalisation capabilities are demonstrated by the evaluation of the learning curve. The learning curve maps how well the model learns over time or with experience. It can also be utilised to investigate underfit, overfit or well-fit models. A learning curve determines the number of samples for training necessary to fit the model with the trade-off between bias and variance. Figure 4 depicts learning curves for datasets 1, 2 and 3 with the size of the training set on the *x*-axis and accuracy score on the *y*-axis, respectively. In dataset 1, the training score of the learning curve was high and steady on addition of training set, while the cross-validation score was low initially and then increases gradually with the training size. For dataset 2, the training score falls rapidly on 100 and 225 training size before attaining stability above 300 training sample. The cross-validation score increases initially with slight deep around 250 set size and then moves almost linearly. The training score of dataset 3 decreases abruptly around 20 training samples and increases further with the set size. The cross-validation score also had break point near 20 sample sizes and obtained a constant score with a 0.56 accuracy. The duration of time needed for the model to fit the estimator with the training dataset determines the model's scalability. Training examples are plotted on the *x*-axis, and fit_times are plotted on the *y*-axis. The amount of time the model needs to fit the classifier using the training examples for each cross-validation is known as fit_times. The curve of dataset 1 and dataset 2 increases gradually with the training examples and attains peaks at fit_times 0.30 and 0.05, respectively, while the curve on dataset 3 achieves scalability with 0.004 fit_times.

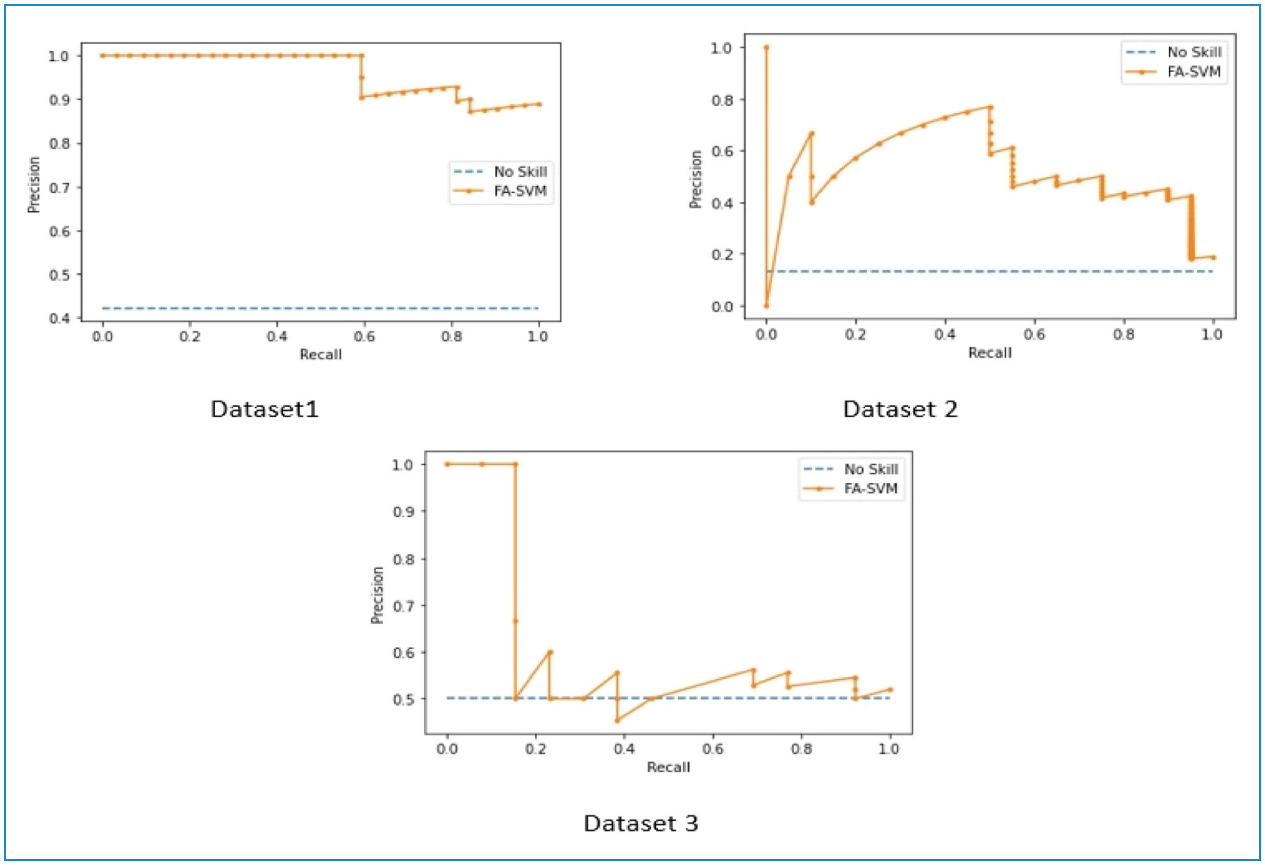


Figure 3. Precision–recall curves of firefly-SVM predictive model on three datasets. It is plotted graphically with recall in the *x*-axis and precision in the *y*-axis. The blue dashed line denotes the no-skill line just above the base, and the orange colour line represents the model skill before touching the baseline.

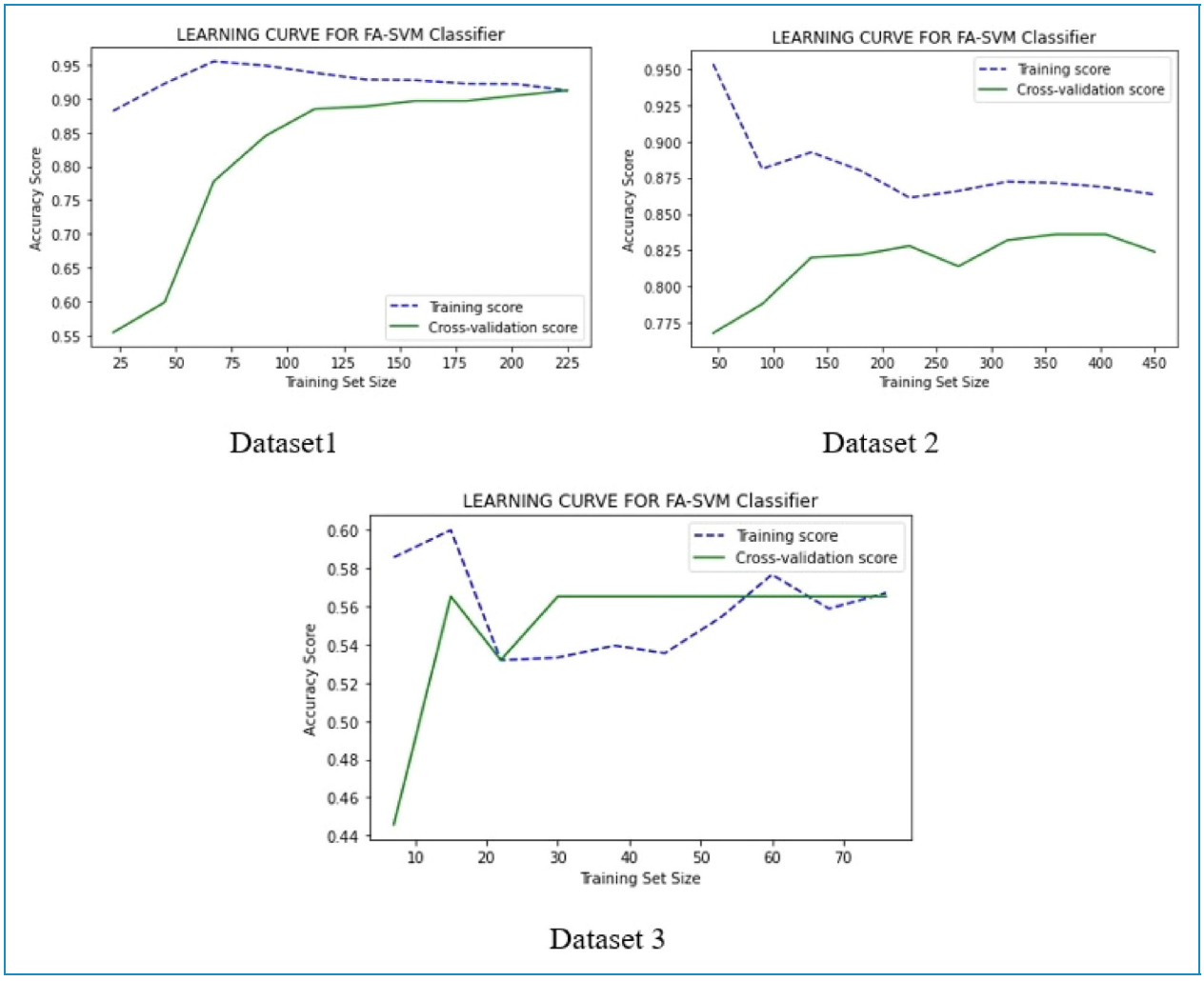


Figure 4. Learning curve of firefly-SVM predictive model on three datasets. The learning curve is plotted with training set size in the *x*-axis and accuracy score in the *y*-axis, respectively.

Fit_times versus test score was used to assess the performance of the model. The dataset 1, 2 and 3 attain stability at test score above 0.50, 0.80 and 0.55, respectively. Figures 5 and 6 demonstrate the scalability and effectiveness of the hybrid model on three datasets.

Comparison with other existing models

The effectiveness of the newly developed firefly-SVM model for identifying breast cancer was compared with other existing hyperparameter tuning hybrid models of traditional Grid-SVM model, GA-SVM, and PSO-SVM. The details of these existing hybrid models were covered in the previous section. On three datasets—two from Africa and one from India—Table 2 compares the classification accuracy of existing models and the firefly-SVM model. Firefly-SVM model's classification accuracy on datasets 1, 2 and 3 was 93.4, 86.6 and 69.6, respectively, which clearly outperformed the other existing model results. Tables 3–5 represent the optimised hyperparameters values like kernel, C and gamma of SVM on firefly-SVM model and other classic hybrid models on dataset 1, 2 and 3. Over the years, majority of the literature reporting firefly algorithm for tuning hyperparameters of SVM have adopted RBF as the base kernel due to its reliability and ease of adaptability in parameter handling.[47–49] Owing to achieve the best hyperparameter combination of SVM, in this paper, different kernel functions have been provided as alternatives and implemented in python so that the best kernel function can be evolved as an automated SVM parameter rather than considering RBF as an ultimate choice. The kernel functions obtained on dataset 1, 2 and 3 were radial basis, linear and sigmoid functions. A greater *C* parameter value signifies that the firefly-SVM model on dataset 1 seeks to minimise the misclassified samples as a result of a high penalty value, while a smaller gamma value of the model indicates substantial similarity radius, allowing for more points to fit into a certain class. This is reinforced

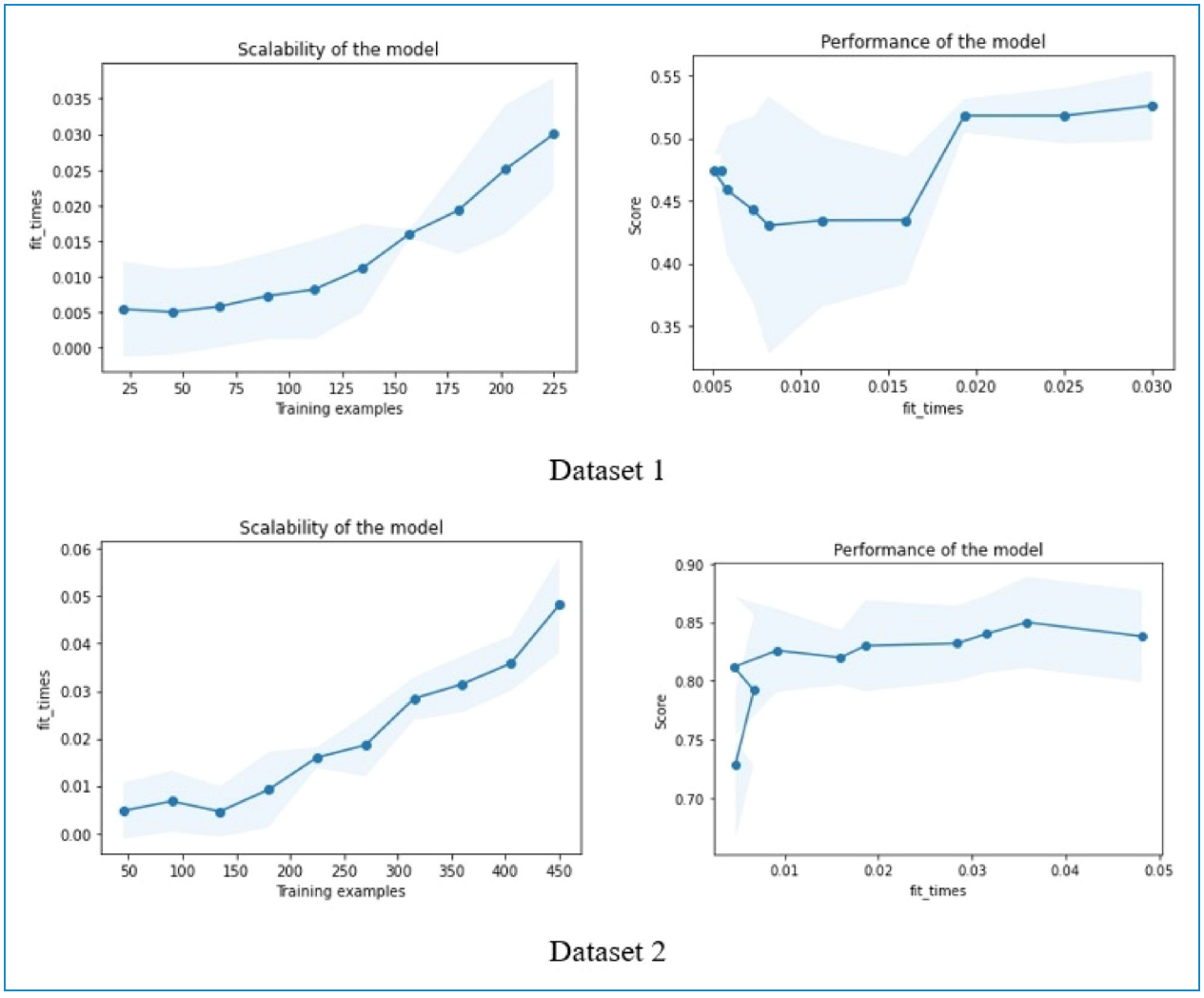


Figure 5. Scalability and performance of firefly-SVM predictive model on dataset 1 and 2. The scalability of the model denotes the time required by the model to fit the estimator with the training dataset. The blue-shaded region in the scalability graph indicates the region of fit_times mean +&- fit_times standard deviation. The performance of the model represents the test score with respect to fit_times. The blue-shaded region indicates the region of test scores mean + & - test scores standard deviation.

by the fact that dataset 1's higher classification accuracy is 93.4%. A lower *C* value on datasets 2 and 3 indicates that the SVM decision boundary has a big margin to accommodate more misclassification. As a result, the classification accuracy is lower than dataset 1 with 86.6% and 69.6%, respectively. The findings showed that the recommended firefly-SVM classification model outperformed other existing models in terms of prediction accuracy for automated SVM parameter selection.

Statistical analysis

Statistical measure heatmap or correlation matrix is often employed in identifying the association between the clinicopathological parameters and their significance in breast cancer classification. Heatmaps are plotted to visualise the amount of dependency between the clinicopathological parameters and are denoted by colours of varying intensity. Correlation heatmaps are depicted to analyse the association between the clinicopathological attributes, that is, the positive and negative association among all the clinicopathological features and are represented by the blue and red colours, respectively. Larger correlation magnitude is identified with stronger colour shades. The diagonal in the heatmap is shaded with dark blue colour, which indicates the correlation of the same variable with itself. In python, the seaborn package is used to draw heatmap. The correlation heatmap on dataset 1, 2 and 3 are delineated in Figures 7, 8 and 9, respectively. The heatmap on dataset 1 reveals that the menopausal state, comorbidity, nutritional status and hypertension are all strongly positively correlated with age. Apart from these, there also exists a relationship

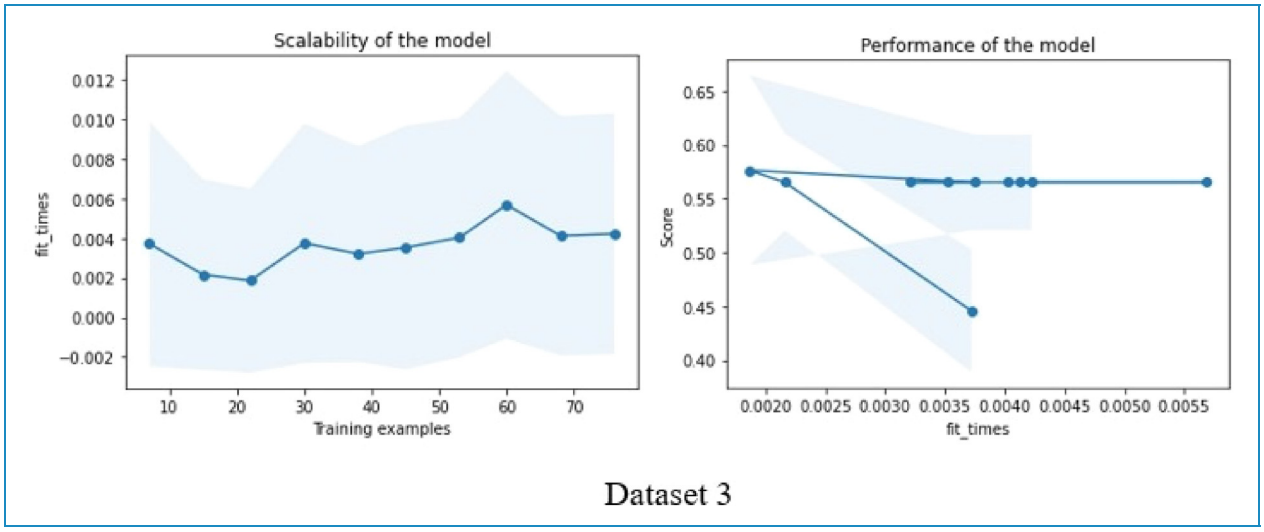


Figure 6. Scalability and performance of firefly-SVM predictive model on dataset 3. The scalability of the model denotes the time required by the model to fit the estimator with the training dataset. The blue-shaded region in the scalability graph indicates the region of fit_times mean +&- fit_times standard deviation. The performance of the model represents the test score with respect to fit_times. The blue-shaded region indicates the region of test scores mean + & - test scores standard deviation.

Table 2. Classification accuracy of firefly-SVM model with other existing models on datasets 1, 2 and 3.

	Classification accuracy		
Models	Dataset 1	Dataset 2	Dataset 3
Firefly-SVM	93.4	86.6	69.6
GA-SVM	91.4	82.0	53.8
PSO-SVM	91.4	82.0	53.8
Grid-SVM	90.3	82.3	50.0

Dataset 1= Lagos Teaching University, Nigeria, breast cancer dataset.
Dataset 2= National Institute of Oncology, Rabat, Morocco, breast cancer dataset.
Dataset 3= NKP Salve Institute of Medical Sciences and Research Centre, Nagpur.

between histology type, disease stage and metastasis. Age, menopause, the number of full-time pregnancies, hormone therapy and lymph nodes, tumour size with surgery type and tumour advancement are all factors that positively correlate in dataset 2. Strength of association also prevails between size at presentation, duration in months and quadrant on the heatmap of dataset 3. Further, there is also a strong correlation between tumour size and surgery type, radiotherapy and adjuvant chemotherapy, as well as nodal stage and TNBC/non-TNBC.

The Pearson's Chi-square test is another statistical method for evaluating the relationship between the clinicopathological characteristics of breast cancer patients. The

Table 3. Optimised hyperparameter values like kernel, *C* and gamma of SVM on firefly-SVM model and other classic hybrid models on dataset 1.

Models	Kernel	*C*	Gamma
FA-SVM	rbf	177.82	0.000177
GA-SVM	rbf	31.62	0.001
PSO-SVM	rbf	5623.41	5.62
Grid-SVM	rbf	10	0.001

Dataset 1= Lagos Teaching University, Nigeria, breast cancer dataset.

square of the difference between each categorical parameter's actual value and expected value divided by that parameter's expected value determines the chi-square statistics.

$$x^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \cdots + \frac{(O_{mn} - E_{mn})^2}{E_{mn}} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{6}$$

Here, τ denotes chi-square value, O_{ij} = observed value and E_{ij} = expected value of the parameters. It attempts to determine whether a feature's observed value differs from its expected value by chance or because of a real relationship that exists between them. The degree of freedom that varies with the number of features label and the class label is adjusted in the chi-square statistics. The chi-square

Table 4. Optimised hyperparameter values like kernel, *C* and gamma of SVM on firefly-SVM model and other classic hybrid models on dataset 2.

Models	Kernel	C	Gamma
FA-SVM	linear	0.38	
GA-SVM	rbf	8.25	0.21
PSO-SVM	rbf	8.25	0.21
Grid-SVM	rbf	100	0.001

Dataset 2= National Institute of Oncology, Rabat, Morocco, breast cancer dataset.

Table 5. Optimised hyperparameter values like kernel, *C* and gamma of SVM on firefly-SVM model and other classic hybrid models on dataset 3.

Models	Kernel	C
FA-SVM	sigmoid	1.0
GA-SVM	sigmoid	1.0
PSO-SVM	sigmoid	1.0
Grid-SVM	sigmoid	1000

Dataset 3= NKP Salve Institute of Medical Sciences and Research Centre, Nagpur.

test was carried out using Python version 3.11.2, and the output values include the chi-square score, chi-square *p*-value, *F*-score, *F*-score *p*-value and mutual information within the clinicopathological parameters. According to the findings from dataset 1, clinicopathological characteristics such as patients' height, family history of breast cancer, body mass index, comorbidities allergies and hormone receptor status were shown to be statistically significant ($p < 0.05$) in differentiating TNBC from non-TNBC cases. Clinically significant ($p < 0.05$) clinicopathological factors for classification in dataset 2 were hormone therapy and progression (metastasis/relapse). The chi-square test of dataset 3 revealed that the clinicopathological parameters duration in months, axillary mass, duration of breastfeeding in months, tumour stage and node stage were statistically relevant ($p < 0.05$). The statistical analysis results justify the aggressive nature of breast cancer, which are characterised with metastasis/relapse, hormone therapy, hormone receptor status and other risk factors. The detailed comparative analyses of clinicopathological features between TNBC and non-TNBC subgroups from statistical perceptive were explored in the original studies.[30,32,34]

Discussion

The AUROC, precision–recall curve, MSE, logarithmic loss, F1-score and learning curve were used to quantify the performance of the firefly-SVM hybrid model. The results demonstrated improved classification accuracy when compared to the hyperparameter tuning models of GA-SVM, PSO-SVM, and the conventional Grid-SVM model. The statistical analysis has also identified the significant risk factors establishing the aggressivity of breast cancer. Hence, the hybrid model's overall performance illustrates the potency of classifying patient groups into those with and without TNBC. Similar type of study employing SVM-FFA hybrid model has been reported by Sudheer Ch et al.[50] for estimating the malaria incidence in Bikaner and Jodhpur districts of Rajasthan, India. A. Kazem et al.[51] applied chaotic firefly algorithm for optimising SVR hyperparameters in stock price prediction. These studies had adopted radial basis kernel function as a basis for hyperparameter tuning of SVM, while our present study have contemplated different kernel functions as alternatives aiming to achieve the best kernel function regardless of choice. Huang et al.[52] assessed the capability of SVM and SVM-ensembles for breast cancer prediction on large- and small-scale datasets. Besides these, various models of firefly algorithm and SVM have been developed and applied in diverse domain like solar radiation prediction,[53] forecasting soil temperature at several depths[54] and fraud detection.[55] Our study is likely the first to report on the automated generation of SVM parameters with a firefly-SVM predictive model to classify patient groups into TNBC and without TNBC based on clinicopathological criteria. The advantage of using hybrid model lies in unifying the complementary parameters of all models involved, thereby reducing the weakness incurred by the individual classifiers.[56] The use of ML technologies in the analysis of medical data has become indispensable due to high-dimensionality and heterogeneity in medical datasets. To tackle such complexity, hybrid ML models have started to appear in literature. Taghizadeh et al.[57] employed an integrated machine learning approach (HMLA) that entails an in-depth search for the identification of the best HMLAs, comprising feature selection methods, a feature extraction technique and classifiers for breast cancer diagnosis. Hybrid ML models reporting the fusion of firefly algorithm and SVM for breast cancer prediction can be found in literature.[58–61] In medical sciences, breast cancer subtype classification traditionally deals with immunohistochemical staining, imaging and radiomics.[62–65] However, application of ML hybrid models with improved classification accuracy provides a framework for categorising TNBC vs. non-TNBC tumours effectively and can be recognised as an alternative, complementing the medical procedures.

According to GLOBOCAN 2020,[66] 186,598 new cases of breast cancer and 85,787 cases of high mortality have been reported in Africa. Moreover, breast cancer was

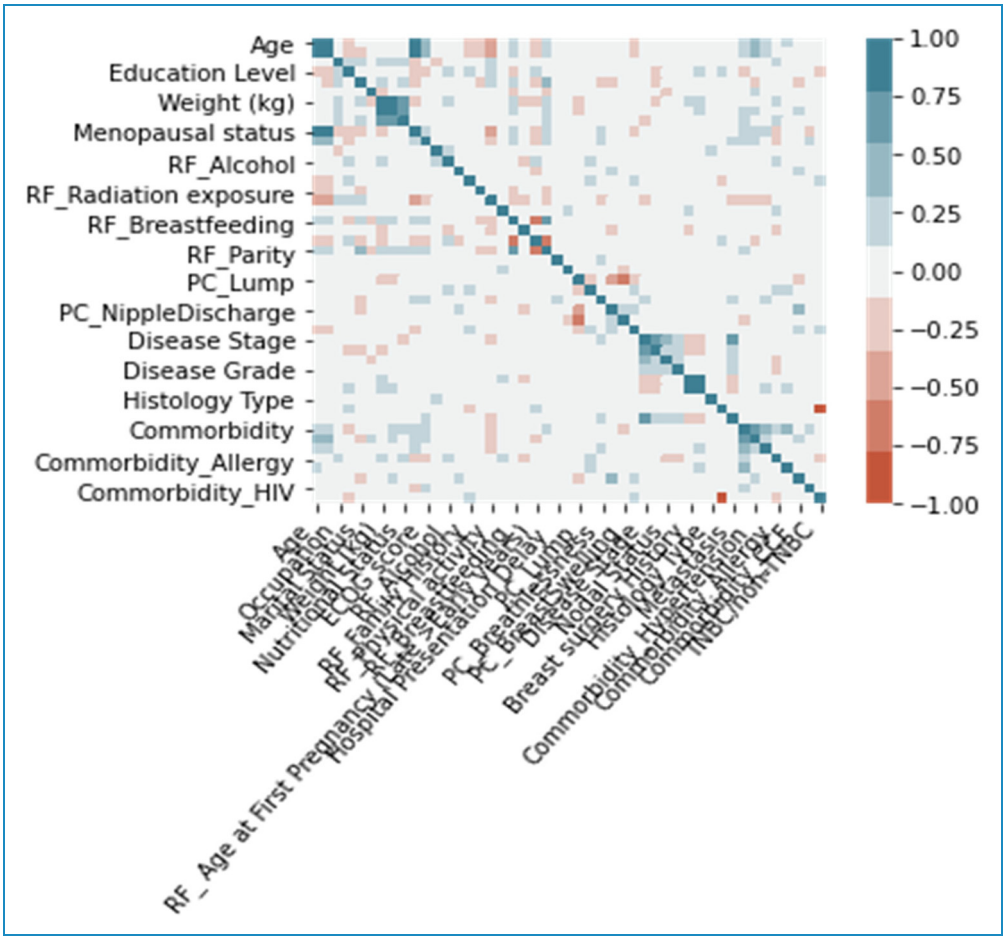


Figure 7. The correlation heat map of dataset 1. The higher correlation value among the clinicopathological parameters was indicated with the stronger colour shades. The dark blue colour heatmap diagonal signifies the correlation of the same variable with itself.

recorded as the most prevalent cancer in Africa ahead of cervix uteri. The breast cancer incidence among the females was estimated at 531,086 cases with 74.3 per 100,000 women, while the death count was 19.4 per 100,000, which shows the significant breast cancer burden in Africa. TNBC, a destructive variant of breast cancer, is characterised by high incidence among premenopausal younger females, associated with aggressive tumour, greater chance of recurrence within the first 3 years and low survival rates upon metastasis. Being chemo-sensitive in nature, surgery in combination with chemotherapy is often considered as the available treatment modalities despite the fact that there are no specific targeted medicines that have been approved by the Food and Drug Administration. This impels us to classify TNBC vs. non-TNBC breast cancer subgroups in African countries.

With the extensive use of bioinformatics, statistics and ML tools, there has been an intensive development in knowledge-based diagnostic system for cancer identification. Future outcomes of breast tumours can be precisely predicted through ML, which has the ability to recognise, discover patterns, and develop relationships from complicated medical data. Moreover, application of ML techniques in cancer classification provides an informative base for cancer prognosis and prediction that can easily be precisely tested within a short time.[67] Recently, many research articles pertaining to ML application in molecular classification of breast tumour have started to come out.[68–70] By understanding the breast cancer molecular subtypes, doctors can decide for appropriate treatment, thereby preventing the side effects of unnecessary medication as well as saving the financial burden of patient's party.[71] Due to digitisation of medical records and its availability to medical practitioners, the medical decision-making paradigm has now been shifted to data-driven diagnostic systems. There lies a necessity of maintaining privacy issues related to electronic health care data in clinical practise before applying ML. Using ML in breast cancer predictive analysis also supports proper treatment plans, clinical patient assessment, determining surgery methods and necessary adjuvant therapies. ML is found to provide satisfactory results in clinical patient management.[72–74] This compels us to develop an integrated ML model for classifying breast cancer with clinicopathological features

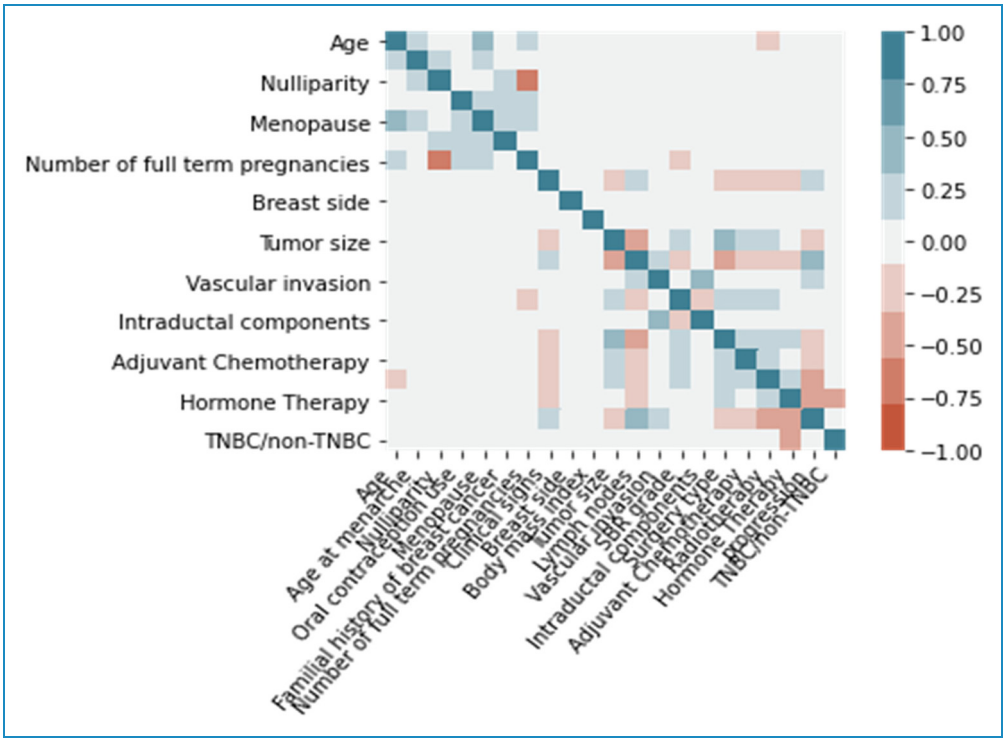


Figure 8. The correlation heat map of dataset 2. The higher correlation value among the clinicopathological parameters was indicated with the stronger colour shades. The dark blue colour heatmap diagonal signifies the correlation of the same variable with itself.

of patients possessing breast cancer in tertiary care hospitals or oncological centres.

The diagnosis and identification of breast cancer using microarray gene expression profiling has received substantial research and is considered as the golden method,[75] but the dynamic characteristics of genes within an individual may cause misclassification errors, which results to its incapability to accurately classify data into various molecular subtypes.[76–78] Numerous imaging techniques have been shown to be quite beneficial for the early diagnosis of breast cancer. In early stages, digital mammography stands as the most popular breast cancer screening and diagnostic tool.[79,80] The radiation risks from digital mammography, however, are dangerous to the patient's body and could potentially increase their chance of breast cancer.[81] The major drawback of mammography is its incompetency of characterising reducing sensitivity and low specificity in younger women with thick breast. In clinical practise, ultrasonography has recently gained popularity as a mammography alternative.[82–84] Another useful tool widely used in clinical practise is the Breast Imaging Reporting and Data System (BI-RADS). However, due to physicians' varying levels of experience and subjective dependencies, there is still a high rate of misinterpretation in the therapeutic setting. As a result, the use of computer-aided diagnosis (CAD) systems can help the clinicians to make more accurate diagnoses of breast tumours. In recent years, numerous CAD methods for the early diagnosis of breast cancer have been put forth.[85–87] SVM has been employed as an effective classifier in several of these CAD systems.[88–90] Breast tumours were classified as benign or malignant applying SVM with 28 textural features on an ultrasound image by Huang et al.[91] To automatically identify and categorise tumours, a novel CAD system based on stepwise regression selection of features and fuzzy SVM was developed.[92] The majority of data-driven diagnostic systems developed with SVM for the detection and investigation of breast lumps are centred on ultrasound imaging, while the focus of the current study is on setting up a data-driven diagnostic model that incorporates clinicopathological and demographic information collected from multiple tertiary care cancer hospitals or oncological centres for the classification of breast cancer subtypes.

Our present investigation with firefly-SVM hybrid model plays a pivotal role in identification of lethal variants of breast cancer TNBC, which are to be prioritised with better treatment regimen apart from classifying breast cancer subtypes with higher prediction accuracy. The significance of the present study also lies in validation with multicentric datasets of diverse geography, which can be look upon as a lacuna in previously reported studies. Lastly, hybridised ML models for classification of breast cancer with hospital-based patient's dataset consisting of demographic, clinical and pathological risk factors are rarely reported in literature. Hybridisation of firefly-SVM model has led to data-driven diagnostic system, which

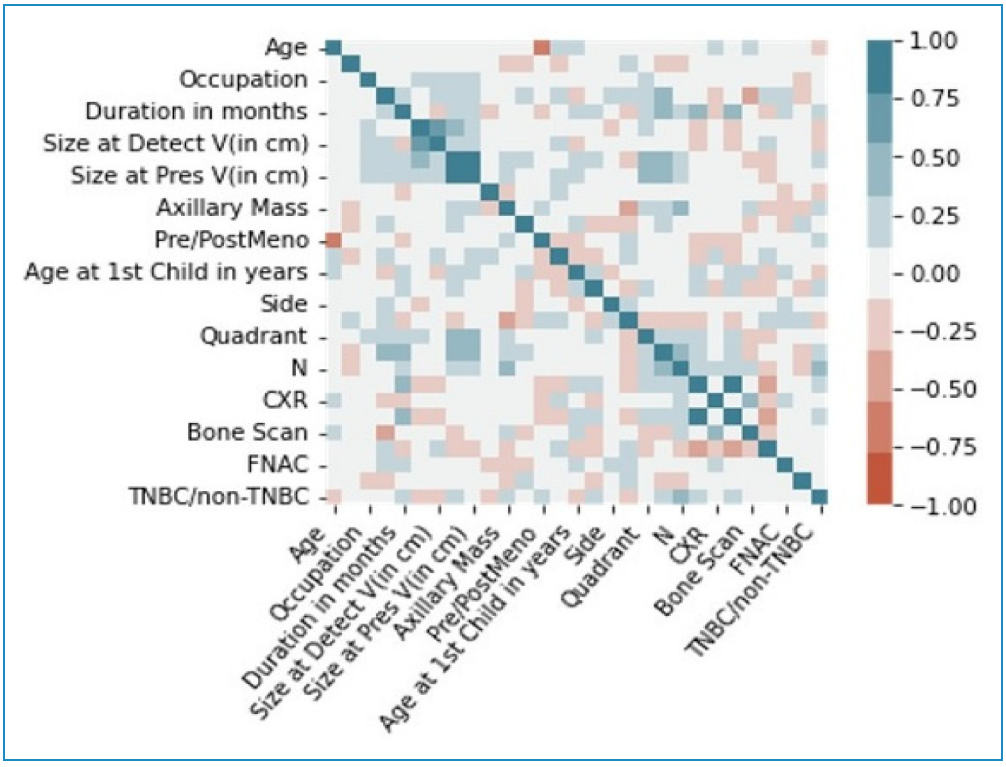


Figure 9. The correlation heat map of dataset 3. The higher correlation value among the clinicopathological parameters was indicated with the stronger colour shades. The dark blue colour heatmap diagonal signifies the correlation of the same variable with itself.

could assist the medical practitioners for clinical assessment of patients and admissible treatment lay-out.

Our hybrid model has categorised breast cancer into TNBC and non-TNBC subtypes, but in reality, there exist several variants of TNBC and non-TNBC subgroup that have not been explored in this research. Analysis of clinicopathological breast cancer characteristics from hospital-gathered datasets yields smaller dataset size. The possible reason for this paucity of hormone receptor data might be the financial hindrance of patients in less economically developed countries of Africa and India. This study has performed analysis on three datasets of which National Institute of Oncology's breast cancer datasets in Morocco constitute an unbalanced design of TNBC cases, and the remaining two datasets, namely Nigeria and India datasets, have almost balanced TNBC cases of 47.4% and 43.7%, respectively. Appropriate technique for handling the imbalance dataset may be incorporated in future work. Additionally, the effects on the various patient populations' varying demographics were not taken into account. Hence, the possible shortcomings of the study have been realised.

Conclusion

The current firefly-SVM predictive model can be used as an alternative approach for correctly classifying TNBC and non-TNBC subgroups of breast cancer, which would also help the health care professionals to manage the patients with the best possible treatment and better diagnostic results. Combining several risk factors in breast cancer prediction modelling might aid in the early identification of the disease and the development of essential treatment regimens. Since predictive models can quickly identify high-risk individuals using known clinical and demographic risk factors, they are vital for personalised medications. Investigation of more predictive models is recommended for improvement in prediction and accuracy to identify personalised medication for the dreadful variant of breast cancer TNBC.

Acknowledgements: The authors are grateful to Dr Murtaza Akhtar, Ex-Professor and Head, Department of Surgery, N.K.P Salve Institute of Medical Sciences, Nagpur, India, for supplying the required data to conduct this research.

Contributorship: SS was involved in the conceptualisation, investigation, methodology, and writing—original draft. KM was involved in the formal analysis, investigation, supervision, and correction of the original draft. All authors met the requirements as outlined by the ICMJE guidelines for co-authorship, and all co-authors have reviewed and approved the final manuscript.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This article does not require any ethical approval.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

Guarantor: SS.

ORCID iD: Suvobrata Sarkar https://orcid.org/0000-0002-1902-1050

Supplemental material: Supplemental material for this article is available online.

References

1. Giaquinto AN, Sung H, Miller KD, et al. Breast cancer statistics, 2022. *CA Cancer J Clin* 2022; 72: 524–541.
2. Dietze EC, Sistrunk C, Miranda-Carboni G, et al. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat Rev Cancer* 2015; 15: 248–254.
3. Gonçalves H, Guerra MR, Cintra JRD, et al. Survival study of triple-negative and non-triple-negative breast cancer in a Brazilian cohort. *Clin Med Insights Oncol* 2018; 12: 1–10.
4. Nedeljkovíc M and Damjanovíc A. Mechanisms of chemotherapy resistance in triple-negative breast cancer-how we can rise to the challenge. *Cells* 2019; 8: 957.
5. Vapnik V and Lerner A. Pattern recognition using generalized portrait method. *Autom Remote Control* 1963; 24: 774–780.
6. Wu J and Hicks C. Breast cancer type classification using machine learning. *J Pers Med* 2021; 11: 61.
7. Huang S, Chai N, Pacheco PP, et al. Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* 2018; 15: 41–51.
8. Ferroni P, Zanzotto FM, Riondino S, et al. Breast cancer prognosis using a machine learning approach. *Cancers* 2019; 11: 28.
9. 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: 230–238.
10. Bai F, Wei C, Zhang P, et al. Use of peripheral lymphocytes and support vector machine for survival prediction in breast cancer patients. *Transl Cancer Res* 2018; 7: 978–987.
11. Mihaylov I, Nisheva M and Vassilev D. Application of machine learning models for survival prognosis in breast cancer studies. *Information* 2019; 10: 93.
12. Beni G and Wang J. Swarm intelligence in cellular robotic systems. In: *Robots and biological systems: towards a new bionics*. Berlin Heidelberg: Springer, 1993, pp.703–712.
13. Kennedy J and Eberhart R. The particle swarm optimization: social adaptation in information processing. In: Corne D, Dorigo M and Glover F (eds) *New ideas in optimization*. London, UK: McGraw Hill, 1999, pp.379–387.
14. Dorigo M and DiCaro G. The ant colony optimization metaheuristic. In: Corne D, Dorigo M and Glover F (eds) *New ideas in optimization*. London, UK: McGraw Hill, 1999, pp.11–32.
15. Korošec P, Šilc J and Filipič B. The differential ant-stigmergy algorithm. *Inf Sci* 2012; 192: 82–97.
16. Karaboga D and Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J Global Optimiz* 2007; 39: 459–471.
17. Fister I, Fister I, Brest J, et al. Memetic artificial bee colony algorithm for large-scale global optimization. In: *IEEE congress on evolutionary computation*. CEC. DOI: 10.1109/CEC.2012.6252938.
18. Yang XS and Deb S. Cuckoo search via levey flights. In: *World congress on nature and biologically inspired computing*. NABIC 2009 - Proceedings: 2009, pp.210–214. DOI: 10.1109/NABIC.2009.5393690.
19. Yang XS. A new metaheuristic bat-inspired algorithm. In: Gonzlez J, Pelta D, Cruz C, Terrazas G and Krasnogor N (eds) *Nature inspired cooperative strategies for optimization (NISCO2010). Vol. 284 of studies in computational intelligence*. Berlin Heidelberg: Springer, 2010, pp.65–74.
20. Yang XS. *Firefly algorithm, nature-inspired metaheuristic algorithms*. 2nd Edition. UK: Luniver Press, 2008.
21. Gandomi AH and Alavi AH. Krill herd: a new bio-inspired optimization algorithm. *Commun Nonlinear Sci Numer Simulat* 2012; 17: 4831–4845.
22. Hatamlou A, Abdullah S and Nezamabadi -pour H. A combined approach for clustering based on k-means and gravitational search algorithms. *Swarm Evol Comput* 2012; 6: 47–52.
23. Hatamlou A. Black hole: a new heuristic optimization approach for data clustering. *Inf Sci* 2013; 222: 175–184.
24. Jain M, Maurya S, Rani A, et al. Owl search algorithm: a novel nature-inspired heuristic paradigm for global optimization. *J Intell Fuzzy Syst: Appl Eng Technol* 2018; 34: 1573–1582.
25. Menga Z and Pan JS. Monkey king evolution: a new memetic evolutionary algorithm and its application in vehicle fuel consumption optimization. *Knowl Based Syst* 2016; 97: 144–157.
26. Heidari AA, Mirjalili S, Faris H, et al. Harris hawks' optimization: algorithm and applications. *Future Gener Comput Syst* 2019; 97: 849–872.
27. Črepinšek M, Mernik M and Liu S. Analysis of exploration and exploitation in evolutionary algorithms by ancestry trees. *Int J Innovative Comput Appl* 2011; 3: 11–19.
28. Yang XS. Firefly algorithms for multimodal optimization. In: Watanabe O, Zeugmann T (eds) *Stochastic algorithms: foundations and applications*. Berlin Heidelberg: Springer, 2009, vol. 5792, pp.169–178.
29. https://www.ebi.ac.uk/biostudies/.
30. Mouh FZ, Slaoui M, Razine R, et al. Clinicopathological, treatment and event-free survival characteristics in a Moroccan population of triple-negative breast cancer. *Breast Cancer (Auckl)* 2020; 14: 1–10.
31. *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/SEPMC7218339?Query=Clinicopathological%2C%20Treatment%20and%20EventFree%20Survival%20Characteristics%20in%20a%20Moroccan%20Population%20of%20TripleNegative%20Breast%20Cancer%20Fatima%20Zahra%20Mouh.
32. 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: 165–172.

33. *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.
34. Akhtar M, Dasgupta S and Rangwala M. Triple negative breast cancer: an Indian perspective. *Breast Cancer (Dove Med Press)* 2015; 7: 239–243.
35. Zhao M, Fu C, Ji L, et al. Feature selection and parameter optimization for support vector machines: a new approach based on genetic algorithm with feature chromosomes. *Expert Syst Appl* 2011; 38: 5197–5204.
36. Aich U and Banerjee S. Modeling of EDM responses by support vector machine regression with parameters selected by particle swarm optimization. *Appl Math Model* 2014; 38: 2800–2818.
37. Korovkinas K, Danenas P and Garšva G. Support vector machine parameter tuning based on particle swarm optimization metaheuristic. *Nonlinear Anal: Model Control* 2020; 25: 266–281.
38. Goldberg DE and Holland JH. Genetic algorithms and machine learning. *Mach Learn* 1988; 3: 95–99.
39. Kennedy J and Eberhart R. Particle swarm optimization. *Int Conf Neural Netw ICNN'95* 1995; 4: 1942–1948.
40. Syarif I, Prugel-Bennett A and Wills G. SVM Parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika* 2016; 14: 1502.
41. Wang H, Cui Z, Sun H, et al. Randomly attracted firefly algorithm with neighborhood search and dynamic parameter adjustment mechanism. *Soft Comput* 2017; 21: 5325–5339.
42. Marichelvam MK, Prabaharan T and Yang XS. A discrete firefly algorithm for the multi-objective hybrid flow shop scheduling problems. *IEEE Trans Evol Comput* 2014; 18: 301–305.
43. Marichelvam MK and Geetha M. A hybrid discrete firefly algorithm to solve flow shop scheduling problems to minimize total flow time. *Int J Bio-Insp Comput* 2016; 8: 318–325.
44. Yang XS. Firefly algorithm, levy flights and global optimization. In: Bramer M, Ellis R and Petridis M (eds) *Research and development in intelligent systems*. London: Springer, 2010, pp.209–218.
45. *Scikit-learn*. Machine Learning in Python. Available at: https://scikit-learn.org/stable/.
46. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010; 5: 1315–1316.
47. Sivapragasam C, Liong SY and Pasha F. Rainfall and runoff forecasting with SSA-SVM approach. *J Hydroinform* 2001; 3: 141–152.
48. Choy K and Chan C. Modelling of river discharges and rainfall using radial basis function networks based on support vector regression. *Int J Sys Sc* 2003; 34: 763–773.
49. Yu X, Liong S and Babovic V. EC-SVM approach for real-time hydrologic forecasting. *J Hydroinform* 2004; 6: 209–223.
50. Chintalapati S, Sohani SK, Kumar D, et al. A support vector machine-firefly algorithm based forecasting model to determine malaria transmission. *Neurocomputing* 2014; 129: 279–288.
51. Kazem A, Sharifi E, Hussain FK, et al. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Appl Soft Comput* 2013; 13: 947–958.
52. Huang MW, Chen CW, Lin WC, et al. SVM And SVM ensembles in breast cancer prediction. *PLoS One* 2017; 12: e0161501.
53. Olatomiwa L, Mekhilef S, Shamshirband S, et al. A support vector machine–firefly algorithm- based model for global solar radiation prediction. *Sol Energy* 2015; 115: 632–644.
54. Shamshirband S, Esmaeilbeiki F, Zarehaghi D, et al. Comparative analysis of hybrid models of firefly optimization algorithm with support vector machines and multilayer perceptron for predicting soil temperature at different depths. *Eng Appl Comput Fluid Mech* 2020; 14: 939–953.
55. Singh A, Jain A and Biable SE. Financial fraud detection approach based on firefly optimization algorithm and support vector machine. *Appl Comput Intell Soft Comput* 2022; 2022: 1–10.
56. Castillo W, Melin O and Pedrycz P. Hybrid intelligent systems: Analysis and design. In: *Studies in fuzziness and soft computing*. Berlin Heidelberg: Springer, 2007, pp.55–64.
57. Taghizadeh E, Heydarheydari S, Saberi AH, et al. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinform* 2022; 23: 10.
58. Gomathi B and Sujatha R. Prediction of breast cancer using data fusion of SVM with optimization technique. *J Inf Comput Sci* 2021; 9: 504–511.
59. Chao CF and Horng MH. The construction of support vector machine classifier using the firefly algorithm. *Comput Intell Neurosci* 2015; 2015: 1–8.
60. Thawkar S and Ingolikar R. Classification of masses in digital mammograms using firefly based optimization. International journal of image. *Graphics Signal Process* 2018; 10: 25–33.
61. Sahmadi B, Boughaci D, Rahmani R, et al. A modified firefly algorithm with support vector machine for medical data classification. In: *6th IFIP international conference on computational intelligence and its applications (CIIA)*. Oran, Algeria, May 2018, pp.232–243. DOI: 10.1007/978-3-319-89743-1_21.
62. Fanizzi A, Basile TM, Losurdo L, et al. Ensemble discrete wavelet transform and gray-level co-occurrence matrix for microcalcification cluster classification in digital mammography. *Appl Sci* 2019; 9: 5388.
63. Losurdo L, Fanizzi A, Basile TMA, et al. Radiomics analysis on contrast-enhanced spectral mammography images for breast cancer diagnosis: a pilot study. *Entropy* 2019; 21: 1110.
64. Conti A, Duggento A, Indovina I, et al. Radiomics in breast cancer classification and prediction. *Semin Cancer Biol* 2021; 72: 238–250.
65. Forgia DA, Fanizzi A, Campobasso F, et al. Radiomic analysis in contrast-enhanced spectral mammography for predicting breast cancer histological outcome. *Diagnostics (Basel)* 2020; 10: 08.
66. Anyigba CA, Awandare GA and Paemka L. Breast cancer in sub-Saharan Africa: the current state and uncertain future. *Exp Biol Med (Maywood)* 2021; 246: 1377–1387.
67. Weston AD and Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 2004; 3: 179–196.
68. Montazeri M, Montazeri M, Montazeri M, et al. Machine learning models in breast cancer survival prediction. *Technol Health Care* 2016; 24: 31–42.
69. Wu T, Sultan LR, Tian J, et al. Machine learning for diagnostic ultrasound of triple-negative breast cancer. *Breast Cancer Res Treat* 2019; 173: 365–373.
70. 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: 41–52.

71. Tao M, Song T, Du W, et al. Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes (Basel)* 2019; 10: 00.
72. Mintz Y and Brodie R. Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol* 2019; 28: 73–81.
73. Qian Z, Li Y, Wang Y, et al. Differentiation of glioblastoma from solitary brain metastases using radiomic machine-learning classifiers. *Cancer Lett* 2019; 451: 128–135.
74. Tan A, Huang H, Zhang P, et al. Network-based cancer precision medicine: a new emerging paradigm. *Cancer Lett* 2019; 458: 39–45.
75. Peppercorn J, Perou CM and Carey LA. Molecular subtypes in breast cancer evaluation and management: divide and conquer. *Cancer Invest* 2008; 26: 1–10.
76. Gusterson B. Do 'basal-like' breast cancers really exist? *Nat Rev Cancer* 2009; 9: 128–134.
77. Pusztai L, Mazouni C, Anderson K, et al. Molecular classification of breast cancer: limitations and potential. *Oncologist* 2006; 11: 868–877.
78. Weigelt B, Baehner FL and Reis-Filho JS. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J Pathol* 2010; 220: 263–280.
79. Chen Z, Strange H, Oliver A, et al. Topological modeling and classification of mammographic microcalcification clusters. *IEEE Trans Biomed Eng* 2015; 62: 1203–1214.
80. Choi JY. A generalized multiple classifier system for improving computer-aided classification of breast masses in mammography. *Biomed Eng Lett* 2015; 5: 251–262.
81. Yin T, Ali FH and Reyes-Aldasoro CC. A robust and artifact resistant algorithm of ultrawideband imaging system for breast cancer detection. *IEEE Trans Biomed Eng* 2015; 62: 1514–1525.
82. Ungi T, Gauvin G, Lasso A, et al. Navigated breast tumor excision using electromagnetically tracked ultrasound and surgical instruments. *IEEE Trans Biomed Eng* 2016; 63: 600–606.
83. Sahiner B, Chan HP, Roubidoux MA, et al. Malignant and benign breast masses on 3D US volumetric images: effect of computer-aided diagnosis on radiologist accuracy. *Radiol* 2007; 242: 716–724.
84. Costantini M, Belli P, Lombardi R, et al. Characterization of solid breast masses - Use of the sonographic breast imaging reporting and data system lexicon. *J Ultrasound Med* 2006; 25: 649–659.
85. Zhou SC, Shi J, Zhu J, et al. Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image. *Biomed Signal Process Control* 2013; 8: 688–696.
86. Yang MC, Moon WK, Wang YCF, et al. Robust texture analysis using multi-resolution gray-scale invariant features for breast sonographic tumor diagnosis. *IEEE Trans Med Imag* 2013; 32: 2262–2273.
87. Cai LY, Wang X, Wang YY, et al. Robust phase-based texture descriptor for classification of breast ultrasound images. *Biomed Eng Online* 2015; 14: 26.
88. Huang QH, Yang FB, Liu LZ, et al. Automatic segmentation of breast lesions for interaction in ultrasonic computer aided diagnosis. *Inf Sci* 2015; 314: 293–310.
89. Huang YL, Chen DR, Jiang YR, et al. Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound. *Ultrasound Obstetrics Gynaecol* 2008; 32: 565–572.
90. Huang QH, Luo YZ and Zhang QZ. Breast ultrasound image segmentation: a survey. *Int J Comput Assisted Radiol Surg* 2017; 12: 493–507.
91. Huang YL, Wang KL and Chen DR. Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines. *Neural Comput Appl* 2006; 15: 164–169.
92. Shi XJ, Cheng HD, Hu LM, et al. Detection and classification of masses in breast ultrasound images. *Digital Signal Process* 2010; 20: 824–836.