Shankar Thawkar* and Ranjana Ingolikar

Classification of Masses in Digital Mammograms Using the Genetic Ensemble Method

https://doi.org/10.1515/jisys-2018-0091
Received February 13, 2018; previously published online July 26, 2018.

Abstract: All over the world, breast cancer is the second leading cause of death in women above 40 years of age. To design an efficient classification system for breast cancer diagnosis, one has to use efficient algorithms for feature selection to reduce the feature space of mammogram classification. The current work investigates the use of hybrid genetic ensemble method for feature selection and classification of masses. Genetic algorithm (GA) is used to select a subset of features and to evaluate the fitness of the selected features, Adaptive boosting (AdaBoost) and Random Forest (RF) ensembles with 10-fold cross-validation are employed. The selected features are used to classify masses into benign or malignant using AdaBoost, RF, and single Decision Tree (DT) classifiers. The performance evaluation of classifiers indicates that AdaBoost outperforms both RF and single DT classifiers. AdaBoost achieves an accuracy of 96.15%, with 97.32% sensitivity, 95.90% specificity, and area under curve of $A_Z = 0.982 \pm 0.004$. The results obtained with the proposed method are better when compared with extant research work.

Keywords: Digital mammography, decision Tree, feature selection, classification, genetic algorithm, ensembles, AdaBoost, Random Forest, receiver operating characteristics curve.

1 Introduction

Breast cancer is considered the second leading cause of death in women above 40 years of age all over the world. Presently, no technique has been discovered for the prevention of breast cancer; hence, detection of breast cancer in its primary stage is very important. Mammography is the finest tool available for the detection of breast cancer in its initial stage [27]. Radiologists diagnose breast cancer by reading the mammogram; however, reading of mammograms is a very challenging task. The suspicious mass is detached for clinical inspection by means of a biopsy procedure. Statistics indicates that more than 60%–70% of the suspicious cases turn out to be benign. This problem can be minimized to a certain extent with the use of a computer-based diagnosis system. Such systems act as a second reader for radiologists and help to minimize unnecessary biopsies. With the use of image processing and machine learning techniques, detection and classification of masses become easier, but still, it is a challenging area of research, especially the detection and classification of masses into benign and malignant. One of the factors that influence the performance of classifiers is feature selection. The basic objective of feature selection techniques is to remove irrelevant or redundant features from the set of features.

Currently, ensemble-based classification is an active area of research in machine learning and pattern recognition [29]. The basic idea of ensemble is to train multiple models using the same learning algorithm. Several studies about ensemble classifiers demonstrate that the combination prototype of classifiers is better than an individual classifier [18]. Nowadays, bagging and boosting ensemble methods have gained wide popularity [31]. In bagging technique, prediction of base classifiers is achieved using the majority voting method.

***Corresponding author: Shankar Thawkar,** Electronics and Computer Science, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, Maharashtra, India, e-mail: shankar.thawkar@gmail.com. https://orcid.org/0000-0002-0118-9605
Ranjana Ingolikar: Department of Computer Science, S. F. S College, Nagpur, Maharashtra, India

Random Forest (RF) is one the most popular bagging techniques. Boosting combines weak classifiers to build a strong classifier. In the boosting method, classifiers are trained on the weighted version of dataset and then combined to produce the final prediction. Adaptive boosting (AdaBoost) is the most popular adaptive boosting algorithm. In this article, genetic algorithm (GA), in conjunction with AdaBoost and RF ensembles, was used for feature selection. To classify the masses, AdaBoost, RF, and single Decision Tree (DT) classifiers were used.

The remaining part of the article is organized as follows: Review of related work is in Section 2; Section 3 describes the research methodology for feature selection and classification of masses; Section 4 contains the Results and Discussion; Section 5 describes the computational complexity of the proposed method. Conclusion regarding the proposed method is discussed in Section 6.

2 Related Work

A sophisticated version of classical floating search algorithm for feature selection is presented by Somol et al. [32]. The new adaptive floating search algorithms have a potential to find the solution close to the optimal one.

Applications of a rough set approach for feature selection in pattern recognition using reducts and dynamic reducts are presented by Swiniarski and Skowron [34]. The features selected by rough set and principal components analysis (PCA) methods are used for face and mammogram recognition using an artificial neural network (AAN).

A two-stage method for the detection of microcalcifications is presented by Fu et al. [10]. In the first stage, a mathematical model is used for the detection of location and shape. The second stage consists of feature selection and classification. A sequential forward search (SFS) method is used to select features relevant to microcalcifications. Features selected by SFS are used to train and test two classifiers, namely, a regression neural network (GRNN) and a support vector machine (SVM). The results of the classifiers are evaluated using area under receiver operating characteristics (ROC) curve.

In Ref. [20], an AAN-based ensemble algorithm was proposed for the detection of speculated masses. Four feature images were obtained from a single mammogram. Each image belonged to one of the four features obtained from a single pixel of the mammogram. The features obtained from a single pixel of the mammogram were mean pixel brightness, standard deviation of pixel brightness, standard deviation of gradient orientation histogram, and standard deviation of the folded gradient orientation. The feature images were partitioned into small blocks, and 10-fold cross-validation was performed on each partition. Finally, for each validation, a neural network was constructed using the bagging technique to detect speculated masses.

A novel multi-objective GA and an ensemble classifier were proposed by Zhang and Yang [39]. A multi-objective GA was used for optimal feature selection. An ensemble classifier consists of a DT classifier, an ANN, and an SVM classifier. The proposed method was tested on three benchmark datasets – sonar, ionosphere, and soybean. A fivefold cross-validation technique was used to evaluate the overall accuracy of classifiers. Experimental results showed that the proposed method is suitable and useful for feature optimization and classification.

A multi-resolution approach for automated classification of mammograms was presented by Dong and Wang [6]. A Gabor filter of different frequencies was used for feature extraction and classification. The dimensionality of features selected by Gabor filters is reduced with the use of statistical test, namely, t-test and its p-values.

A diverse study on various machine learning algorithms such as Naïve Bayes, Multilayer Perceptron, DT, ensemble methods based on bagging, boosting, and a combination of the best base classifiers using meta-learning techniques of stacking and voting was presented in Ref. [35]. The performance of these classifiers was evaluated on different medical datasets. The outcome of the study was the guideline for selecting the best classifier for a particular dataset. It has been observed that not every classification technique is suitable for all kinds of databases because of their dimensionality, multiple classes, and noisy data. The study concluded that voting technique is the most powerful technique among all the techniques studied.

In Ref. [7], an ensemble of Bayesian classifiers was presented for the classification of masses in digital mammograms. The ensembles combined the prediction of three Bayesian classifiers: Tree Augmented Naïve

Bayes, Markov Blanket Estimation, and Ensemble of Bayesian classifiers. The method used confidence score for the selection of the best model. The ensemble with the highest confidence won. The method was compared with multilayer perceptron neural network (MLPNN) classifier, and it has been observed that the ensemble classifier is better than MLPNN. It achieves an accuracy of 91.83% on training and 90.63% on testing.

A method for detecting and classifying curvilinear structures [2] in mammograms was proposed. Detection and classification were achieved with dual tree complex wavelet transform (DT-CWT) and RF. The DT-CWT performed best for all three tasks: curvilinear structure detection, orientation estimation, and spicule classification. The method achieved an area under curve of $A_z = 0.923$ for curvilinear structure detection and $A_z = 0.761$ for classification.

A mass detection method using RF was proposed in Ref. [28]. The method was tested on 120 mammograms of CC view with 60 benign and 60 malignant cases. Texture patterns such as entropy, energy, sum average, sum variance, and cluster tendency were used to analyze the region of interest (ROI). The best features were selected using GA. The proposed method achieved an area under curve of $A_z = 0.90$.

Zhang et al. [40] proposed an ensemble system for the classification of suspicious masses into malignant or benign. In a segmentation process, multiple contours are generated from ROI. Fourteen shape-based features are extracted from each of the segmented contours for classification. The dataset was partitioned into four subsets based on young age, old age, small ROI size, and large ROI size. Then for each subset, an ensemble was built for the classification of masses. The proposed method achieves an accuracy of 72%.

Luo and Cheng [22] proposed ensemble-based techniques for the accurate prediction of breast cancer. An optimal feature set was obtained using two techniques: forward selection and backward selection. The classifiers DT, SVM-sequential minimal optimization (SVM-SMO), and their ensembles were trained using these optimal features for the prediction of breast cancer. The performance of the classifier was tested on the breast cancer dataset obtained from the Institute of Radiology of the University of Erlangen-Nuremberg. The prediction performance of the classifiers was evaluated with 10-fold cross-validation technique. The best results were achieved with DT ensemble. It achieved an accuracy of 83.4% with area under curve of $A_z = 0.866 \pm 0.004$. The authors conclude that ensemble-based classifiers are better than single classifiers.

A clustered ensemble neural network method for the classification of masses in digital mammograms was proposed by Mc Leod and Verma [24]. The technique uses K-mean classifier to generate clusters by partitioning the data based on different seeding points. A neural network classifier was then trained with each cluster generated by k-mean for each layer. This layer-based training provided diversity into the classifiers. Finally, the network of clustered ensemble was created by combing the outputs using the majority voting technique. The experiment employed over 100 masses obtained from the Digital Database for Screening Mammography (DDSM).

Jothi et al. [15] proposed a tolerance rough set model for the classification of mammograms. An optimal feature set is selected using the Tolerance Rough Set PSO-based Quick Reduct (STRSPSO-QR) and Tolerance Rough Set-PSO-based Relative Reduct (STRSPSO-RR) methods. The results achieved helped to improve the diagnosis accuracy of breast cancer.

Choi et al. [5] proposed an ensemble classification algorithm for the reduction of false-positive detections on mammography database. For the detection of ROI, a contour-based unsupervised learning using multilevel thresholding and wavelet model based supervised learning was used. The results of both supervised and unsupervised methods were combined for the segmentation of ROI. The texture, shape, intensity, and speculation index-based features were extracted from the segmented ROI. An ensemble classifier based on adaptive boosting algorithm was developed for the detection of masses. It was observed that the proposed ensemble-based classification system significantly reduces false positive detection when it was tested on the mammogram database.

An ensemble supervised classification method was proposed in Ref. [1]. Twenty texture-based features of benign and malignant breast cases were extracted using gray level co-occurrence matrix. Then the maximum difference method was employed to select a subset of six features, the method select features based on the maximum difference between benign and malign dataset. These features were used to classify masses using three supervised classifiers: K-nearest neighbors (KNN), Naive Bayes, and SVM. Finally, results were obtained using the voting method. The experiment was conducted on 200 mammograms obtained from the DDSM.

Mafarja and Mirjalili [23] proposed two hybrid models for feature selection based on the Whale Optimization Algorithm (WOA). In the first model, the simulated annealing (SA) algorithm was embedded in WOA. The best solution found at each iteration of WOA was used in the second model. The performance of the model was evaluated using standard benchmark datasets, and the results were compared with well-known wrapper feature selection methods.

Mohanty et al. [26] proposed a feature selection technique based on the Forest Optimization Algorithm (FOA). FOA belongs to the family of wrapper-based feature selection technique. The optimal feature set selected by FOA was used to classify mammograms using different classifiers, namely, SVM, k-NN, Naïve Bayes, and C4.5.

The inspirations driving our proposed technique are as follows:

- poor performance of existing ensemble-based classifiers
- requirement to reduce type I error [false positive rate (FPR)]
- justification of the statement "Ensembles are better than single Decision Tree classifiers"
- requirement to use balance as well as large dataset for experiment
- The techniques must be useful to improve breast cancer diagnosis.

3 Proposed Framework

GA with AdaBoost and RF algorithms was used for feature selection. The features selected were used for the classification of masses using AdaBoost, RF, and single DT classifiers. The proposed methodology is as shown in Figure 1.

3.1 Feature Selection Using the Genetic Ensemble Method

The features extracted for the classification of masses are categorized into three types: intensity based, texture based, and shape based. A total of 25 features extracted is shown in Table 1 [36, 37].

The main objective of feature selection is to remove irrelevant or unnecessary features from the extracted feature set. Feature selection techniques select the most relevant features based on four criteria: discrimination, reliability, independence, and optimality [19, 30].

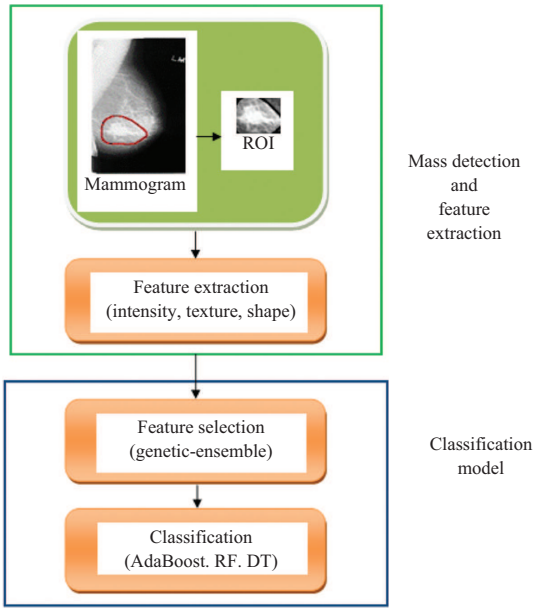

Figure 1: Proposed Framework.

Table 1: Extracted Features.

1	Average gray level	14	Homogeneity
2	Average contrast	15	Sum average
3	Smoothness	16	Sum variance
4	Skewness	17	Sum entropy
5	Uniformity	18	Area
6	Entropy1	19	Perimeter
7	Energy	20	Compactness
8	Entropy2	21	Normalized standard deviation
9	Contrast	22	Area ratio
10	Mean	23	Contour roughness
11	Standard deviation	24	Normalized residual value
12	Variance	25	Overlapping ratio
13	Correlation		

In this article, GA with AdaBoost and RF ensemble was used to select the most relevant features that will improve the performance of a classifier. A detailed description about GA is given in the following section.

3.2 Genetic Algorithm

GA based on the Darwin theory of evolution signifies survival of the fittest. It is a computerized search and optimization technique based on natural selection and natural genetics. The evolutionary nature of the technique leads to various models for solving optimization problems [8, 11, 14, 25]. The process of feature selection using GA with AdaBoost and RF is as shown in Figure 2. It consists of the following steps:

3.2.1 Population Encoding Technique

In this work, binary encoding technique was used for encoding each element or chromosome in a population. Each chromosome or element in the population represents a possible solution of the feature selection

```
1.  Load feature data set        //  651 samples of  25 features
2.  Initialize GA  parameters
      n                  // maximum iterations
      ps                 // population size
      nf                 // number of feature
      pc                 // Crossover probability
      pm                 // Mutation Probability
      np                 //  Number of parents

3.  Define fitness function,  f(Ci) = mean (accuracy)      // classifier accuracy
4.  Generate initial solution (i = 1,2,3,……ps)
5.  Encode the initial  solution using Binary encoding  technique
6.  Compute  fitness of  initial solution using classification accuracy of
    AdaBoost/Random Forest classifiers
7.  Select Best solution
8.      for i = 1 to n
9.          for k= 1 to np/2
10.             Select two parents p1 and p2 using Roulette wheel selection
11.             Apply Crossover operation on p1 and p2
12.             Evaluate new offspring's  // calculate fitness
13.         end  loop k
14.         for k= 1 to np
15.             Select parent
16.             Apply mutation on selected parent
17.             Evaluate parent after mutation
18.         End loop k
19.       store the best solution ever found for current iteration
20.   end loop i
21.   Obtain the final result as optimal features
```

Figure 2: GA Ensemble-Based Feature Selection Algorithm.

problem. An individual element in a population consists of 25 genes, one gene for each feature. A gene is represented by two values. A value of '1' indicates that the corresponding feature is selected, and a value of '0' indicates that the feature is not selected. The search space consists of 2^{25} chromosomes [38].

3.2.2 Fitness Evaluation

The fitness of each chromosome in a population was evaluated using AdaBoost and RF classifiers with 10-fold cross-validation. The fitness function of this GA ensemble is the classification accuracy of the classifier. It is defined as follows:

$$\text{fitness}(C_i) = \frac{\sum_1^n \text{accuracy}(C_i)}{n}, \tag{1}$$

where C_i is the subset of features and n is the number of iterations.

3.2.3 Reproduction Operation

Reproduction is the first operator applied on the population. It is also called selection operator. The Roulette wheel selection strategy is used to select the best chromosomes based on their fitness value for crossover operation. In this method, a chromosome is chosen from the mating pool (population) with a probability proportional to the fitness. Thus, the probability of selecting the i^{th} chromosome is as follows:

$$P_i = \frac{F_i}{\sum_{j=1}^n F_j}, \tag{2}$$

where n is the number of chromosomes in the population and F_i is the fitness value of the chromosome. The cumulative probability P_i of each chromosome is computed as follows:

$$P_i = \sum_{j=1}^{i} p_j \tag{3}$$

The chromosome will be selected for crossover if $P_{i-1} < C < P_i$. Parameter C is a random number between value 0 and 1.

3.2.4 Crossover Operator

Once the chromosomes are selected, a crossover operation is performed on them. The crossover operation is performed with the hope that the new generation will be better than the previous one. It occurs with crossover probability (pc). In this work, one-point or two-point crossover operation was performed on the selected chromosomes. The one-point or two-point crossover was decided using the Roulette wheel selection strategy.

3.2.5 Mutation Operation

Mutation is performed just after a crossover operation. A mutation operation is used to thwart the fall of all solutions in a population into a local optimum of solved problem. It maintains diversity in the population. Binary mutation is used to changes bits of the new child (new chromosome) randomly. The amount of bits flipped is determined by mutation probability (pm).

The summary of the features selected using genetic ensemble method is as shown in Table 2.

Table 2: Summary of Features Selected by Genetic Ensemble Method.

Input data size (features × no. of cases)	Number of iterations/method	Pop size	Feature set	No. of features selected	Feature selected	Accuracy of feature selection (%)
25 × 651	50 (AdaBoost)	10	F1	12	1, 6, 7, 8, 11, 13, 14, 15, 19, 20, 21, 23	94.08
		20	**F2**	**11**	**1, 2, 6, 7, 13, 14, 17, 19, 20, 23, 24**	**96.15**
		30	F3	15	1, 3, 6, 7, 8, 9, 12, 13, 15, 16, 17, 18, 20, 22, 23	92.47
	50 (Random Forest)	10	F4	10	1, 6, 9, 13, 14, 17, 18, 20, 21, 23	92.78
		20	F5	15	1, 2, 6, 7, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20, 23	92.70
		30	F6	9	2, 4, 6, 8, 13, 16, 17, 20, 23	93. 30

Feature set F2 appears to be best among all subsets.

3.3 Classification

The proposed method used ensemble classifiers based on boosting and bagging techniques for the classification of masses. An ensemble classifier uses multiple base classifiers to form a final classifier. A base classifier may be DT, ANN, or SVM. The final classifier (ensemble) is better than the single classifier. The most popular ensemble techniques used for classification nowadays are bagging and boosting.

3.3.1 Bagging

Bagging is one of the earliest ensemble techniques, also known as bootstrap aggregating [3]. In this technique, each base classifier was trained on a subset of the training set. The training data of each base classifier were created by sampling with replacement. Suppose D is the dataset, then each base classifier is trained using D_K samples. In bagging technique, prediction of base classifiers is achieved using the majority voting method, i.e. the base classifier with a maximum number of votes is selected as the final classification. RF [4] is an improved bagging algorithm. The RF algorithm uses DTs as base classifiers with sampling and replacement method for generating the training data for each base classifier. Each independent base classifier is trained using these bootstrap replicas. The RF consists of arbitrary number of trees, each vote for the most popular class to determine the outcome. It is robust against over-fitting. Detailed description about bagging algorithm is found in Ref. [4].

3.3.2 Boosting

In boosting, each base classifier is created sequentially where more weights are assigned to the next classifiers according to the error in the previous classifier. In this study, we used AdaBoost [9], which is one of the most popular boosting algorithms. It takes training set $S = \{(i_1, t_1),\ldots, (i_n, t_n)\}$ as input, where each i_i belongs to I and each label t_i belongs to $T = \{-1, 1\}$. For each iteration, $k = 1,\ldots$, K, AdaBoost calls a given base classifier, which acknowledges a succession of preparing illustrations S with a distribution or a set of weights over the preparation case, $D_k(i)$. The base classifier predicts the labels for such given inputs. In AdaBoost, a strong classifier is formed by combing weak classifiers, i.e. by calling the weak learner repeatedly with a different distribution of the training set. At each round, the weight of all incorrect cases is increased so that the next iteration classifier receives greater weights on all incorrect cases. This will improve the predictive power of weak classifiers. The predictions of all the classifiers are combined through weighted majority vote to form a final prediction. The detailed description about the AdaBoost algorithm is found in Ref. [9].

4 Results and Discussion

The proposed method was tested on 651 mammogram images, with 314 benign cases and 337 malignant cases, obtained from the DDSM. The database is available at www.marathon.csee.usf.edu/Mammography/Database.html [12, 13].

4.1 Parameter Selection for GA

As discussed in Section 3.1, optimal features are selected using genetic ensemble method. The GA uses the following parameters:

- Number of iterations: 50
- Population size: 10, 20, 30
- Crossover probability (Pc): 0.9
- Mutation Probability (Pm): 0.1

The parameter values of GA were selected by the trial and error method. An algorithm was tested for a population size of 10, 20, 30, 50, and 100, respectively, with a fixed number of iterations. The results obtained for population sizes of 10, 20, and 30 were found to be best in terms of accuracy of feature selection and computational time. For a population of size 50, an algorithm selects 13 features with an accuracy of 94.1%. Similarly, for a population of size of 100, an algorithm selects 14 features with an accuracy of 95.5%. It was observed that there was no significant improvement in the accuracy of feature selection with respect to increase in population size, but computational time increases very significantly. This is the main reason of selecting the population sizes of 10, 20, and 30.

An algorithm is also tested on crossover probability values of 0.7, 0.8, 0.9, respectively, and mutation probability values of 0.05, 0.1, and 0.2. The best outcome was obtained for Pc 0.9 and Pm 0.1. Both of these parameters influence the behavior and performance of GA [21]. These values are problem-specific.

4.2 Analysis of the Feature Selected Using Genetic Ensemble

The selection of the optimal feature subset depends on the fitness value. The fitness value of a solution is the mean classification accuracy of AdaBoost or RF classifiers. One can observe from Table 2 that a total of six feature subsets were selected using GA ensemble. Subsets F1 to F3 were selected using GA AdaBoost, and F4 to F6 were selected using GA RF. All the subsets included varying numbers of features, selected based on the mean accuracy of the classifier. Feature set F2 is the most significant feature subset having an accuracy of 96.15% with 11 features. This indicates that GA AdaBoost selects the best feature subset than does GA RF. Although GA RF selects feature subsets with fewer number of features compared with GA AdaBoost, the selection accuracy of GA AdaBoost is better. Frequency of selection of each feature is as shown in Figure 3. One can see from Figure 3 that feature numbers 6 (entropy), 13 (correlation), 20 (compactness), and 23 (contour roughness) appear in every subset. Likewise, feature numbers 1 (average gray level), 7 (energy), 14 (homogeneity), and 17 (sum entropy) appear in at least four subsets out of six. These features play a significant role in feature subset F2.

One interesting observation is that feature number 5 (uniformity) and feature number 25 (overlapping ratio) have zero frequency. These features have no impact in the proposed method.

4.3 Training of Classifiers

Feature subset F2 was used to train AdaBoost, RF, and DT classifiers, which were used to characterize masses. The training dataset (F2) consists of 651 samples of 11 features as shown in Figure 4.

The legitimacy of the outcomes created by the classifiers was guaranteed by a 10-fold cross-validation method. In 10-fold validation, input samples are randomly divided into 10 samples. Out of the 10 samples,

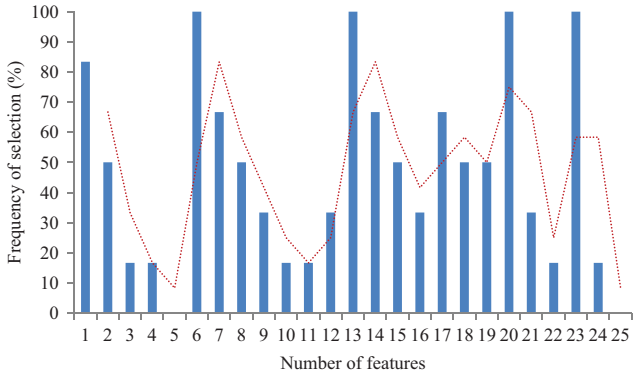

Figure 3: Rate of Features Selection.

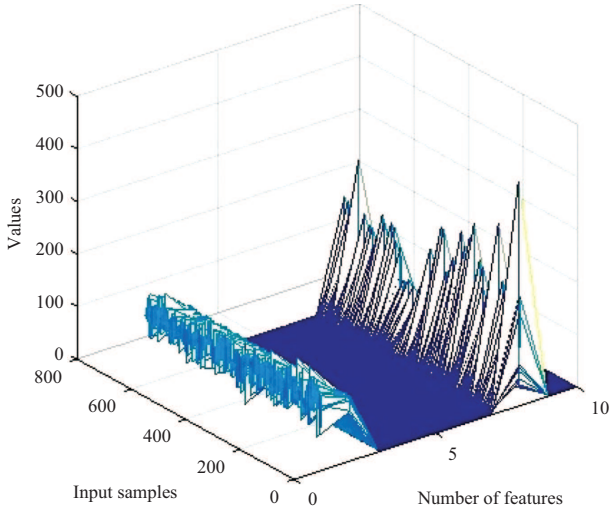

Figure 4: Training Data.

nine samples are used for training the classifier and the remaining one is used for testing the classifier. The proposed AdaBoost algorithm was implemented with '500' trees with a depth value of '30'; RF algorithm was implemented with '500' trees with three decision variables.

The progress of the error rate with respect to the number of tree for AdaBoost is as shown in Figure 5. One can observe from Figure 5 that the training error was reduced significantly from the 20th tree to the 150th tree, then remained steady up to the 500th tree. The training and out-of-bag (OOB) error for AdaBoost are 0.011 and 0.033, respectively.

The OOB error for RF algorithm is as shown in Figure 6. It has been observed that the error was reduced significantly with the induction of trees, and it was 0.158 at the 500th tree. These errors are useful for determining the optimal number of trees required for the construction of a classifier. From Figure 5, one can observe that the optimal number of trees required for implementing AdaBoost algorithm is 200, because the error rate is steady from the 200th tree to the 500th tree. Similarly from Figure 6, the optimal trees required for implementing RF algorithm were 325, as the error was slightly increased after the induction of trees from 325 to 500.

The single DT classifier was trained and tested using feature set F2. It was implemented with a minimum split value of 20, a maximum depth of 20, and a minimum number of buckets of 7.

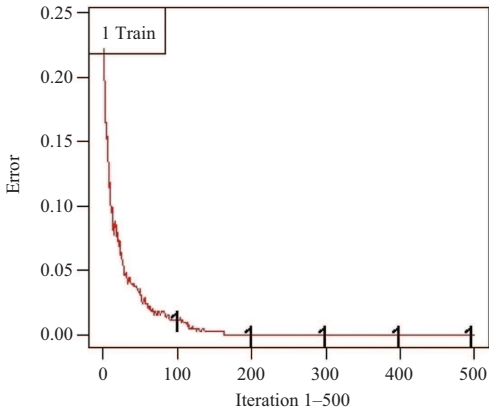

Figure 5: Training Error of AdaBoost Classifier.

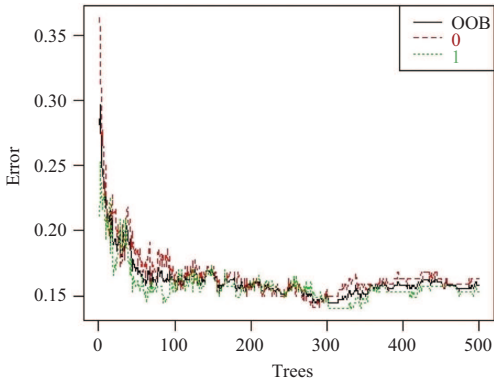

Figure 6: OOB Error During Training of RF.

4.4 Performance Evaluation Parameters

The parameters illustrated in Eq. (4) to Eq. (9) were used to evaluate the performance of AdaBoost, RF, and DT classifiers.

True positive rate (TPR) is also called sensitivity. It is defined as follows:

$$\text{TPR} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}} \tag{4}$$

True negative rate (TNR) is also called specificity. It is defined as follows:

$$\text{TNR} = \frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}} \tag{5}$$

Accuracy (ACC) is defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

Error is the rate of misclassification. It is defined as follows:

$$\text{Error} = \frac{\text{False Positive} + \text{False Negative}}{\text{True Positive} + \text{True Negative}} \tag{7}$$

Type I error (false alarm) is a defined as FPR:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{8}$$

Type II error (false alarm) is a defined as false negative rate (FNR):

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{9}$$

Area under the ROC curve is another important parameter to measure the accuracy of classifiers. It is a plot of TPR (sensitivity) against FPR (1-specificity), and its value lies between zero and one. The classification model is said to be 100% accurate if its value is one [33].

4.5 Performance Analysis of Classifiers for the Entire Dataset

The performance of the classifiers for the entire dataset consists of 25 features and 651 cases, as shown in Table 3. One can observe from Table 3 that the AdaBoost classifier is better than both RF and DT classifiers. AdaBoost achieves an accuracy of 91.55% with overall misclassification rate of 8.45%.

4.6 Performance Analysis of Classifiers for Optimal Feature Set

The summary of the classifiers performance for selected (optimal) feature set F2 is as shown in Table 4. One can observe from Table 4 that the AdaBoost ensemble is better than both RF and DT classifiers, while RF is better than DT with respect to all the parameters considered for evaluating the performance of classifiers. AdaBoost achieved the highest classification accuracy of 96.15% with 97.32% sensitivity and 95.9% specificity. RF outperformed the single DT classifier; it achieved an accuracy of 95.08% with 96.14% sensitivity and 93.94% specificity. The single DT classifier showed the lowest performance with an accuracy of 85.4%.

As far as misclassification rate (error) is concerned, AdaBoost outperformed both RF and single DT classifiers. AdaBoost has a misclassification rate of 3.85%, while RF and DT classifiers have 4.92% and 14.6%, respectively. The FPR or type I error and FNR or type II error for AdaBoost classifier have 5.09% and 2.67%, respectively. Similarly, type I and type II errors of RF classifier have 6.05% and 3.85%, respectively. For the DT classifier, it is 14.96% and 14.24%, respectively.

The above discussion shows that AdaBoost and RF ensemble classifiers are far better than individual classifiers such as single DT. These classifiers improve the classification rate and FPR very significantly. AdaBoost classifier improves the classification accuracy from 85.40% to 96.15%, a raise by 10.75% as of the single DT classifier. As far as type I is concerned, it reduces from 14.96% to 5.09% for AdaBoost.

The comparative analysis of results presented in Table 3 and Table 4 shows that the performance of the classifiers significantly improves with the use of the feature set selected by GA ensemble algorithm as compared to the entire dataset. The overall misclassification rate for AdaBoost reduces to 3.85% with the use of feature set F2 select by the GA ensemble as compared to 8.45% for the entire dataset.

Table 3: Performance of Classifiers for Entire Dataset.

Classifier	Number of features	TP (malign)	FN	TN (benign)	FP	Sensitivity TPR (%)	Specificity TNR (%)	Accuracy (%)	Error (%)
AdaBoost	25	313	24	283	31	92.87	90.12	91.55	8.45
Random Forest	25	310	27	282	32	91.98	89.80	90.93	9.07
Single DT	25	277	60	253	61	82.19	80.57	81.41	18.59

Table 4: Performance of Classifiers Using Feature Set F2.

Classifier	Number of Feature	TP (malign)	FN	TN (benign)	FP	Sensitivity TPR (%)	Specificity TNR (%)	Accuracy (%)	Error (%)
AdaBoost	11	328	09	298	16	97.32	95.90	96.15	3.85
Random Forest	11	324	13	295	19	96.14	93.94	95.08	4.92
Single DT	11	289	48	267	47	85.75	85.03	85.40	14.60

4.7 Performance Analysis of Classifiers for the Area Under ROC Curve

An area under ROC curve is another important parameter to measure the accuracy of classifiers. The ROC curve for the AdaBoost, RF, and DT classifiers is as shown in Figure 7, and calculated area under ROC curve (AUC) with 95% confidence interval is shown in Table 5. The area under curve for AdaBoost is $A_Z = 0.982 \pm 0.004$, and for the RF and DT classifiers, it is $A_Z = 0.986 \pm 0.004$ and $A_Z = 0.868 \pm 0.015$, respectively. It has been observed that both AdaBoost and RF ROC curve values are close to 1 and both are significantly better than that of the DT classifier. The RF classifier is slightly better than the AdaBoost classifier by 0.4%.

4.8 Impact of Intensity, Texture, and Shape features

The impact of the intensity, texture, and shape features on classifier performance was analyzed using 651 cases for reduced feature set F2. The results of analysis are as shown in Table 6. From Table 6, one can observe that shape features are better than both intensity and texture features. It achieves the highest accuracy of 92.16% for the AdaBoost classifier, while intensity features are, marginally, better than texture features. Both intensity and texture features achieved the highest accuracy of 91.7% and 90.93%, respectively. Analysis shows that shape-based features have significant impact on the classifiers' performance.

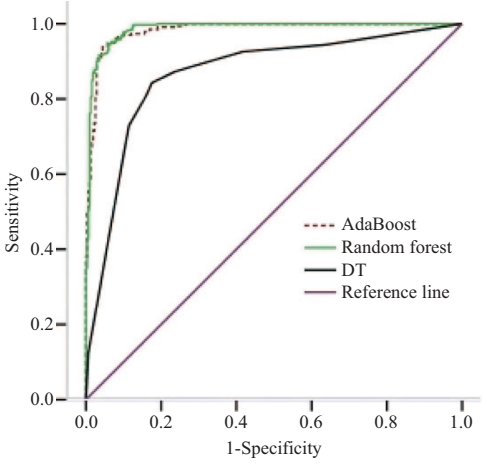

Figure 7: Comparison of ROC Curves of AdaBoost, RF, and DT Classifiers.

Table 5: Area Under Curve for AdaBoost, Random Forest, and DT Classifiers.

Classifier	Area (AUC)	Std. error	Asymptotic Sig	Asymptotic 95% confidence interval	
				LB	UB
AdaBoost	0.982	0.004	0.000	0.974	0.991
Random Forest	0.986	0.004	0.000	0.979	0.994
Decision Tree	0.868	0.015	0.000	0.838	0.897

Table 6: Impact of Intensity, Texture, and Shape Features on Classifiers Performance.

Features	AdaBoost			Random Forest			Decision Tree		
	TPR (%)	TNR (%)	Acc (%)	TPR (%)	TNR (%)	Acc (%)	TPR (%)	TNR (%)	Acc (%)
Intensity	90.20	90.76	90.47	91.69	91.71	91.7	67.35	83.75	71.12
Texture	90.20	90.12	90.16	91.39	90.44	90.93	75.66	71.97	73.88
Shape	92.87	91.40	92.16	91.98	91.40	91.7	75.37	80.57	77.8

Acc, accuracy; TNR, true negative rate/specificity; TPR, true positive rate/sensitivity.

Table 7: Comparison of Ensemble-based Methods.

Authors	Database	Total Cases used	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC Value
Elsayad [7]	DDSM	961	92.23	90.93	91.53	0.914
Mc Leod and Verma [24]	DDSM	100	–	–	91	–
Zhang et al. [40]	DDSM	543	–	–	72	–
Luo and Cheng [22]	DDSM	961	82.4	81.3	82.1	–
Banaem et al. [1]	DDSM	200	96.66	97.50	97	–
Proposed (GA-AdaBoost)	DDSM	651	97.32	95.90	96.15	0.982

4.9 Comparative Analysis of Proposed Method with Existing Ensemble-Based Systems

The results of the proposed method, compared with other studies in literature, are as shown in Table 7.

The comparative study shows that the proposed GA ensemble method is far better than the methods presented by other researchers. These methods are compared using statistical parameters like accuracy, sensitivity, specificity, and area under ROC curve. The method proposed in Ref. [1] showed a slightly better performance than the proposed one in terms of accuracy and specificity and a slightly lower performance for sensitivity. Even though the accuracy of proposed method is down by 0.85%, the AdaBoost method is better. As the success of the classifier depends on the feature selection and the number of cases used for testing the performance of the classifiers, the use of very few and unbalanced cases (benign and malignant) affects the performance of the classifier as well as feature selection process [16, 17]. Another reason is that the validity of the model is not tested using parameters such as area under ROC curve. The proposed GA ensemble method uses a large and balanced dataset, and it is validated using area under ROC curve. The GA ensemble method appears to be the best among all the methods presented in Table 7.

5 Computational Complexity

The computational complexity of an algorithm plays an important role in the design of an efficient algorithm. The complexity of the proposed GA ensemble algorithm (Figure 2) is O (N^{2}⋆ O(f)), where N^{2} is the complexity of GA (line nos. 8 and 9) and O(f) is the complexity of fitness function (line no. 12).

6 Conclusion

In this article, authors examined the efficient methods for feature selection and classification of masses in digital mammograms. The six subsets of important features were chosen using the genetic ensemble method with 10-fold cross-validation. The most significant feature subset that achieved the highest classification accuracy was used as input for training and testing three classifiers – AdaBoost, RF, and single DT. The results of the methods demonstrate that AdaBoost is better than both RF and single DT. Both AdaBoost and RF achieved high classification accuracy and low FPR. The outcome of the proposed method proves that ensemble classifiers are far better than single classifiers. The proposed method helps to improve breast cancer diagnosis. In the future, we will focus on minimizing the misclassification error.

Bibliography

[1] H. Banaem, A. Dehnavi and M. Shahnazi, Ensemble supervised classification method using the regions of interest and grey level co-occurrence matrices features for mammograms data, *Iranian J. Radiol.* **12** (2015), 1–8.

[2] M. Berks, Z. Chen, S. Astley and C. Taylor, Detecting and classifying linear structures in mammograms using random forests. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6801 LNCS, pp. 510–524, (2011).

[3] L. Breiman, Bagging predictors, *Machine Learning* **24** (1996), 123–140.
[4] L. Breiman, Random Forests, *Machine Learning* **45** (2001), 5–32.
[5] J. Y. Choi, D. H. Kim, K. N. Plataniotis and Y. M. Ro, Computer-aided detection (CAD) of breast masses in mammography: combined detection and ensemble classification, *Phys. Med. Biol.* **59** (2014), 3697.
[6] A. Dong and B. Wang, Feature selection and analysis on mammogram classification, in: *IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing – Proceedings*, pp. 731–735, Canada, 2009.
[7] A. Elsayad, Predicting the severity of breast masses with ensemble of Bayesian classifiers, *J. Comput. Sci.* **6** (2010), 576–584.
[8] D. B. Fogel, *Evolutionary computation: The fossil Record*, IEEE Press Piscataway, NJ, 1998.
[9] Y. Freund and R. E. Schapire, A desicion-theoretic generalization of on-line learning and an application to boosting, in: *European Conference on Computational Learning Theory*. pp. 23–37, Springer, Berlin, Heidelberg, 1995.
[10] J. Fu, S. Lee, S. Wong, J. Yeh, A. Wang and H. Wu, Image segmentation feature selection and pattern classification for mammographic microcalcifications, *Comput. Med. Imaging Graphics* **29** (2005), 419–429.
[11] D. Goldberg, *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, Reading, MA, 1989.
[12] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer Jr, R. Moore, K. Chang and S. Munishkumaran, Current status of the digital database for screening mammography, in: *Digital Mammography. Computational Imaging and Vision*, vol. 13, N. Karssemeijer, M. Thijssen, J. Hendriks and L. van Erning, eds., pp. 457–460, Springer, Dordrecht, 1998.
[13] M. Heath, K. Bowyer, D. Kopans, R. Moore and P. Kegelmeyer, The digital database for screening mammography, in: *Proceedings of the Fifth International Workshop on Digital Mammography*, 2001, 212–218.
[14] J. Holland, *Adaptation in natural and artificial systems*, Ann Arbor MI University of Michigan Press, Ann Arbor, 1975, 183.
[15] G. Jothi, H. Inbarani and A. Azar, Hybrid tolerance rough set: PSO based supervised feature selection for digital mammogram images, *Int. J. Fuzzy System Appl.* **3** (2013) 15–30.
[16] M. Kubat and S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, *In ICML* **97** (1997), 179–186.
[17] M. Kubat, R. C. Holte and S. Matwin, Machine learning for the detection of oil spills in satellite radar images, *Machine Learning* **30** (1998), 195–215.
[18] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, John Wiley & Sons, Hoboken, NJ, USA, 2004.
[19] H. Li, Y. Wang, K. Liu, S. Lo and M. Freedman, Computerized radiographic mass detection – part II: decision support by featured database visualization and modular neural networks, *IEEE Trans. Med. Imaging* **20** (2001), 302–313.
[20] N. Li, H. Zhou, J. Ling and Z. Zhou, Spiculated lesion detection in digital mammogram based on artificial neural network ensemble, in: *International Symposium on Neural Networks*, pp. 790–795, Springer, Berlin, Heidelberg, 2005.
[21] W. Lin, W. Lee and T. Hong, Adapting crossover and mutation rates in genetic algorithms, *J. Inf. Sci. Eng.* **19** (2003), 889–903.
[22] S.-T. Luo and B.-W. Cheng, Diagnosing breast masses in digital mammography using feature selection and ensemble methods, *J. Med. Syst.* **36** (2012), 569–577.
[23] M. Mafarja and S. Mirjalili, Hybrid whale optimization algorithm with simulated annealing for feature selection, *Neurocomputing* **260** (2017), 302–312.
[24] P. Mc Leod and B. Verma, Clustered ensemble neural network for breast mass classification in digital mammography, in: *Proceedings of the International Joint Conference on Neural Networks*, Australia (2012).
[25] M. Mitchell, *An introduction to genetic algorithms (complex adaptive systems)*, p. 221, The MIT Press, Cambridge, MA, 1998.
[26] F. Mohanty, S. Rup, B. Dash, B. Majhi and M. N. S. Swamy, (2018). Mammogram classification using contourlet features with forest optimization-based feature selection approach, *Multimedia Tools Appl.* 1–30. https://doi.org/10.1007/s11042-018-5804-0.
[27] National Cancer Institute, *Cancer Stat Fact Sheets: Cancer of the Breast*. (2009). Available at: http://seer.cancer.gov/statfacts/html/breast.html.
[28] R. Ramos, M. Nascimento and D. Pereira, Texture extraction: an evaluation of ridgelet, wavelet and co-occurrence based methods applied to mammograms, *Expert Syst. Appl.* **39** (2012), 11036–11047.
[29] J. Rodríguez, L. Kuncheva and C. Alonso, Rotation forest: a new classifier ensemble method, *IEEE Trans. Pattern Anal. Machine Intell.* **28** (2006), 1619–1630.
[30] M. Sameti, R. Ward, J. Morgan-Parkes and B. Palcic, A method for detection of malignant masses in digitized mammograms using a fuzzy segmentation algorithm, *Eng. Med. Biol. Soc.* 1997. *Proceedings of the 19th Annual International Conference of the IEEE,* **2** (1997), pp. 513–516, vol. 2.
[31] R. Schapire, The strength of weak learnability, *Machine Learning* **5** (1990), 197–227.
[32] P. Somol, P. Pudil, J. Novovičová and P. Paclík, Adaptive floating search methods in feature selection, *Pattern Recognit. Lett.* **20** (1999), 1157–1163.
[33] J. Swets, Measuring the accuracy of diagnostic systems, *Science* **240** (1988), 1285–1293.
[34] R. Swiniarski and A. Skowron, Rough set methods in feature selection and recognition, *Pattern Recognit. Lett.* **24** (2003), 833–849.

[35] A. Tanwani, J. Afridi, M. Shafiq and M. Farooq, Guidelines to select machine learning scheme for classification of biomedical datasets, *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5483 LNCS, (2009), pp. 128–139.
[36] S. Thawkar and R. Ingolikar, Automatic detection and classification of masses in digital mammograms, *Int. J. Intell. Eng. Syst.* **10** (2017) 65–74.
[37] S. Thawkar and R. Ingolikar, Efficient approach for the classification of masses in digital mammograms, *Int. J. Innovative Computing Inf. Control* **13** (2017) 967–978.
[38] J. Yang and V. Honavar, Feature subset selection using a genetic algorithm, *IEEE Intell. Syst. Appl.* **13** (1998), 44–49.
[39] Z. Zhang and P. Yang, An ensemble of classifiers with genetic algorithm based feature selection, *IEEE Intell. Inf. Bull.* **9** (2008), 18–24.
[40] Y. Zhang, N. Tomuro, J. Furst, and D. Raicu, Building an ensemble system for diagnosing masses in mammograms, *Int. J. Comput. Assisted Radiol. Surgery* **7** (2012), 323–329.