

Precision Measurement and Feature Selection in Medical Diagnostics using Hybrid Genetic Algorithm and Support Vector Machine

Gowri Subadra K¹, Sathish Babu P²

¹Department of IT, Vel Tech High Tech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamil Nadu, India, jeyagowrikalyan@outlook.com

²Department of EEE, University College of Engineering Panruti, Panruti, Tamil Nadu, India

Abstract: The rapid development of intelligent systems has had a significant impact on healthcare, forensics, and medicine, offering innovative solutions to critical problems. Breast cancer, which affects a large number of women each year, requires effective methods for early detection and accurate diagnosis to improve patient outcomes. This study introduces a hybrid feature selection method based on genetic algorithm (GA) and Bucket of Models (BoM) approach to improve breast cancer detection and classification. In the proposed method, GA is used to identify the most relevant features from the breast cancer diagnosis data, to improve the efficiency of the classification process. BoM is then used to select the optimal classification model from a set of candidates, to further improve the accuracy of diagnosis. The support vector machine (SVM) is used as the primary classifier due to its robustness in classifying medical data. The GA feature selection process includes encoding chromosomes, initializing the population, evaluating fitness, and iterating through reproduction steps, that systematically evaluate and select the most informative features for breast cancer diagnosis. In this study, a breast cancer detection accuracy of 97.16 % was achieved, which is a superior performance compared to existing state-of-the-art methods. This study contributes to the development of more accurate and efficient breast cancer screening tools to help healthcare providers make informed diagnostic decisions.

Keywords: breast cancer detection, genetic algorithm, feature selection, support vector machine, classification accuracy

1. INTRODUCTION

Breast cancer [1] is one of the most prevalent and consequential diseases in the world today, affecting millions of people, especially women. Breast cancer is not just a health problem or a tumor [2], but a critical disease where early detection and necessary treatment are especially important. In the early stages, breast cancer does not show any significant symptoms [3], but only small lesions on the skin, swelling in the breast or armpit, which are highly unlikely to be detected. In modern times, advanced methods of treating breast cancer have developed, but these treatments require an adequate system for early detection. According to a report published by the World Health Organization (WHO), 2.3 million new cases of breast cancer were registered in 2024, of which 93 % of patients are not aware of their infection in the earlier stages, as the disease does not show any decisive symptoms.

The three conventional methods for detecting breast cancer are physical examination, mammography, and biopsy. Breast cancer is detected using imaging techniques such as mammography [4], magnetic resonance imaging (MRI) [5] or 3-D mammography. Mammography images are a type of X-ray images [6] of the human breast.

Advances in technology and the incorporation of machine learning (ML) algorithms are helping and contributing to the detection of breast cancer at an earlier stage. Some of the well-known ML algorithms are k-nearest neighbor (KNN) algorithm [7], decision tree (DT) algorithm [8], logistic regression (LR), Naive Bayes (NB) algorithm [9], and random forest (RF) classifier [10]. The biggest challenge with the existing methods for early detection of breast cancer is the imbalance dataset. The self-generated imbalance dataset consists of a large number of features that may be essential and redundant and interfere with the classification process.

On the other hand, it is not only painful but also expensive, invasive, and time-consuming [7]. It is imperative that improved methods are used to detect breast cancer [8]. It has been mentioned that the chances of successfully treating breast cancer are significantly higher if the disease is detected at an earlier stage [9]. The optimization of feature selection from medical data is achieved by using a hybrid GA, which ultimately leads to an improvement in the accuracy of detection and classification models [10]. The main contributions of this proposed work are as follows:

- The novel framework performs an optimized feature selection and supports a robust classification process.
- The introduced ensemble model performs breast classification with minimal computational cost.
- The proposed model reduces the redundancy in terms of features and exhibits better computational accuracy.

The manuscript is organized into an analysis of recent related work in Section 2 and a detailed description of the proposed work in Section 3. The performance analysis and comparative analysis are presented in Section 4 with conclusive remarks in Section 5.

2. RELATED WORK

Basaad et al. (2024) used the characteristics of graph neural networks (GNN) [11] for the detection of breast cancer. The accuracy of this model is reported to be 83.16 %. Supriya et al. (2024) proposed a traditional ML and deep learning (DL) model for breast cancer prediction using the federated learning (FL) method [12]. The framework used the Wisconsin diagnostic breast cancer (WDBC) dataset for training the model and for testing purposes. The accuracy of this model was measured to be 94.73 %. Chen et al. (2024) proposed a modality specific information disentanglement (MoSID) method [13] for the earlier prediction of breast cancer. The major drawback of this model is that the MRI image can only be used for certain women and not for all diseased women.

Furtney et al. (2023) have developed a model for breast cancer detection using the multi-relational directed graph method [14]. This method accepts the MRI image of the patient and evaluates the features using relational graph convolutional neural networks for detecting probabilities of molecular subtypes. Wang et al. (2023) proposed a novel method of dynamic contrast enhanced magnetic resonance (DCE-MR) imaging [15] for breast cancer detection using MRI images. This model suppressed the excessive false negative results of other modern methods and showed an accuracy of 89.61 %. Panigrahi et al. (2024) used the ML algorithm using the minimum redundancy maximum relevance (MRMR) method [16] for feature selection. The model includes four different classifiers namely support vector machine (SVM), decision tree, multilayer perceptron and RF to achieve higher accuracy. The accuracy of this model was measured to be 94.09 % and the computational time was improved.

Thakur et al. (2023) proposed a hybrid model of convolutional neural networks (CNN) and recurrent neural networks (RNN) [17] for detection of cancer in multiple body parts, namely breast, kidney, uterus, etc. This model used VGG-19 and VGG-16 models for training the model and achieved an accuracy of 85.31 %. The main challenge with this model is the large amount of data in the dataset, which leads to dataset imbalance and overfitting. Almaslukh et al. (2024) proposed a computer-aided diagnosis model that incorporates the DL method [18] for early detection of breast cancer. The DL approach used the random search algorithm together with the DenseNet-121 transfer model and achieved an accuracy of 96.42 %.

David et al. (2024) showed a lower effectiveness when using CNN [19], which may be a consequence of problems related to overfitting. Duan et al. (2024) investigated breast

cancer using LASSO regression analysis and classification was performed using a hybrid SVM, LR. This hybrid model showed an accuracy of 93.6 % and an F-score of 88.9 %. The objectives of the proposed model are:

- To optimize feature selection in breast cancer detection and improve prediction performance.
- To improve classification accuracy using the ensemble model.
- To reduce the computational complexity of the breast cancer classification by feature selection.

The proposed work focuses on a multi-class classification problem for breast cancer detection, distinguishing between Normal, cancer stage I, II, and III based on mammogram images. The study proposes a hybrid model that integrates a genetic algorithm (GA) for optimal feature selection and a SVM as the primary classifier with a Bucket of Models (BoM) ensemble approach to improve classification accuracy.

3. PROPOSED METHOD

The proposed detection of early stage breast cancer using GA and SVM is triple folded. In the proposed method, an ensemble strategy is used for breast cancer diagnosis. There are two ensembles. One is for model selection and the other is for attribute selection. An SVM classifier is used to implement this ensemble strategy. Another GA is specifically designed to select the attributes from the models selected by the previous genetic algorithm. The proposed model consists of three stages. The model selection is done in the first stage. This model generates ten models as output. These outputs are responsible for classification. The GA is used for selecting the BoM technique. The classification of the model is done based on the ML technique.

The best configuration can also be selected based on the combination of the models. One such technique is the BoM. The possible model configurations are given by different forms. They are the base classifier and the set of parameters for these base classifiers. The BoM selects the model that fits best from the other available models. Governing the BoM is done using the genetic algorithm. It generates new populations. These populations are nothing more than a group of individuals. Fig. 1 shows the individual steps of the BoM method.

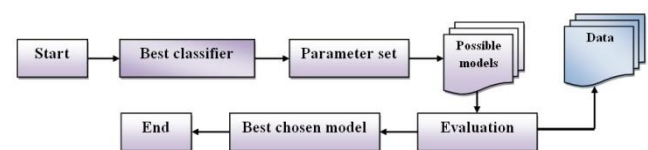


Fig. 1. Steps involved in the BoM method.

The best classifier is selected and then the parameters are defined. Based on the parameters, the possible models are selected and sent for evaluation. The reason for using the GA in the BoM methodology and feature selection is that it does not require much data for evaluation. The first process is the pre-processing stage which consists of several steps, namely image resizing, intensity normalization, noise reduction, grayscale conversion and data augmentation. In the proposed work, no image cropping or RoI extraction is performed but the entire image is used as image cropping would lead to loss of minute features.

Feature selection using genetic algorithm

The result of the previous steps would be the best models selected to fit the problem of classification. The proposed model does not spend much time searching for regions that are not certain, saving the resources and thus minimizing the time required for computation as presented in Fig. 2.

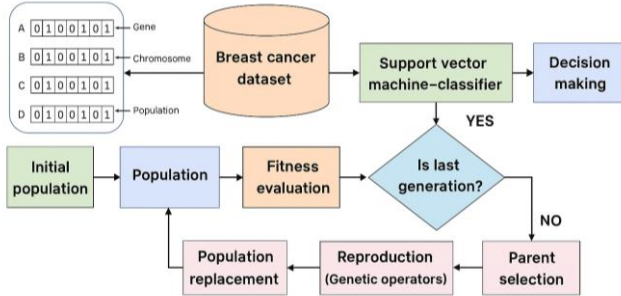


Fig. 2. Proposed breast cancer detection using GA and SVM.

The GA generates a population of potential chromosomes representing a subset of features related to breast cancer. The chromosomes were selected by assigning a gene value of 1, and the non-selected chromosome is labeled with a gene value of 0. The features are defined as shown in (1).

$$p = \{f_1, f_2, f_3, \dots, f_n\} \quad (1)$$

where i represents the feature f_i with 1 as selected and $f_i = 0$ as non-selected chromosome. The fitness function is defined as in (2).

$$S = \arg \max_f f(p) \quad (2)$$

where S is the feature subset, and $f(p)$ is the fitness function of the features. The fitness function $f(p)$ depends on the error rate as defined in (3).

$$f(p) = \frac{1}{1 + E(x)} \quad (3)$$

where $E(x)$ is the error rate and a minimum error rate is required for an improved fitness function. One of the most common methods of feature selection is the roulette wheel selection method, in which chromosomes are selected with a probability proportional to the fitness of the chromosome, as defined in (4).

$$p(f) = \frac{f(c)}{\sum_{i=1}^N f_i(c)} \quad (4)$$

where $p(f)$ is the probability for the selected features and $f(c)$ is the fitness of chromosome i , while N is the total number of chromosomes. The selection of the parent chromosomes is followed by the crossover function as defined in (5) and (6).

$$p^1(c) = \{p_1^1(c); p_2^1(c); p_3^1(c); \dots, p_n^1(c)\} \quad (5)$$

$$p^2(c) = \{p_1^2(c); p_2^2(c); p_3^2(c); \dots, p_n^2(c)\} \quad (6)$$

The crossover leads to the production of offspring with new chromosomes as defined in (7) and (8).

$$O^1(c) = \{p_1^1(c); p_2^1(c); p_3^1(c); \dots, p_k^1(c); p_{k+1}^2(c); p_{k+2}^2(c); p_{k+3}^2(c); \dots, p_n^2(c)\} \quad (7)$$

$$O^2(c) = \{p_1^2(c); p_2^2(c); p_3^2(c); \dots, p_k^2(c); p_{k+1}^1(c); p_{k+2}^1(c); p_{k+3}^1(c); \dots, p_n^1(c)\} \quad (8)$$

The crossover function is followed by the core function of the mutation process. The mutation process randomly flips the bits of the chromosomes and is represented by the m_i defined in (9).

$$m_i = \begin{cases} 1; & \text{if } p(c) = 0 \text{ and mutation occurs} \\ 0; & \text{if } p(c) = 1 \text{ and mutation occurs} \end{cases} \quad (9)$$

This process is repeated until the condition converges and the fitness function is stabilized. The performance of the GA in selecting the feature is improved by implementing the adaptive capability in the genetic algorithm. The adaptive process in the GA is defined in (10).

$$P_{\text{best}} = \begin{cases} P_{\text{max}} \times \left(\frac{P_c - P_i}{P_c - P_{\text{avg}}} \right); & \text{if } P_i < P_{\text{avg}} \\ P_{\text{min}} \times \left(\frac{P_i - P_c}{P_i - P_{\text{avg}}} \right); & \text{if } P_i > P_{\text{avg}} \end{cases} \quad (10)$$

where, P_i is the initial fitness of the parent, P_c is the fitness of the current solution, and P_{avg} is the average fitness of the parent.

In contrast to the existing methods that rely on manual feature selection, the proposed work introduces a two-level ensemble strategy that combines a GA and a BoM for model selection. Fractal analysis is performed during the feature extraction to identify and measure the complexity and irregularities of breast tissue patterns.

4. PERFORMANCE ANALYSIS AND DISCUSSION

The proposed work was trained and tested on the Kaggle dataset – breast ultrasound images dataset (BUID) [23], which consists of 10200 mammogram images with breast cancer. Of these 10200 images, 80 % (8160) were used for training, while the remaining 20 % (2040) were used for the testing process. The proposed method is intended for mammogram classification and the performance of the ML technique was improved by using the GA for feature selection. The performance of the proposed method was calculated by evaluating parameters such as accuracy, precision, recall and F1 score. All experiments were performed on a system with an Intel Core i7 processor (3.6 GHz), 32 GB RAM, and an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM) running Windows 10 (64-bit). The implementation was carried out using Python 3.8 with key libraries such as Scikit-learn, NumPy, and OpenCV.

True positive is the number of positive instances that were correctly identified as positives. True negative is the number of negative instances that were correctly identified as negative instances. False negative is the number of negative instances that were incorrectly identified as positives. False positive is the number of negative instances that were incorrectly identified as negatives. Table 1 shows the selection models for cancer detection based on the area under the curve. The 8LTP and the 3LTP are the local ternary pattern (LTP) used for texture analysis in the image processing and classification application.

Table 1. Selection models for cancer detection based on the area under the curve.

S. No.	Features	Selection of model (Intermediate selection)	Selection of features (Eventual selection)
1.	8LTP + Wavelets + Fractals	81.12	95.87
2.	8LTP + Fractals	81.12	97.16
3.	GLCM	81.12	95.38
4.	2LTP + Fractals + GLCM	76.00	84.98
5.	3LTP + Fractals	74.71	84.98
6.	8LTP + GLCM	69.80	97.16

Note: GLCM – grey level co-occurrence matrix

The 2LTP + Fractals + GLCM have a much lower intermediate and final selection values compared to the previous methods. Table 2 shows the performance evaluation of the different feature groups.

Table 2. Performance evaluation of different feature groups.

S. No.	Features	F1 score [%]	Accu [%]	Sensy [%]	Specy [%]
1.	8LTP + Wavelets + Fra	95.88	95.62	98.44	94.11
2.	8LTP + Fractals	89.47	90.75	91.03	94.11
3.	GLCM	95.36	95.62	95.62	95.88
4.	2LTP + Fractals + GLCM	93.17	93.70	93.70	91.39
5.	3LTP + Fractals	96.39	96.52	96.52	98.44
6.	8LTP + GLCM	96.90	97.16	97.16	98.44

The GLCM technique has an F1 score of 95.36 %, an accuracy and sensitivity of 95.62 % and a specificity of 95.88 % and is depicted in Fig. 3.

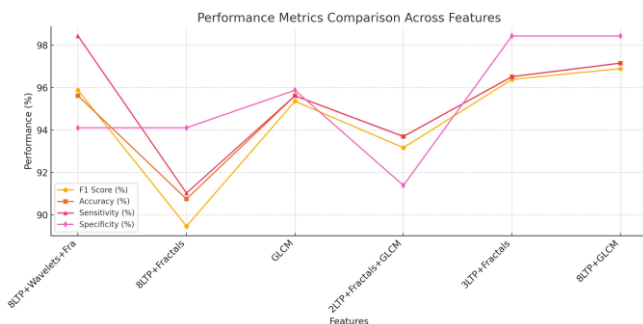


Fig. 3. Graphical representation of the performance evaluation of the different feature groups.

Table 3 contains the training dataset performance evaluation based on the number of genes. Different numbers of genes are considered for calculating the performance measures: 5502, 4096, 2048, 1024, 512, 256, 128, 64, 32 and 16.

Table 3. Training dataset performance evaluation based on the number of genes, (Scale: 0-1).

Gene count	Accuracy	Precision	Recall	Specificity	F1 score
5502	0.91	0.52	0.88	0.91	0.66
4096	0.91	0.53	0.89	0.92	0.66
2048	0.93	0.57	0.87	0.93	0.69
1024	0.92	0.54	0.88	0.92	0.67
512	0.91	0.52	0.90	0.91	0.66
256	0.92	0.54	0.88	0.92	0.68
128	0.90	0.50	0.79	0.91	0.64
64	0.88	0.45	0.76	0.89	0.57
32	0.79	0.28	0.65	0.81	0.39
16	0.75	0.23	0.62	0.76	0.34

When the number of genes is 5502, the accuracy value is 0.91 and when the number of genes is reduced to 16, the accuracy value is 0.75.

The highest recall value is achieved when the number of genes is 512. The values for the second highest recall value and the second lowest recall value are 0.89 and 0.65, respectively.

Table 4 shows the testing dataset performance evaluation as a function of the number of genes. The accuracy value is highest when the number of genes is higher. When the number of genes is 5502, the accuracy value is 0.83 and when the number of genes is reduced to 16, the accuracy value is 0.70. The highest recall value of 0.76 is reached when the number of genes is 2048, 1024 and 128. The second highest recall value of 0.68 is achieved when the number of genes is 5502, 256, 64 and 16. The F1 score values of 0.43, 0.49, 0.48 and 0.38 are obtained when the number of the genes is 512, 256, 128 and 64, respectively.

Table 4. Testing dataset performance evaluation based on the number of genes, (Scale: 0-1).

Gene count	Accuracy	Precision	Recall	Specificity	F1 score
5502	0.83	0.34	0.68	0.81	0.45
4096	0.86	0.38	0.59	0.72	0.46
2048	0.85	0.39	0.76	0.84	0.51
1024	0.87	0.42	0.76	0.82	0.53
512	0.84	0.34	0.59	0.75	0.43
256	0.86	0.39	0.68	0.77	0.49
128	0.83	0.36	0.76	0.86	0.48
64	0.77	0.27	0.68	0.86	0.38
32	0.72	0.20	0.51	0.82	0.28
16	0.70	0.22	0.68	0.90	0.32

Fig. 4 shows the graphical representation of the training dataset performance evaluation as a function of the number of genes. Fig. 5 shows the graphical representation of the testing dataset performance evaluation as a function of the number of genes.

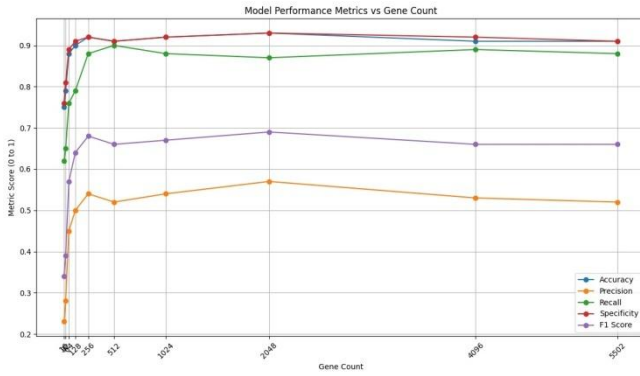


Fig. 4. Graphical representation of the training dataset performance evaluation based on the number of genes.

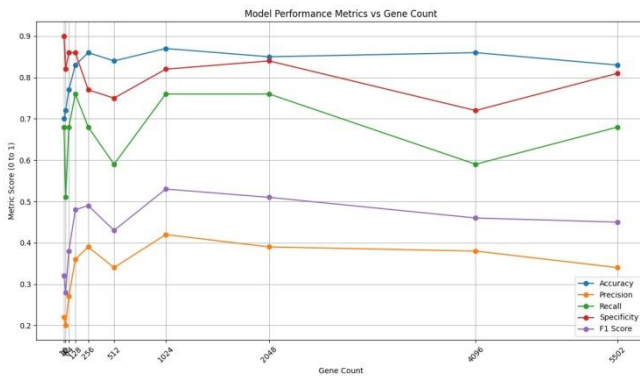


Fig. 5. Graphical representation of the testing dataset performance evaluation based on the number of genes.

Table 5 contains the 15 most important types of genes for differentiating breast cancer. The designations such as 6q26.2-q26.3 stand for the number of the chromosome and the region in which the gene is located on this chromosome.

Table 5. Top 15 types of genes for differentiating breast cancer.

Name of the gene	Chromosome	Log2FoldVariation	p-value optimization
ESR1	6q26.2-q26.3	-9.966061532	0.003
MLPH	2q38.4	-7.235698423	0.005
FSIP1	15q15	-7.762415635	0.008
C5AR2	20q14.33	-5.963125489	0.012
GATA3	11p15	-6.462539781	0.016
TBC1D9	4q32.22	-5.723641265	0.008
CT62	15q24	-9.213658914	0.002
TFF1	22q23.4	-14.23658974	0.002
PRRR15	7q15.4	-7.251323646	0.003
CA12	15q23.3	-7.156982345	0.005
AGR3	7p22.2	-12.36548921	0.001
SRARP	1p37.14	-13.23654897	0.015
AGR2	7p22.2	-9.362145789	0.022
BCAS1	21q13.3	-7.362145587	0.027
LINC00504	5p16.34	-8.256987451	0.001

Table 6 shows the performance comparison of the proposed GA in combination with information gain and information ratio for different classifiers. The harmonic mean of precision and recall is called F1 score.

Table 6. Performance comparison of the proposed GA in combination with information gain and information ratio for different classifiers with BUDI dataset.

Classifier	Parameter	All features	IG	IG-GA	IGR	IGR-GA
SVM [20]	Accuracy	53.59	75.24	85.56	70.08	83.48
	Recall	51.00	74.90	85.35	69.68	83.23
	Precision	27.30	75.42	85.70	70.22	83.62
	F1 score	35.47	75.16	85.52	69.95	83.45
NB [21]	Accuracy	49.46	56.67	56.74	55.65	63.90
	Recall	47.94	54.48	56.55	53.72	61.87
	Precision	46.32	63.95	71.64	57.24	80.32
	F1 score	47.12	58.83	71.64	55.42	68.89
KNN [22]	Accuracy	55.68	72.14	63.19	65.96	86.62
	Recall	55.44	71.53	90.70	64.59	86.99
	Precision	56.79	72.94	90.78	70.32	89.42
	F1 score	56.12	72.23	90.68	67.33	91.73
DT [23]	Accuracy	58.74	68.02	90.73	61.83	91.44
	Recall	58.74	67.72	87.62	61.52	92.32
	Precision	58.26	67.98	87.72	61.68	91.88
	F1 score	58.53	67.87	88.08	61.60	94.82
RF [24]	Accuracy	64.93	87.72	87.70	88.64	94.82
	Recall	64.56	87.52	90.70	88.50	94.81
	Precision	64.86	87.72	90.66	89.72	94.81
	F1 score	64.72	87.67	90.67	88.62	94.81
GA+SVM [25]	Accuracy	71.25	88.64	91.26	92.65	96.84
	Recall	70.25	87.91	91.03	91.49	95.84
	Precision	71.62	88.03	90.64	92.06	95.02
	F1 score	71.03	88.56	91.59	92.12	96.01

For the SVM classifier, the accuracy has improved from 53.59 % for all features to 85.56 % for IG-GA. Recall has increased from 51.00 % to 85.35 % with IG-GA and precision has improved from 27.30 % to 85.70 %.

The accuracy of the NB method has increased from 49.46 % for all features to 63.90 % for IGR-GA. Recall increased from 47.94 % to 61.87 % and precision improved from 46.32 % to 80.32 %.

Accuracy [26] and recall [27] in the KNN classifier improved from 55.68 % to 86.62 % and from 55.44 % to 86.99 %, respectively.

Fig. 6 shows the graphical representation of the performance comparison of SVM [28] and NB classifiers. Fig. 7 shows the graphical representation of the performance comparison of KNN, DT and RF classifiers.

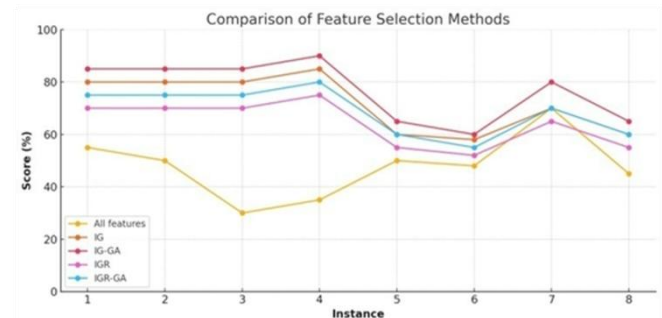


Fig. 6. Graphical representation of the performance comparison of SVM and the NB classifiers.

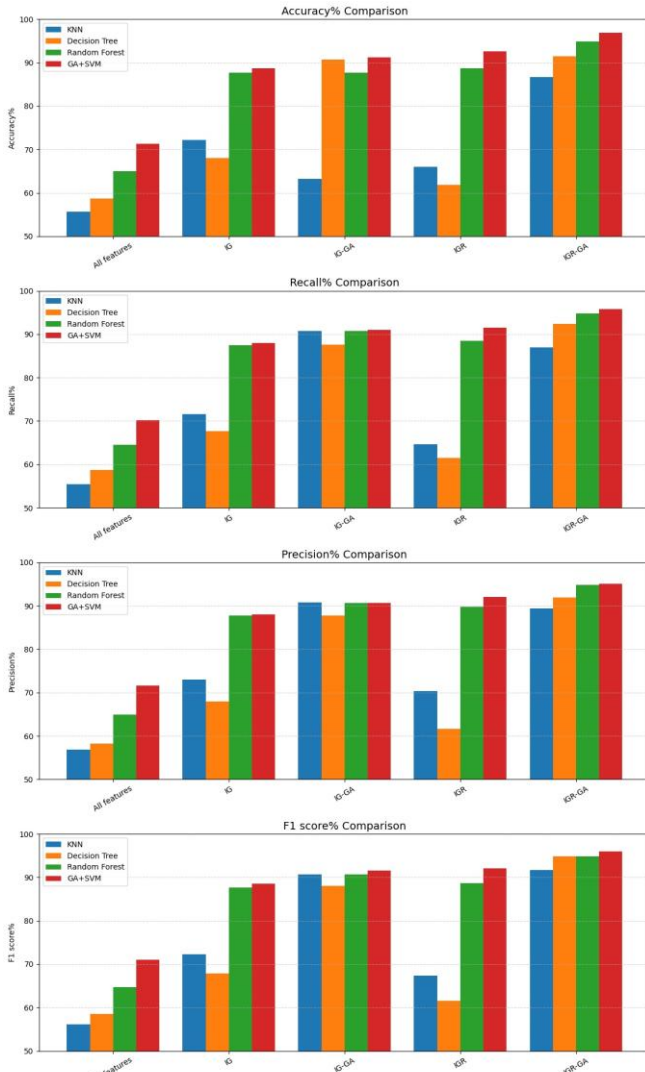


Fig. 7. Graphical representation of the performance comparison of KNN, DT, and RF classifiers.

Table 7 shows the number of selected features before and after applying the GA with different classifiers. The classifier performance is affected by the feature selection methods such as the binary particle swarm optimization (BPSO), information gain (IG), IG-GA, information gain ratio (IGR) and IGR-GA. The number of features after applying IGR-GA is 625, 605, 614, 624 and 619 in the case of SVM, NB, KNN, DT, and RF, respectively. The calculated accuracy values are around 617.

Table 7. Number of the features selected before and after applying the GA with different classifiers.

Dataset	Classifier	All features	After applying GA			
			IG	IG-GA	IGR	IGR-GA
Breast dataset	SVM	24.592	1225	612	1225	625
	NB	24.592	1225	643	1225	605
	KNN	24.592	1225	622	1225	614
	DT	24.592	1225	603	1225	624
	RF	24.592	1225	611	1225	619
	Average	24.592	1225	618	1225	617

Table 8 shows the comparison of the percentage of accuracy values of the proposed method with the existing methods. The genetic improved SVM recursive feature elimination (GI-SVM-RFE) method has achieved an accuracy of 91 % using the RF classifier.

Table 8. Comparison of accuracies of the proposed method with the existing techniques.

Classifier	Proposed method		GI-SVM- RFE	Fusion	PCC- GA	PCC- BPSO
	IG-GA	IGR-GA				
	[%]	[%]	[%]	[%]	[%]	[%]
SVM	95.72	98.63	NA	96.00	98.63	98.63
KNN	86.87	98.63	88.51	NA	96.25	98.63
DT	86.72	88.21	72.51	NA	NA	NA
RF	72.20	83.48	91.00	89.68	96.26	86.72

The performance of the proposed work is tested on different mammogram images of infected and normal breast tissue. The results are shown in Table 9.

From Table 9, it is clear that the proposed GA + SVM model is able to detect and classify breast cancer more effectively than the existing works. The main reason for the improved accuracy is the dataset balance provided by the genetic algorithm. GA efficiently reduces the number of features, which is crucial for high-dimensional datasets such as gene expressions and mammography images.

Sample images were used as a test image for the proposed model and the result was categorized into four classes, namely Normal cell, Cancer – stage I, II, and III. Training images 1 and 2 were assigned to stage II, while image 3 is assigned to stage I in the initial stage. Image 5 is widespread and is assigned to stage III, while image 4 is a normal cell.

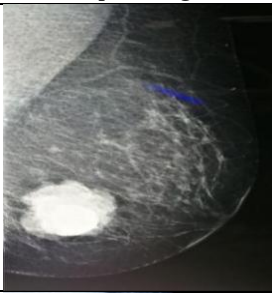
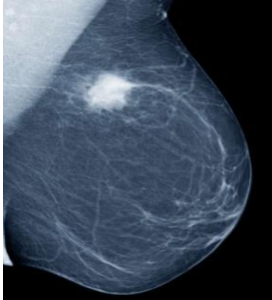
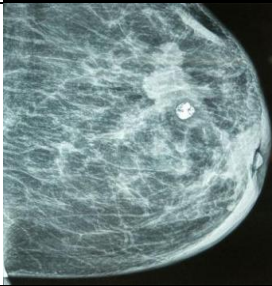
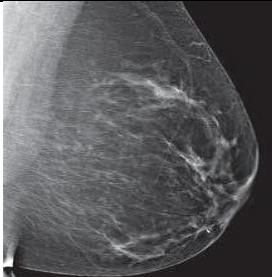
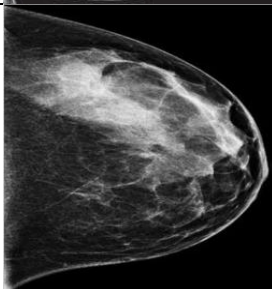
The classification into stages I, II and III was based on the following information on the infected tissue.

- Stage I: Tumor is small in size (< 2 cm) and no lymph node involvement.
- Stage II: Tumor is of 2-5 cm, with limited spreading to nearby lymph nodes.
- Stage III: Larger tumor with dimension > 5 cm, with considerable lymph node involvement.

5. CONCLUSION

Breast cancer is one of the main types of cancer affecting many women. It is the unwanted growth of cells in the breast that leads to a tumor. In the proposed method, a genetic algorithm-based feature selection technique has been used in combination with BoM methodology for feature selection. These methods are used for selecting the best features among the different features from the collected datasets. The proposed GA showed the best performance when it was tested with the information gain and information ratio gain filtering techniques. The effectiveness of the proposed model was evaluated based on various performance metrics. Different classifiers were tested with the proposed GA and the results were analyzed.

Table 9. Performance analysis of the proposed work.

S. No.	Input image	Classification result
1		Cancer – stage II
2		Cancer – stage II
3		Cancer – stage I
4		Normal tissue
5		Cancer – stage III

REFERENCES

- [1] Lupat, R., Perera, R., Loi, S., Li, J. (2023). Moanna: Multi-omics autoencoder-based neural network algorithm for predicting breast cancer subtypes. *IEEE Access*, 11, 10912-10924. <https://doi.org/10.1109/ACCESS.2023.3240515>
- [2] Megha, R., Geethapriya, Radhakrishna, S., Eranki, A. (2024). Breast tumor heterogeneity quantification using 3D ultrasound texture. In *2024 IEEE South Asian Ultrasonics Symposium (SAUS)*. IEEE. <https://doi.org/10.1109/SAUS61785.2024.10563639>
- [3] Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J. (2023). HoVer-Trans: Anatomy-aware HoVer-Transformer for ROI-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*, 42 (6), 1696-1706. <https://doi.org/10.1109/TMI.2023.3236011>
- [4] Lamprou, C., Katsikari, K., Rahmani, N., Hadjileontiadis, L. J., Seghier, M., Alshehhi, A. (2024). StethoNet: Robust breast cancer mammography classification framework. *IEEE Access*, 12, 144890-144904. <https://doi.org/10.1109/ACCESS.2024.3473010>
- [5] Felício, J. M., Martins, R. A., Costa, J. R., Fernandes, C. A. (2024). Microwave breast imaging for cancer diagnosis: An overview [Bioelectromagnetics]. *IEEE Antennas and Propagation Magazine*, 66 (4), 85-97. <https://doi.org/10.1109/MAP.2024.3411480>
- [6] Prabakaran, D., Sheela, K. (2021). A strong authentication for fortifying wireless healthcare sensor network using elliptical curve cryptography. In *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*. IEEE, 249-254. <https://doi.org/10.1109/MysuruCon52639.2021.9641546>
- [7] Batool, A., Byun, Y.-C. (2024). Toward improving breast cancer classification using an adaptive voting ensemble learning algorithm. *IEEE Access*, 12, 12869-12882. <https://doi.org/10.1109/ACCESS.2024.3356602>
- [8] Xie, X., Wu, L., Su, Z., Sun, Z., Cao, X., Hou, Y. (2024). CORONet: A cross-sequence joint representation and hypergraph convolutional network for classifying molecular subtypes of breast cancer using incomplete DCE-MRI. *IEEE Journal of Biomedical and Health Informatics*, 28 (4), 2103-2114. <https://doi.org/10.1109/JBHI.2024.3355111>
- [9] Tiryaki, V. M., Tutkun, N. (2024). Breast cancer mass classification using machine learning, binary-coded genetic algorithms and an ensemble of deep transfer learning. *The Computer Journal*, 67 (3), 1111-1125. <https://doi.org/10.1093/comjnl/bxad046>
- [10] Li, Z.-Z., Wang, F.-L., Qin, F., Yusoff, Y. B., Zain, A. M. (2024). Feature selection of gene expression data using a modified artificial fish swarm algorithm with population variation. *IEEE Access*, 12, 72688-72706. <https://doi.org/10.1109/ACCESS.2024.3402652>
- [11] Basaad, A., Basurra, S., Vakaj, E., Aleskandarany, M., Abdelsamea, M. M. (2024). GraphX-Net: A graph neural network-based Shapley values for predicting breast cancer occurrence. *IEEE Access*, 12, 93993-94007. <https://doi.org/10.1109/ACCESS.2024.3424526>

- [12] Supriya, Y., Chengoden, R. (2024). Breast cancer prediction using Shapely and game theory in federated learning environment. *IEEE Access*, 12, 123018-123037.
<https://doi.org/10.1109/ACCESS.2024.3424934>
- [13] Chen, Q., Zhang, J., Meng, R., Zhou, L., Li, Z., Feng, Q. (2024). Modality-specific information disentanglement from multi-parametric MRI for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 43 (5), 1958-1971. <https://doi.org/10.1109/TMI.2024.3352648>
- [14] Furtney, I., Bradley, R., Kabuka, M. R. (2023). Patient graph deep learning to predict breast cancer molecular subtype. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20 (5), 3117-3127.
<https://doi.org/10.1109/TCBB.2023.3290394>
- [15] Wang, S., Sun, K., Wang, L., Qu, L., Yan, F., Wang, Q. (2023). Breast tumor segmentation in DCE-MRI with tumor sensitive synthesis. *IEEE Transactions on Neural Networks and Learning Systems*, 34 (8), 4990-5001.
<https://doi.org/10.1109/TNNLS.2021.3129781>
- [16] Panigrahi, A., Pati, A., Sahu, B., Das, M. N., Nayak, D. S. K., Sahoo, G. (2023). En-MinWhale: An ensemble approach based on MRMR and whale optimization for cancer diagnosis. *IEEE Access*, 11, 113526-113542.
<https://doi.org/10.1109/ACCESS.2023.3318261>
- [17] Thakur, T., Batra, I., Malik, A., Ghimire, D., Kim, S.-H., Sanwar Hosen, A. S. M. (2023). RNN-CNN based cancer prediction model for gene expression. *IEEE Access*, 11, 131024-131044.
<https://doi.org/10.1109/ACCESS.2023.3332479>
- [18] Almaslukh, B. (2024). A reliable breast cancer diagnosis approach using an optimized deep learning and conformal prediction. *Biomedical Signal Processing and Control*, 98, 106743.
<https://doi.org/10.1016/j.bspc.2024.106743>
- [19] Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R. C., Gambhir, S. S., Kuhn, P., Rebbeck, T. R., Balasubramanian, S. (2022). Early detection of cancer. *Science*, 375 (6586).
<https://doi.org/10.1126/science.aay9040>
- [20] Rashid, T. A., Majidpour, J., Thinakaran, R., Batumalay, M., Arrova Dewi, D., Hassan, B. A. (2024). NSGA-II-DL: Metaheuristic optimal feature selection with deep learning framework for HER2 classification in breast cancer. *IEEE Access*, 12, 38885-38898.
<https://doi.org/10.1109/ACCESS.2024.3374890>
- [21] Mirimoghaddam, M. M., Majidpour, J., Pashaei, F., Arabalibeik, H., Samizadeh, E., Roshan, N. M., Rashid, T. A. (2024). HER2GAN: Overcome the scarcity of HER2 breast cancer dataset based on transfer learning and GAN model. *Clinical Breast Cancer*, 24 (1), 53-64.
<https://doi.org/10.1016/j.clbc.2023.09.014>
- [22] Issa, A. S., Ali, Y. H., Rashid, T. A. (2023). Review on hybrid swarm algorithms for feature selection. *Iraqi Journal of Science*, 64 (10), 5331-5344.
<https://doi.org/10.24996/ij.s.2023.64.10.38>
- [23] Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, 28, 104863.
<https://doi.org/10.1016/j.dib.2019.104863>
- [24] Long, J., Zheng, Z., Wang, J., Ng, C. K., Liu, C., Ji, W. (2024). Decision tree based automated detection of breast cancer. In *2024 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*. IEEE, 549-554.
<https://doi.org/10.1109/CIS-RAM61939.2024.10673070>
- [25] Rahman, S., Siregar, D., Syah, R. B. Y., Setiawan, H., Maulana, A. E., Hamsiah. (2023). The effective breast cancer classification with the random forest algorithm. In *2023 International Conference of Computer Science and Information Technology (ICOSNIKOM)*. IEEE.
<https://doi.org/10.1109/ICoSNIKOM60230.2023.10364529>
- [26] Khater, T., Hussain, A., Bendardaf, R., Talaat, I. M., Tawfik. H., Ansari, S., Mahmoud, S. (2025). An explainable artificial intelligence model for the classification of breast cancer. In *IEEE Access*, 13, 5618-5633.
<https://doi.org/10.1109/ACCESS.2023.3308446>
- [27] Maouche, I., Terrissa, L. S., Benmohammed K., Zerhouni, N. (2023). An explainable AI approach for breast cancer metastasis prediction based on clinicopathological data. In *IEEE Transactions on Biomedical Engineering*, 70 (12), 3321-3329.
<https://doi.org/10.1109/TBME.2023.3282840>
- [28] Shukla, V., Kaarthika, Mathur, A., Narayan P., Kishor, K. (2025). A multi-modal approach for the molecular subtype classification of breast cancer by using Vision Transformer and novel SVM polyvariant kernel. In *IEEE Access*, 13, 97545-97558.
<https://doi.org/10.1109/ACCESS.2025.3575126>

Received November 14, 2024

Accepted June 9, 2025