

Enhancing breast cancer diagnosis: a comparative analysis of feature selection techniques

Salsabila Benghazouani¹, Said Nouh¹, Abdelali Zakrani²

¹Department of Mathematics and Computer Science, Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco

²Department of Computer Science Engineering, ENSAM, Hassan II University, Casablanca, Morocco

Article Info

Article history:

Received Jan 30, 2024

Revised Mar 2, 2024

Accepted Mar 20, 2024

Keywords:

Breast cancer

Classification

Feature selection

Genetic algorithm

Particle swarm optimization

ABSTRACT

Breast cancer is a significant contributor to female mortality, emphasizing the importance of early detection. Predicting breast cancer accurately remains a complex challenge within medical data analysis. Machine learning (ML) algorithms offer valuable assistance in decision-making and diagnosis using medical data. Numerous research studies highlight the effectiveness of ML techniques in improving breast cancer prediction. Feature selection plays a pivotal role in data preprocessing, eliminating irrelevant and redundant features to minimize feature count and improve classification accuracy. This study focuses on optimizing breast cancer diagnostics through feature selection methods, specifically genetic algorithms (GA) and particle swarm optimization (PSO). The research involves a comparative analysis of these methods and the application of a diverse set of ML classification techniques, including logistic regression (LR), support vector machine (SVM), decision tree (DT), and ensemble methods like random forest (RF), AdaBoost, and gradient boosting (GB), using a breast cancer dataset. The models' performance is subsequently evaluated using various performance metrics. The experimental findings illustrate that PSO achieved the highest average accuracy, reaching 99.6% when applied to AdaBoost, while GA attained an accuracy rate of 99.5% when employed with both AdaBoost and RF.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Salsabila Benghazouani

Department of Mathematics and Computer Science, Faculty of Sciences Ben M'Sick, Hassan II University
Casablanca, Morocco

Email: bengahouani.salsabila239@gmail.com

1. INTRODUCTION

In 2020, the number of diagnosed breast cancer cases surpassed those of lung cancer, establishing it as the most prevalent form of cancer. Female breast cancer has witnessed 2.3 million new cases reported, alongside an escalation in the mortality rate [1]. This cancer type has profoundly affected women's lives, and its mortality rate can be mitigated through enhanced awareness, early detection, and diagnosis [2]. In modern healthcare, the substantial volume of diverse disease data generated plays a crucial role in facilitating analysis and predictive tasks.

Machine learning (ML) assumes a significant role within healthcare systems [3], greatly assisting doctors and pathologists in making precise predictions. This not only aids in averting additional medical expenses but also ensures the provision of appropriate treatment, with the potential to save lives through early detection [4]. Nevertheless, utilizing ML methods for breast cancer prediction presents a significant challenge in clinical data analysis [3]. Identifying the most effective features to differentiate patients from healthy individuals remains a key obstacle in the development of ML techniques for early breast cancer prediction.

Numerous studies have been conducted to detect and forecast breast cancer diagnoses. For instance, Minnoor and Baths [5] employed the random forest (RF) algorithm for training and underwent hyperparameter tuning to attain efficient and accurate breast cancer diagnosis. Furthermore, the performance of these RF models was compared against four other supervised learning techniques. The findings conclusively establish RF as the superior method for diagnosing breast cancer. Massari *et al.* [6] introduced an ontological model based on the decision tree (DT) method, showcasing its reliability in predicting breast cancer. By extracting discriminative rules from the DT algorithm, this model effectively distinguishes between malignant and benign breast cancer cases. These rules are seamlessly integrated into an ontological reasoner using the semantic web rule language (SWRL). The achieved prediction accuracy for this ontological model is notably high, standing at 97.10%. Similarly, in another study by Nemade and Fegade [4], a variety of ML classification techniques were applied to a breast cancer dataset and assessed using various performance metrics. The findings revealed that, among all models, both the DT and XGBoost classifiers achieved the highest accuracy of 97%, with the XGBoost classifier achieving the maximum area under the curve (AUC) score of 99.90%.

Several studies have investigated the effectiveness of diverse ML approaches. Kabiraj *et al.* [7] introduced a predictive system for breast cancer risk using both RF and XGBoost methods. According to Kaul and Sharma [8], four different ML algorithms were utilized, including DT classifiers, RF, K-nearest neighbors (KNN), and support vector machine (SVM). Among these, the SVM exhibited the highest classification accuracy, reaching 97% for diagnosing breast cancer in women. Additionally, Naji *et al.* [9] found that the SVM and RF algorithms achieved accuracy rates exceeding 96% when identifying malignant tumors.

Numerous researchers have delved into the significance of feature selection in enhancing the performance of various supervised ML methods. Dhanya *et al.* [10] employed an ensemble model with F-test feature selection to predict breast cancer. This research combines several supervised ML algorithms, including SVM, naive Bayes, and KNN, and integrates feature selection methods such as variance threshold and F-test to enhance the ensemble model's accuracy in breast cancer prediction. According to Alnowami *et al.* [11], a wrapper feature selection method is utilized to explore biomarkers for early breast cancer detection. By integrating three classification algorithms (SVM, RF, and DT) into a sequential backward selection model, the study identifies an optimal set of biomarkers for breast cancer prediction, with the SVM model being the primary contributor. Additionally, Birchha and Nigam [12] employed a back-propagation neural network (BPNN) as a ML model, utilizing the breast cancer Wisconsin original dataset (WBC). They also incorporated principal component analysis and dimension reduction techniques to enhance the performance of the model.

Based on the literature review provided in this section, and several other studies [13]–[15], the researchers emphasize the risks and challenges of breast cancer and the significance of feature selection in identifying the most impactful features to enhance the performance of various supervised ML methods for early breast cancer prediction. The particle swarm optimization (PSO) and genetic algorithms (GA) are metaheuristic techniques that have demonstrated success in solving optimization problems, especially in the realm of feature selection. These methods are widely utilized for their ability to efficiently identify informative subsets of features, thereby enhancing the performance of ML techniques. This article assesses the importance of feature selection, specifically focusing on PSO and GA, in enhancing the performance of different supervised ML methods for forecasting breast cancer and compares these methods with other feature selection techniques. Additionally, the study evaluates the predictive power of the ML techniques using both the original dataset and feature-selected dataset. We outline the contributions of this study as follows: i) investigating the effectiveness of feature selection techniques utilizing PSO and GA for breast cancer prediction and contrasting them with alternative feature selection methods; ii) utilizing various ML techniques and evaluating their performance to develop an optimal predictive model; iii) applying multiple feature selection techniques, such as PSO, GA, recursive feature elimination (RFE), and SelectFromModel, to each ML technique, and evaluating their performance based on precision, recall, F1-score, accuracy, AUC, and the percentage of features that have been removed or reduced from the initial set.

The following sections of the paper are arranged as follows: section 2 outlines the methodology, presenting the proposed approach for breast cancer prediction and detailing the research procedure. Section 3 offers a detailed discussion of the results obtained from various ML techniques, including an analysis using different performance metrics. Finally, section 4 encapsulates the paper, presenting its conclusions and offering insights into future perspectives.

2. MATERIALS AND METHODS

In this section, we will discuss the materials and methods employed to obtain our results. This section is subdivided into five subsections, covering aspects such as dataset description, the proposed methodology, feature selection techniques, algorithms employed for comparison, and performance metrics. Each subsection provides detailed insights into the specific methods and processes used.

2.1. Methodology

This section outlines the steps and research methodology applied to enhance breast cancer diagnosis through feature selection methods. Initially, The experimental process started with acquiring the chosen dataset from the UCI ML repository. Subsequently, data cleaning and preprocessing are performed to ensure that only relevant data is retained. Furthermore, we normalize the dataset and randomly split it into a training set (70% of the dataset) and a testing set (30% of the dataset). Afterwards, various ML techniques, including logistic regression (LR), AdaBoost, RF, SVM, gradient boosting (GB), and DT are trained on the training set and assessed on the original dataset using different metrics. To further improve outcomes, each feature selection method-PSO, GA, RFE, and SelectFromModel—was employed to identify significant features of breast cancer. Lastly, each ML model is trained on four feature-selected datasets, and their effectiveness is assessed and compared using various performance metrics such as accuracy, F1 score, recall, precision, and AUC. Additionally, a reduction rate is calculated to gauge the impact of feature selection on the dataset. Figure 1 illustrates the flowchart for the study.

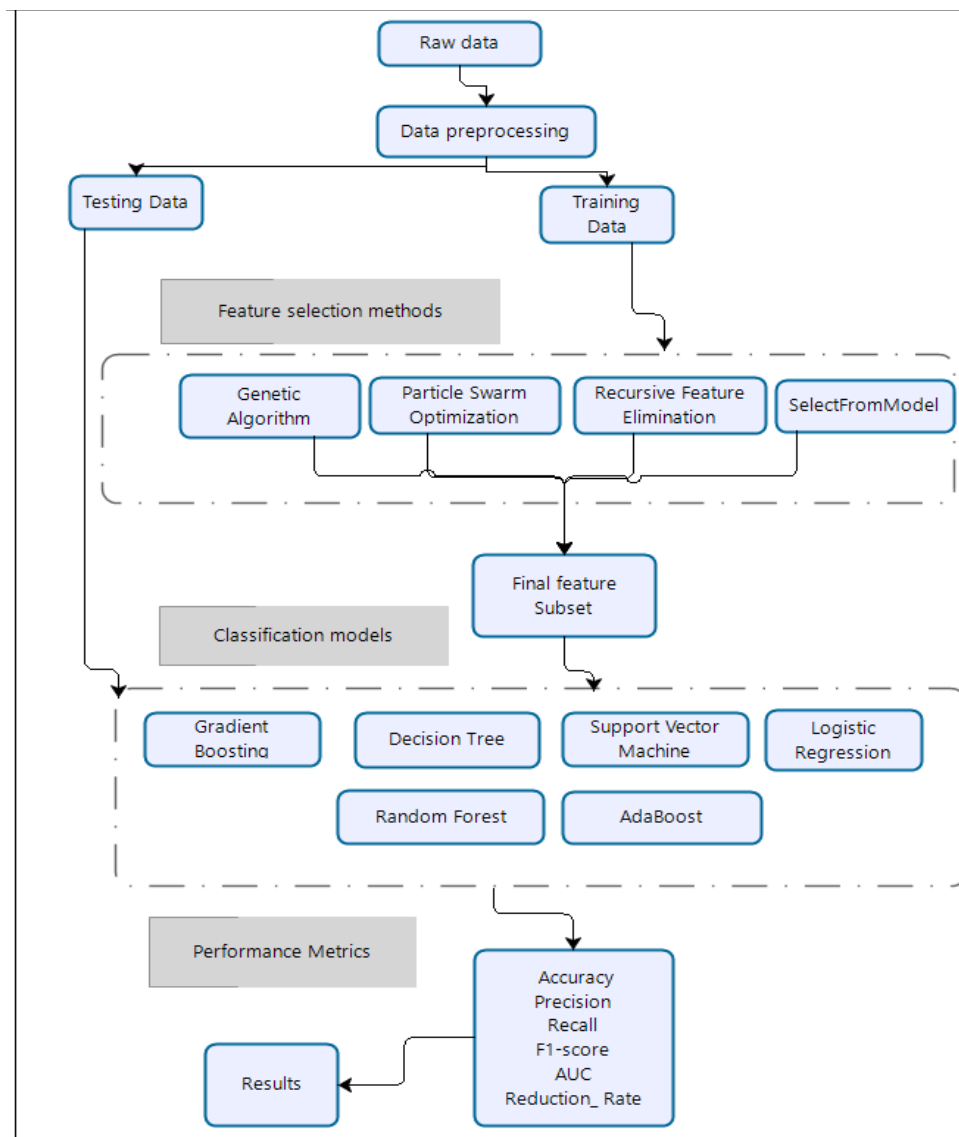


Figure 1. Proposed flow diagram for optimizing breast cancer diagnosis

2.2. Dataset description

A comparative and scientific examination was conducted on the Wisconsin diagnostic breast cancer (WDBC) datasets, which are openly accessible from the UCI ML repository. This dataset, initially provided

by the University of Wisconsin, includes 569 instances, comprising 212 categorized as malignant and 357 as benign. Each instance has 32 distinct characteristics, providing a comprehensive foundation for analysis and comparison.

2.3. Feature selection techniques

Feature selection represents a valuable step in data modeling, aimed at eliminating redundant, unimportant, and noisy data. The current study employs various feature selection techniques, which have been assessed and compared. These techniques enhance the model's performance by focusing on the most relevant data, ultimately improving the accuracy and efficiency of the analysis.

2.3.1. Particle swarm optimization-based feature selection

The PSO-based feature selection represents an evolutionary algorithm that relies on a population of particles, a concept introduced by Eberhart and Kennedy [16]. This approach draws inspiration from the collective movement of natural swarms in search of food. PSO employs both individual particle memory, tracking their personal best, and a global memory of the swarm to determine the most optimal motion [17].

In our PSO process utilized in this study, we follow several crucial stages:

- Particle representation: each particle denotes a possible solution and is defined by its position vector within the search space. Each element of the vector corresponds to a possible value of a feature.
- Initialization: the swarm is initialized with a group of particles; each assigned a random position within the search space. Additionally, each particle is given a random velocity.
- Fitness evaluation: the fitness of each particle is assessed according to its accuracy, favoring those particles that exhibit higher accuracy while utilizing fewer predictors.
- Particle movement: each particle adjusts its position and velocity based on its own best-known position (local best) and the global best-known position among all particles in the swarm.
- Updating best positions: the local best position and global best position are updated based on the fitness of particles.
- Termination conditions: the process ends when a specified number of iterations (100 in this study) is reached. Additionally, termination can occur if there's no improvement in the fitness function for a certain number of iterations.

2.3.2. Genetic algorithm-based feature selection

The GA-based feature selection draws inspiration from the principles of biological evolution, where the fittest individuals are chosen to generate the subsequent generation's offspring. In GA, this concept is applied by using genetic operators like crossover and mutation, with crossovers primarily recombining genetic material from the current population to discover new solutions. Mutation operators introduce novelty by altering existing data. GAs have proven to be highly effective in solving optimization problems, including feature selection challenges, and researchers have proposed various GA variants to tackle these specific problems, emphasizing the adaptability and utility of the GA approach [18]–[21].

The GA process used in this study involves several key stages:

- Individual encoding: chromosomes are represented as binary vectors, where each element denotes the presence or absence of a predictor in the dataset.
- Starting population: a binary matrix is formed with randomly chosen individuals, where rows represent potential solutions and columns correspond to available predictors.
- Fitness function: individuals' fitness is assessed based on accuracy, favoring those with higher accuracy and fewer predictors.
- Leveraging genetic operators for the formation of the next generation:
 - Selection: chromosome pairs for crossover are selected using the roulette wheel selection technique, with higher fitness individuals having a greater likelihood of being chosen.
 - Crossover: selected parent chromosomes exchange elements to create the next generation, with a probability parameter set at 0.8.
 - Mutation: random alterations of gene values in chromosomes promote exploration of the solution space and avoid convergence to local optima, with a uniform mutation probability set at 0.01.
- Termination conditions: the process ends either after a maximum number of generations (set to 100) or when there's no improvement in the fitness function for two consecutive generations.

2.3.3. SelectFromModel

SelectFromModel is a feature selection technique employed to extract vital and pertinent features. It eliminates features with importance values below a specified threshold. This method is compatible with estimators that possess significant features or coefficients [22].

2.3.4. Recursive feature elimination

The RFE enhances the effectiveness of ML models by systematically reducing feature sizes during model training. The RFE is a comprehensive approach that collaborates with learning algorithms like SVM and lasso. It achieves a systematic reduction of features by iteratively removing attributes with low weights [23].

2.4. Algorithms used for study

This section pertains to the ML methodologies employed in this study for the development of a predictive model for breast cancer. The algorithms selected for this study include SVM, LR, DT, GB, RF, and AdaBoost. Each algorithm was chosen based on its proven effectiveness in previous research and its suitability for this dataset.

2.4.1. Support vector machine

The SVM is a broadly adopted method for classification and regression, relying on support vector algorithms. Operating as a supervised learning method, SVM identifies a limited set of crucial representative samples from all categories and constructs a linear discriminant function aimed at achieving maximum separation. SVM effectively separates different groups into discrete categories through the utilization of multi-dimensional hyperplanes [24].

2.4.2. Logistic regression

The LR analysis serves as a fundamental and effective linear algorithm used for assessing multidimensional data and forecasting clinical outcomes, operating as a probabilistic predictive classification model. LR represents a supervised learning method specifically designed for tackling categorization challenges. It is capable of working with both continuous and discrete data, although it doesn't provide continuous output values [25]. LR relies on the sigmoid function to address classification problems.

2.4.3. Decision tree

The DT method, classified under supervised learning, is a versatile tool used for classification problems and can also handle regression tasks. It consists of inner nodes that describe branching structures, a dataset indicating the algorithm's verdict, and leaf nodes that represent outcomes. In this system, decision nodes are employed for making choices and have multiple branches, while leaf nodes provide the final decision outcome with no further branches. Its name is due to its tree-like structure, with a root node as the starting point, branching into sub-trees based on yes or no responses to questions [26].

2.4.4. Gradient boosting

The GB is an advanced prediction technique that sequentially tackles an infinite-dimensional convex optimization problem to create a model expressed as a linear combination of elementary predictors, often DT. This method initially constructs a model and subsequently enhances it by allowing the optimization of any differentiable loss function. It leverages a gradient-descent algorithm to minimize the loss associated with new trees. This technique applies to predictive modeling for both regression and classification tasks [27].

2.4.5. Random forest

The RF is a popular ensemble technique used for pattern recognition, comprising a collection of DT. Notably, it employs semi-random feature selection, making it capable of handling a large number of characteristics and identifying the most important ones. This approach, also referred to as bagging or bootstrap aggregation, combines the outputs from individual trees to provide a consolidated output. In contrast to a single DT, which exhibits low bias and high variance, the RF leverages a wealth of data to reduce variance, ultimately yielding improved results [28].

2.4.6. AdaBoost

AdaBoost stands as an exemplar of boosting ensemble learning techniques, as indicated [29]. Its methodology involves an iterative process that trains a sequence of weak learners, adjusting the sample weights based on regression error rates. This process focuses the subsequent learner's attention on the samples that performed poorly in previous iterations. Eventually, the learners receive weights corresponding to their regression error rates, and the final output is determined by taking a weighted average of the predictions [30].

2.5. Performance measures

The classification evaluation metrics employed include accuracy, precision, recall, F1-score, and ROC-AUC. These measures are derived from the elements of the confusion matrix, which conveys information about predicted and actual values. The performance metrics are expressed using (1)-(5).

Accuracy is a metric used to assess the percentage of accurate predictions, calculated by dividing the number of correct predictions by the total number of predictions. The formula for accuracy is provided as (1):

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

In this context, TP stands for true positives, TN represents true negatives, FP corresponds to false positives, and FN indicates false negatives.

Precision is a metric used to assess the quality of the ML model specifically in terms of positive predictions. It is calculated by dividing the number of true positive predictions by the total number of positive predictions. The formula for precision is as (2):

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

Recall is a metric that quantifies the ML model's capability to identify all pertinent instances in the provided dataset. It is computed by dividing the number of TP predictions by the sum of TP and FN predictions. The formula for recall is presented as (3):

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

The F1-score, also known as the F-measure, serves as a metric for the weighted average that combines the concepts of recall and precision. It offers an accurate assessment of a model's predictive ability in scenarios involving both balanced and imbalanced datasets. The formula for calculating the F1-score is as (4):

$$F1 - score = 2 * \frac{(Precision*Recall)}{(Precision+Recall)} \quad (4)$$

Receiver operating characteristic (ROC)-AUC, which combines the ROC curve and AUC, serves as a measure of the extent of separation in ML model. It quantifies the model's ability to differentiate between distinct classes. The calculation involves dividing the TP rates by the FP rates. The formulas for this calculation are presented as (5):

$$ROC - AUC = \frac{TPR}{FPR} \quad (5)$$

In this context, TPR represents the true positive rate, while FPR corresponds to the false positive rate. The reduction rate of features from the initial feature set determines the proportion of features that have been eliminated or reduced from the original set by subtracting the number of selected features from the initial total features and then dividing this difference by the initial number of features.

3. RESULTS AND DISCUSSION

In this section, we assess the performance of various feature selection methods using multiple classifiers, presenting results in terms of accuracy, precision, recall, F1 score, AUC, and the reduction rate of features from the original set. Each feature selection method is executed five times during each experiment, and we compare the approaches based on the averages of these runs. All these procedures are conducted using Python on a system equipped with an Intel Core i7 CPU and 16 GB of RAM.

3.1. The performance of machine learning models

The initial experiment aims to assess the performance of each algorithm using all attributes available in the dataset for the classification task. Table 1 presents classification results for multiple ML methods without employing feature selection. These results offer insights into the performance of each method in terms of accuracy, precision, recall, F1 score, and AUC. As shown in Table 1, the both AdaBoost and SVM emerge as the top performers across multiple metrics when compared to other ML models. AdaBoost achieves an accuracy of 96.60%, with precision, recall, F1-score, and AUC of 96.20%, 97.00%, 96.59%, and 96.63% respectively. Similarly, SVM achieves slightly higher accuracy at 96.70% and precision at 97.00%, with recall, F1-score, and AUC of 96.50%, 96.72%, and 96.75% respectively.

Table 1. Classification result without feature selection

Method	Accuracy	Precision	Recall	F1-score	AUC
LR	95.60	95.80	95.42	95.61	95.61
AdaBoost	96.60	96.20	97.00	96.59	96.63
SVM	96.70	97.00	96.50	96.72	96.75
DT	95.50	98.20	93.22	95.62	95.68
GB	96.40	97.60	95.36	96.45	96.46
RF	96.30	96.20	96.40	96.30	96.30

3.2. Feature selection performance across machine learning models

The second experiment offers a comprehensive assessment of various feature selection methods combined with different ML algorithms. Table 2 presents classification results for various feature selection techniques, including RFE, SelectFromModel, PSO, and GA. These methods are evaluated alongside popular ML algorithms such as LR, AdaBoost classifier, SVM, DT, GB, and RF. The metrics, including accuracy, precision, recall, F1 score, AUC, and reduction rate.

Table 2. Performance comparison of ML models with various feature selection methods

Model	FS-Method	Accuracy	Precision	Recall	F1-score	AUC	Reduction_rate
LR	GA	99.00	100.0	98.00	98.99	99.00	77.41
	PSO	98.50	100.0	97.00	98.48	98.50	70.32
	RFE	96.20	96.45	96.00	96.60	96.20	51.61
	SelectFromModel	96.30	95.67	96.00	96.59	96.30	45.80
AdaBoost	GA	99.50	99.01	100.0	99.50	99.50	52.25
	PSO	99.60	99.21	100.0	99.6	99.60	60.64
	RFE	97.70	98.02	97.20	97.69	97.70	51.61
	SelectFromModel	97.80	98.21	97.40	97.79	97.80	57.41
SVM	GA	99.00	100.0	98.0	98.99	99.00	70.96
	PSO	98.90	100.0	97.80	98.89	98.90	70.96
	RFE	97.50	97.61	97.40	97.12	97.50	51.61
	SelectFromModel	96.60	96.10	97.20	96.99	96.60	54.83
DT	GA	99.10	99.20	99.00	99.10	99.10	64.51
	PSO	98.10	97.25	99.00	98.12	98.10	69.03
	RFE	96.10	96.11	96.20	95.43	96.10	51.61
	SelectFromModel	95.90	96.21	95.60	95.99	95.90	89.03
GB	GA	99.10	100.0	98.20	99.09	99.10	74.83
	PSO	99.20	99.40	99.00	99.20	99.20	74.19
	RFE	97.50	97.05	98.00	97.90	97.50	51.61
	SelectFromModel	96.50	95.90	97.20	96.52	96.50	83.22
RF	GA	99.50	100.0	99.00	99.50	99.50	58.06
	PSO	98.50	100.0	97.00	98.48	98.50	60.00
	RFE	96.10	96.96	95.20	96.50	96.10	51.61
	SelectFromModel	96.00	97.75	94.20	97.07	96.00	75.48

In Table 2, when feature selection based PSO was employed alongside AdaBoost, it achieved outstanding performance metrics, including an average accuracy of 99.60%, with high precision, recall, F1 score, and AUC of 99.21%, 100.0%, 99.6%, and 99.60% respectively. Similarly, feature selection-based GA exhibited only a marginal decrease of 0.1% in accuracy compared to PSO. Notably, GA demonstrated remarkable results, achieving an accuracy rate of 99.5% when applied to both AdaBoost and RF. Specifically, in AdaBoost, GA achieved precision, recall, F1 score, and AUC of 99.01%, 100.0%, 99.50%, and 99.50% respectively, while in RF, it achieved precision, recall, F1 score, and AUC of 100.0%, 99.00%, 99.50%, and 99.50% respectively. When employing gradient descent, PSO achieved an accuracy of 99.20%, while GA attained an accuracy of 99.10%. Notably, for the ML models LR, SVM, and DT, the feature selection-based GA obtains accuracies of 99.0%, 99.0%, and 99.10%, respectively.

Moreover, Table 2 illustrates the superior performance of feature selection methods based on PSO and GA compared to RFE and SelectFromModel across various ML models. For instance, in AdaBoost, GA achieves an accuracy of 99.50%, while PSO achieves an even higher accuracy of 99.60%. In contrast, RFE and SelectFromModel achieve lower accuracies of 97.70% and 97.80%, respectively. The reduction rate indicates the percentage of features decreased through feature selection. According to Table 2, all feature selection methods significantly reduce dimensionality by opting for only a fraction of the original features, usually less than 50%. PSO and GA typically demonstrate proficiency in feature reduction. For instance, in LR, the reduction rates stand at 77.41% for GA, 70.32% for PSO, 51.61% for RFE, and 45.80% for SelectFromModel.

Figures 2(a) to 2(f) illustrate ROC curves representing different ML classification techniques applied to the breast cancer dataset. As revealed in Figure 2, AdaBoost, when combined with either PSO or GA, achieved the top AUC score of 100%, outperforming RFE and SelectFromModel which achieved scores of 98% each. Additionally, RF reached a 100% AUC score when paired with GA but slightly lower at 99% when using PSO. In contrast, RFE and SelectFromModel achieved lower scores of 96% each. Moreover, SVM and GB both achieved AUC scores of 99 using GA and PSO, while RFE and SelectFromModel obtained slightly lower scores of 98 and 97, respectively. Similarly, LR and DT both achieved AUC scores of 99 with GA, 98 with PSO, and 96 when utilizing RFE and SelectFromModel.

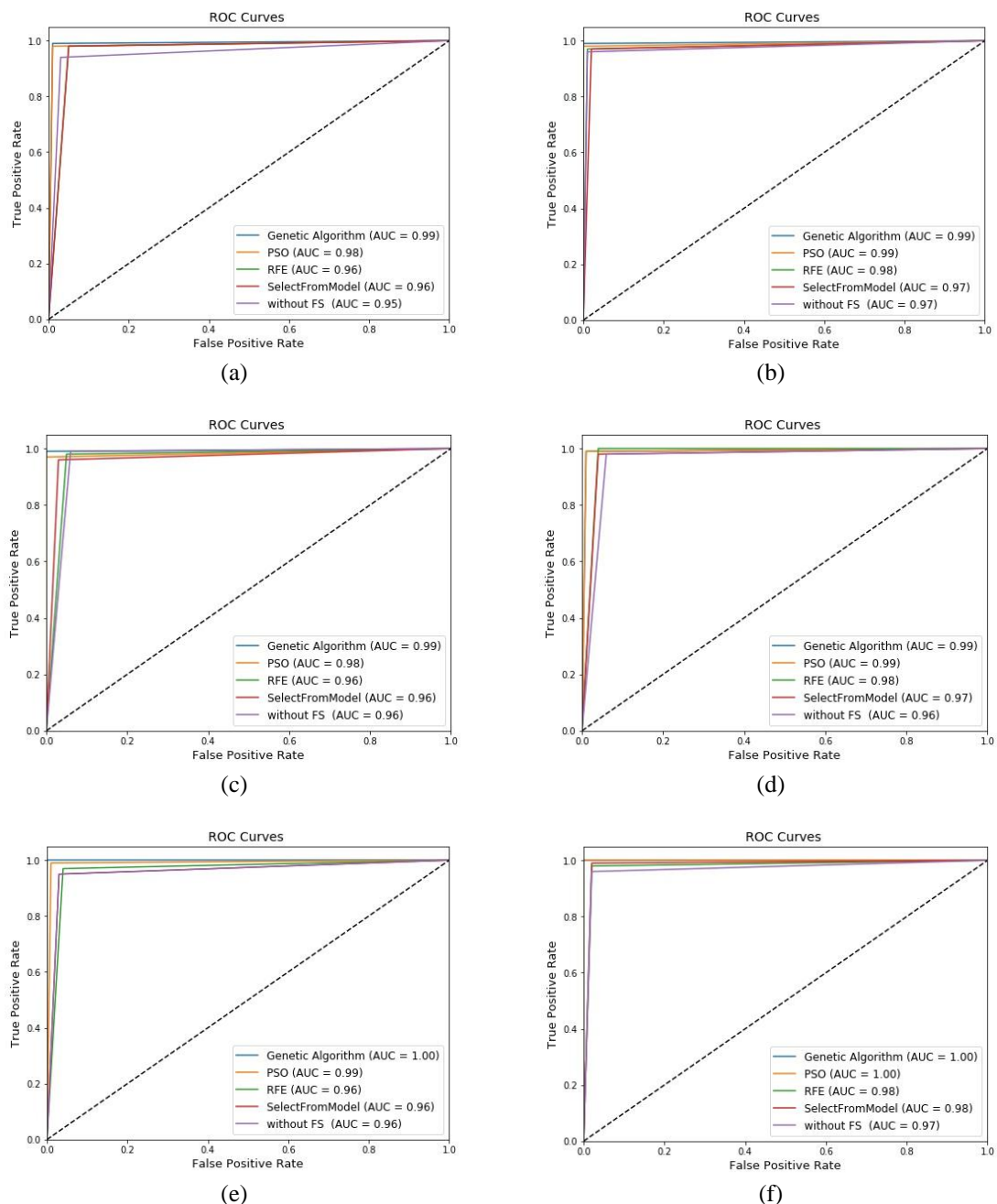


Figure 2. ROC curves for breast cancer dataset using different models: (a) LR model, (b) SVM model, (c) DT model, (d) GB model, (e) RF model, and (f) AdaBoost classifier

Figure 3 provides a detailed comparison of accuracy among four feature selection methods-GA, PSO, RFE, and SelectFromModel-across various ML models. The results demonstrate a consistent trend where GA

and PSO consistently outperform RFE and SelectFromModel in terms of accuracy across all classifiers. For instance, the feature selection based PSO achieved an accuracy of 99.60% in the AdaBoost classifier, securing the top rank with a slight margin compared to GA, which achieved 99.5%. Similarly, in GB, PSO led with an accuracy of 99.20%, followed closely by GA with 99.10%. Conversely, in LR, SVM, RF, and DT, GA-based feature selection dominated with accuracies of 99.00, 99.00, 99.5, and 99.10, respectively, while PSO-based methods ranked second with accuracies of 98.50, 98.90, 98.50, and 98.10.

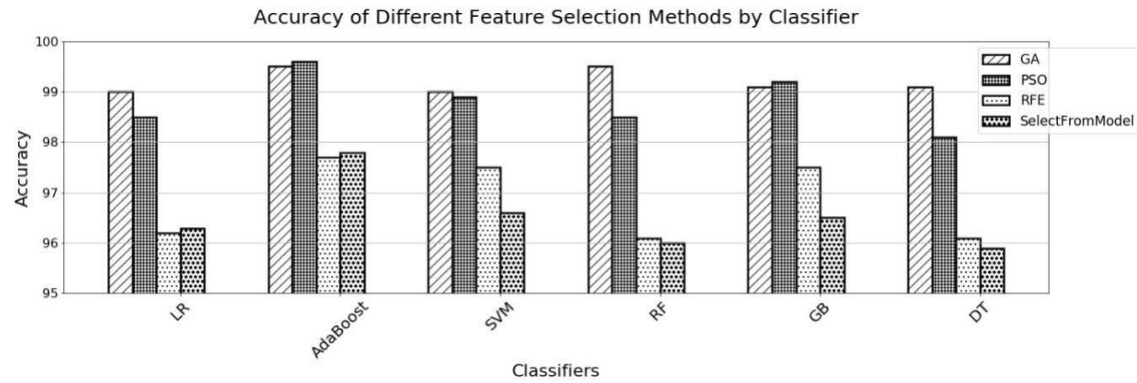


Figure 3. Accuracy achieved by various feature selection methods with different classifier

Table 3 illustrates the comparison between our proposed method and previous studies in breast cancer diagnosis, focusing on accuracy, precision, recall, F-measure, and AUC. It demonstrates that our method significantly surpasses others in these metrics on the WDBC dataset. For instance, when using PSO-based feature selection with AdaBoost, an accuracy of 99.60% was achieved, whereas GA-based feature selection attained 99.50% accuracy when applied to AdaBoost or RF. It's worth noting that in the research conducted by Minnoor and Baths [5], they employed the RF model and obtained an accuracy of 99.30%, which is marginally lower (0.3% difference) than our approach. Similarly, in the study by Lahoura *et al.* [31], they utilized the artificial neural network (ANN) model and achieved an accuracy of 98.68%.

Table 3. Comparative analysis of breast cancer diagnosis studies by various authors

Authors	Year	Dataset collection (samples)	Models	Accuracy	Precision	Recall	F1-score	AUC
Lahoura <i>et al.</i> [31]	2021	569	ANN	98.68	90.54	91.30	81.29	-
Kadhim and Kamil [32]	2023	569	GB	97.36	100	97.87	-	0.99
Nemade and Fegade [4]	2023	569	XGBoost	97.00	-	-	-	99.00
Minnoor and Baths [5]	2023	569	RF	99.30	99.00	100.0	99.00	99.00
Birchha and Nigam [12]	2023	699	Averaged perceptron	98.40	-	100	-	-
Naji <i>et al.</i> [9]	2021	569	SVM	97.20	-	-	-	96.6
Our work	2024	569	Adaboost+PSO	99.60	99.21	100.0	99.60	99.60
			Adaboost+GA	99.50	99.01	100.0	99.50	99.50
			RF+GA	99.50	100.0	99.00	99.50	99.50

4. CONCLUSION

Breast cancer remains a significant cause of women's mortality, but early detection offers a cure. ML-based computer-aided diagnosis (CAD) systems with high accuracy can provide a rapid and cost-effective solution for early recognition. In summary, this research offers a thorough examination of various feature selection methods and their impact on classification performance in breast cancer diagnosis employing various ML classifiers. The results indicate that GA and PSO consistently outperform other feature selection methods, particularly RFE and SelectFromModel, demonstrating their effectiveness in enhancing accuracy across classifiers and reducing the number of feature selections. Both the PSO and GA feature selection methods demonstrate effectiveness in identifying the most relevant features within a training dataset for predicting breast cancer. The study demonstrates PSO's remarkable average accuracy of 99.6% when applied to AdaBoost and GA's 99.5% accuracy with both AdaBoost and RF. This emphasizes the utility of these methods, especially in medical contexts, for predicting breast cancer and aiding in clinical decision-making and risk analysis related

to breast cancer. However, a limitation of this study is the absence of testing the ML models on diverse datasets. Future research should consider evaluating the performance of PSO and GA on different datasets to validate the study's findings. Additionally, exploring deep learning algorithms specifically designed for image analysis holds promise for further investigation in this area.




REFERENCES

- [1] M. Lu, X. Xiao, Y. Pang, G. Liu, and H. Lu, "Detection and localization of breast cancer using UWB microwave technology and CNN-LSTM framework," *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 11, pp. 5085–5094, 2022, doi: 10.1109/TMTT.2022.3209679.
- [2] M. Bibikova and J. Fan, "Liquid biopsy for early detection of lung cancer," *Chinese Medical Journal Pulmonary and Critical Care Medicine*, vol. 1, no. 4, pp. 200–206, 2023, doi: 10.1016/j.pccm.2023.08.005.
- [3] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," *International Journal of Intelligent Networks*, vol. 3, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.
- [4] V. Nemade and V. Fegade, "Machine learning techniques for breast cancer prediction," *Procedia Computer Science*, vol. 218, pp. 1314–1320, 2022, doi: 10.1016/j.procs.2023.01.110.
- [5] M. Minnoor and V. Baths, "Diagnosis of breast cancer using random forests," *Procedia Computer Science*, vol. 218, pp. 429–437, 2022, doi: 10.1016/j.procs.2023.01.025.
- [6] H. E. Massari, N. Gherabi, S. Mhammedi, Z. Sabouri, H. Ghandi, and F. Qanouni, "Effectiveness of applying machine learning techniques and ontologies in breast cancer detection," *Procedia Computer Science*, vol. 218, pp. 2392–2400, 2023, doi: 10.1016/j.procs.2023.01.214.
- [7] S. Kabiraj *et al.*, "Breast cancer risk prediction using XGBoost and random forest algorithm," *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225451.
- [8] C. Kaul and N. Sharma, "High accuracy predictive model on breast cancer using ensemble approach of supervised machine learning algorithms," *2021 International Conference on Computational Performance Evaluation, ComPE 2021*, pp. 71–76, 2021, doi: 10.1109/ComPE53109.2021.9752254.
- [9] M. A. Naji, S. E. Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouahid, and O. Debauche, "Machine learning algorithms for breast cancer prediction and diagnosis," *Procedia Computer Science*, vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.
- [10] R. Dhanya, I. R. Paul, S. S. Akula, M. Sivakumar, and J. J. Nair, "F-test feature selection in Stacking ensemble model for breast cancer prediction," *Procedia Computer Science*, vol. 171, pp. 1561–1570, 2020, doi: 10.1016/j.procs.2020.04.167.
- [11] M. R. Alnowami, F. A. Abolaban, and E. Taha, "A wrapper-based feature selection approach to investigate potential biomarkers for early detection of breast cancer," *Journal of Radiation Research and Applied Sciences*, vol. 15, no. 1, pp. 104–110, 2022, doi: 10.1016/j.jrras.2022.01.003.
- [12] V. Birchha and B. Nigam, "Performance analysis of averaged perceptron machine learning classifier for breast cancer detection," *Procedia Computer Science*, vol. 218, pp. 2181–2190, 2022, doi: 10.1016/j.procs.2023.01.194.
- [13] N. Al-Azzam and I. Shatnawi, "Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer," *Annals of Medicine and Surgery*, vol. 62, pp. 53–64, 2021, doi: 10.1016/j.amsu.2020.12.043.
- [14] S. Benghezouani *et al.*, "Enhancing feature selection with a novel hybrid approach incorporating genetic algorithms and swarm intelligence techniques," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 944–959, 2024, doi: 10.11591/ijece.v14i1.pp944-959.
- [15] M. H. Alshayegi, H. Ellethy, S. Abed, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, 2022, doi: 10.1016/j.bspc.2021.103141.
- [16] R. Eberhart and J. Kennedy, "New optimizer using particle swarm theory," *Proceedings of the International Symposium on Micro Machine and Human Science*, pp. 39–43, 1995, doi: 10.1109/mhs.1995.494215.
- [17] F. Navazi, Y. Yuan, and N. Archer, "An examination of the hybrid meta-heuristic machine learning algorithms for early diagnosis of type II diabetes using big data feature selection," *Healthcare Analytics*, vol. 4, 2023, doi: 10.1016/j.health.2023.100227.
- [18] F. Utaminugrum *et al.*, "Feature selection of gray-level cooccurrence matrix using genetic algorithm with extreme learning machine classification for early detection of pole roads," *Results in Engineering*, vol. 20, 2023, doi: 10.1016/j.rineng.2023.101437.
- [19] M. G. Altarabichi, S. Pashami, S. Nowaczyk, and P. S. Mashhadi, "Fast genetic algorithm for feature selection - a qualitative approximation approach," *GECCO 2023 Companion - Proceedings of the 2023 Genetic and Evolutionary Computation Conference Companion*, pp. 11–12, 2023, doi: 10.1145/3583133.3595823.
- [20] G. Ahn, M. K. Jin, S. B. Hwang, and S. Hur, "Shapelet selection based on a genetic algorithm for remaining useful life prediction with supervised learning," *Heliyon*, vol. 8, no. 12, 2022, doi: 10.1016/j.heliyon.2022.e12111.
- [21] J. O. Onah, S. M. Abdulhamid, M. Abdullahi, I. H. Hassan, and A. Al-Ghusham, "Genetic algorithm-based feature selection and Naïve Bayes for anomaly detection in fog computing environment," *Machine Learning with Applications*, vol. 6, 2021, doi: 10.1016/j.mlwa.2021.100156.
- [22] J. A. Krupinova *et al.*, "Mathematical model for preoperative differential diagnosis for the parathyroid neoplasms," *Journal of Pathology Informatics*, vol. 13, 2022, doi: 10.1016/j.jpi.2022.100134.
- [23] X. Ding, F. Yang, and F. Ma, "An efficient model selection for linear discriminant function-based recursive feature elimination," *Journal of Biomedical Informatics*, vol. 129, 2022, doi: 10.1016/j.jbi.2022.104070.
- [24] G. Wu, C. Li, L. Yin, J. Wang, and X. Zheng, "Compared between support vector machine (SVM) and deep belief network (DBN) for multi-classification of Raman spectroscopy for cervical diseases," *Photodiagnosis and Photodynamic Therapy*, vol. 42, 2023, doi: 10.1016/j.pdpdt.2023.103340.
- [25] A. Erener, A. Mutlu, and H. S. Düzgün, "A comparative study for landslide susceptibility mapping using GIS-based multi-criteria decision analysis (MCDA), logistic regression (LR) and association rule mining (ARM)," *Engineering Geology*, vol. 203, pp. 45–55, 2016, doi: 10.1016/j.enggeo.2015.09.007.
- [26] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short-term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, 2022, doi: 10.1016/j.dajour.2022.100071.
- [27] R. Punmiya and S. Choe, "Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 2326–2329, 2019, doi: 10.1109/TSG.2019.2892595.




- [28] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, "A random forest classifier for lymph diseases," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 465–473, 2014, doi: 10.1016/j.cmpb.2013.11.004.
- [29] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008, doi: 10.1007/s10115-007-0114-2.
- [30] N. Lin, R. Jiang, G. Li, Q. Yang, D. Li, and X. Yang, "Estimating the heavy metal contents in farmland soil from hyperspectral images based on stacked AdaBoost ensemble learning," *Ecological Indicators*, vol. 143, 2022, doi: 10.1016/j.ecolind.2022.109330.
- [31] V. Lahoura *et al.*, "Cloud computing-based framework for breast cancer diagnosis using extreme learning machine," *Diagnostics*, vol. 11, no. 2, 2021, doi: 10.3390/diagnostics11020241.
- [32] R. R. Kadhim and M. Y. Kamil, "Comparison of machine learning models for breast cancer diagnosis," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 415–421, 2023, doi: 10.11591/ijai.v12.i1.pp415-421.

BIOGRAPHIES OF AUTHORS






Salsabila Benghazouani    achieved a master's in data science from the Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca in 2021. Holding an Aggregation degree in Computer Science from Marrakech (2020) and a master's in computer engineering from Mohammed V, Casablanca (2006), she currently serves as a professor in computer science at CPGE Mohamed V, Casablanca. Actively pursuing her Ph.D. at the TIM Lab, Ben M'Sik Faculty of Science, Hassan II University. Her research focus spans artificial intelligence, machine learning, and deep learning. She can be contacted at email: bengahzouani.salsabila239@gmail.com.



Said Nouh    obtained his Ph.D. in computer sciences from ENSIAS, Rabat, Morocco, in 2014. Currently serving as a professor (higher degree research) at the Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco. His research spans artificial intelligence, machine learning, deep learning, and areas such as telecommunications, information, and coding theory. He can be contacted at email: said.nouh@univh2m.ma.



Abdelali Zakrani    earned his B.Sc. and M.Sc. degrees in Computer Science from Hassan II University, Casablanca, Morocco, in 2003 and 2005, respectively. He completed his Ph.D. in the same field at Mohammed V University, Rabat, Morocco, in 2012. Currently, his research focuses on artificial neural network, data mining, and software engineering. He can be contacted at email: abdelali.zakrani@univh2c.ma.