# Modified Artificial Bee Colony Based Feature Selection: A New Method in the Application of Mammogram Image Classification

S. Shanthi[1], V. Murali Bhaskaran[2]
[1] Department of Computer Applications, Kongu Engineering College, Tamil Nadu, India

[2] Principal, Dhirajlal Gandhi College of Technology, Tamil Nadu, India

***Abstract*— Mammography has been one of the most reliable methods for early detection of breast cancer. There are different signs of breast cancer such as microcalcifications, masses, architectural distortions and bilateral asymmetry. To detect all four signs of cancer lesions in a mammogram image, large set of features are extracted based on Gabor filters, fractal analysis, directional analysis, and multiscale surrounding region dependence method. All the extracted features do not help in detection of abnormality in a mammogram, so it is proposed to select the best feature set to improve classification accuracy. A good feature selection method may not only improve the classification accuracy of the final classifier, but also reduce the computational complexity of it. Hence, we proposed modified artificial bee colony based feature selection (MABCFS) technique to select the predominant feature set in classification of breast lesion in mammogram images. The experiments have been conducted on two benchmark data sets (MIAS & DDSM) and SRAN classifier classifies the mammogram into normal, benign or malignant. Performance of MABCFS is compared withArtificial Bee Colony (ABC) optimization, Ant Colony Optimization (ACO), and Genetic Algorithm (GA) and the results have been proved to be progressive.***

***Index Terms: Feature selection. Artificial Bee Colony, Mammogram.***

## I. INTRODUCTION

The extracted feature space is very large and complex due to the wide diversity of the normal tissues and the different signs of abnormalities. Also, the learning algorithm used is slowed down unnecessarily due to higher dimensions of the feature space, while experiencing lower prediction accuracies due to learning irrelevant information. As a result, it is necessary to have the feature selection in CAD system to select the subset of features that represent the whole set without losing of information.

More recently, nature inspired algorithms are used for feature selection. In literature, Genetic Algorithm (GA), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Simulated Annealing (SA) are used in numerous applications for feature selection and also in optimization problems [1], [2].

Artificial bee colony (ABC) is a stochastic, nature inspired, swarm intelligent algorithm proposed by Karaboga [3] for constrained optimization problems. Since its proposal, ABC has been proved to be successful in solving optimization problems in numerous application domains [4]. Also ABC is proved to give promising and enhanced results in the areas where GA, ACO, PSO, differential evolution algorithm and evolution strategies have given already [5].

In this paper we proposed a modified artificial bee colony based feature selection (MABCFS) method to select optimal feature subset from high dimensional data with improved diagnosis ability. The remaining of the paper is organized as follows: Section II describes methodology of the proposed work. Experimental results obtained through the application of the proposed method are discussed in Section III. Finally, the Section IV concludes the work

## II. METHODOLOGY

The proposed method contains the following stages: collection of mammographic image database, removing the label; background; pectoral region (pre-processing), ROI identification, feature extraction, feature selection and classification. Figure 1 shows the sequence of different stages of the proposed methodology.

### *A. Mammographic Image Database*

For the current study, the images are taken from two public and widely known databases: the Mammographic Image Analysis Society (MIAS) database [6] and Digital database for screening mammography (DDSM) database [7].

The MIAS data set consists of 322 images belonging to three big categories: normal, benign and malign which indicate different classes of abnormalities such as calcification (CALC), well-defined circumscribed masses (CIRC), speculated masses (SPIC), ill-defined masses (MISC), architectural distortion (ARCH), asymmetry (ASYM) and normal. The data set consists of 208 normal, 63 benign and 51 malign images.



ISSN: 2278 – 7798 

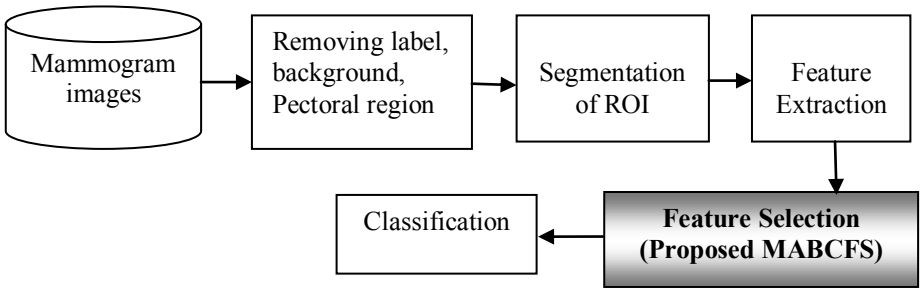


Figure 1 Outline of the proposed method

### B. *Pre-processing, ROI identification and Feature Extraction*

In a typical mammogram, several areas (noise) such as image background, digitization noises, informative marks and pectoral region and so on are present. Prior to ROI identification all the noise should be removed and ROI is identified using intutionistic fuzzy-c means clustering [8].

A combination of features namely multiscale surrounding region dependence method features, fractal dimension features, Gabor features, and directional and morphological features are extracted from the suspicious region or fibroglandular disks of the mammogram images [9]-[12]. Table 1 list the different type of features used in this work.

A total of 84 features are extracted to classify the ROI into normal or benign or malignant.

**Table 1** List of features extracted to classify the abnormalities

| Type of features | Feature description |
|---|---|
| MSRDM features ($F_1$ to $F_{64}$) | Horizontal Weighted Sum (HWS), Vertical Weighted Sum (VWS), Diagonal Weighted Sum (DWS) and Grid Weighted Sum (GWS) |
| Directional features ($F_{65}$ to $F_{69}$) | First order moment, second order angular moments, entropy, dominant orientation and circular variance |
| Morphological features ($F_{70}$ to $F_{81}$) | Seven Hu's moments ($\Phi_1$ *to* $\Phi 7$), area, average density, eccentricity, and stretch |
| Gabor Features ($F_{82}$) | Energy |
| Fractal Dimension features ($F_{83}$to $F_{84}$) | Fractal dimension and Lacunarity |

### C. *Feature Selection*

A classic feature selection procedure has four basic steps, namely, subset generation, subset evaluation, stopping criterion, and result validation. Subset generation is a search procedure that produces feature subsets for evaluation based on a certain search strategy. Each subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Then, the selected best subset is validated by prior knowledge. Feature selection techniques provide three main benefits when constructing predictive models. They are

- Improved model interpretability,
- Shorter training times and
- Enhanced generalization by reducing overfitting.

## 1 Artificial bee colony (ABC) optimization

Artificial Bee Colony (ABC) algorithm was proposed by Karaboga for optimizing numerical problems [3]. The algorithm simulates the intelligent foraging behavior of honey bee swarms. In the ABC model, the colony consists of three groups of bees: employed bees, onlookers and scouts. It is assumed that there is only one artificial employed bee for each food source. Employed bees go to their food source and come back to hive and dance on this area. A bee waiting on the dance area for making a decision to choose a food source is called onlooker bee. The other kind of bee is scout bee that carries out random search for discovering new sources. The position of the food source represents a possible solution to the optimization problem and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution calculated by equation (1).

$$fit_i = \frac{1}{1+f_i} \tag{1}$$

An artificial onlooker bee selects a food source depending on the probability value associated with that food source $p_i$ calculated using equation (2)

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \tag{2}$$

Where SN is the number of food sources equal to the number of employed bees, and $fit_i$ is the fitness of the solution given in equation (1). An employed bee produces a modification on the position of the food source (solution) in her memory depending on local information (visual information) and finds a neighboring food source, and then evaluates its quality. In ABC, finding a neighboring food source is defined in equation (3).

$$v_{ij} = z_{ij} + \emptyset_{ij}(z_{ij} - z_{kj}) \tag{3}$$

In equation (3) j∈{ 1,2,...D} and k∈{1,2,...SN} are randomly chosen indexes that has to be different from i. $\emptyset_{ij}$

  

is uniformally distributed random numbers between [-1,1]. It controls the production of neighbor food sources around $z_{i,j}$ and represents the comparison of two food positions visible to a bee. The food source of which the nectar is abandoned by the bees is replaced with a new food source by the scouts. Assume that the abandoned source is $z_i$ and $j \in \{1, 2, \ldots, D\}$, then the scout discovers a new food source to be replaced with $z_i$. This operation can be defined as in equation (4).

$$z_i^j = z_{min}^j + rand(0,1)(z_{max}^j - z_{min}^j) \quad (4)$$

After each candidate source position $v_{i,j}$ is produced and then evaluated by the artificial bee, its performance is compared with that of its old one. If the new food source has equal or better nectar than the old source, it is replaced with the old one in the memory. Otherwise, the old one is retained in the memory. A general pseudo code for ABC optimization approach is shown in Algorithm1

**Algorithm 1: Artificial bee colony optimization**

1:Initialization phase
2. repeat
3: Employed bees phase
4: Onlooker bees phase
5: Scout bees phase
6: Memorize the best solution achieved so far
7: until (cycle=maximum cycle number or Maximum CPU time)

## 2. Modified Artificial Bee Colony Based Feature Selection

In ABC algorithm, the employed bees explore the new food source and communicate this information to the onlooker bees; whereas the onlooker bees exploit the food sources which are explored by the employed bees. The search equation proposed in ABC algorithm is good at exploration but poor at exploitation, so that will affect the convergence speed of the algorithm. That is, while producing a new solution, $v_i$, changing only one parameter of the parent solution $x_{i,}$ results in a slow convergence rate. In order to improve the exploitation ability of ABC algorithm, the global best solution will be considered in the modified search equation. The search equation (3) in ABC algorithm is modified and defined in equation (5).

$$v_{ij} = z_{ij} + \emptyset_{ij}(z_{ij} - z_{kj}) + \varphi_{ij}(w_j - z_{ij}) \quad (5)$$

Where k∈ {1,2,…. SN} is a random selected index which is different from i; j ∈ { 1,2,...D} is a random selected index; $w_j$ is the jth element of the global best solution; $\emptyset_{ij} \in [-1,1]$ and $\varphi_{ij} \in [0,1.5]$ are both uniformly distributed random numbers.

Differential evoluation (DE) is a population based algorithm, whose main strategy is to generate a new position for an individual by calculating vector differences between other randomly selected members in the population. "DE/current-to-rand/1" is a variant DE mutation strategy, which can effectively maintain population diversity according to randomness of the search equation. Based on this mutation strategy, a new search equation in employee bee stage is proposed as follows.

$$v_{ij} = z_{ij} + \emptyset_{ij}(z_{ij} - z_{kj}) + \varphi_{ij}(w_j - z_{ij}) \quad (6)$$

Where k∈ {1,2,…. SN} is a random selected index which is different from i; j ∈ { 1,2,...D} is a random selected index; $w_j$ is the jth element of the global best solution; $\emptyset_{ij} \in [-1,1]$ and $\varphi_{ij} \in [-1,-1]$ are both uniformly distributed random numbers; $\emptyset_{ij}$ and $\varphi_{ij}$ are both negative or both positive, which can keep the same search direction. A general pseudo code for Modified ABC optimization approach is shown in Algorithm 2.

**Algorithm 2: Modified Artificial bee colony optimization**

1:Initialization phase
2. repeat
3: Employed bees phase: Search food source for employee bee according to the equation (6) and evaluate its quality that is accuracy of the food source.
3: Select the better solution between the new food source and the old food source.
4: If solution does not improve $trail_i$=$trail_i$+1; otherwise $trail_{i=0}$
5: Onlooker bees phase: Search the new food source for onlooker bees using equation (5) and evaluate its quality.
6 Apply greedy selection process and select the better solution between new and old food source.
5: Memorize the best solution achieved so far.
6: Scout bees phase: Find abandoned food source and produce new scout bees.
7: until (cycle=maximum cycle number or Maximum CPU time)

### *D. Classification*

The Self-adaptive Resource Allocation Network (SRAN) is a recently proposed sequential learning algorithm with a self-regulated control mechanism built in it [13]. In SRAN, when the training sample arrives one-by-one, it computes first the difference in knowledge acquired by the network and the information present in current sample. If the difference is more, either the sample participates in the learning or the sample is pushed into the data stack for later use. If the difference is small, all these samples are deleted from the training set in order to avoid over training. Also, based on this difference, the network adapts its control parameters. Since, the SRAN classifier uses explicit classification error in growing/learning criterion and discarding similar samples, it prevents overtraining and provides better generalization performance. The dataset is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average accuracy across all k trials is computed.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

where

Where, TP=>True Positive: Correctly identified.



ISSN: 2278 – 7798 

TN=>True Negative: Correctly rejected.
FN=>False Negative: Incorrectly rejected.
FP =>False Positive: Incorrectly identified.

## III. Experimental Results and Discussion

A combination of 84 features is extracted and all the features may not be used for classification. The proposed feature selection method is applied to select the most discriminating features. Then the SRAN classifier is used for classification and 10 fold cross validation is carried out. The classification accuracy is used to evaluate the proposed feature selection method and the performance of the method is compared with GA, PSO and ABC base feature selection methods.

Table 1 shows the number of features selected in the best solution obtained by different feature selection algorithms. Out of 84 features GA, PSO and PSO selects 50, 62 and 56 respectively, where as the proposed MABCFS selects only 42 features for the images from MIAS dataset. From the Table1 and Table 2 we can see that MABCFS selects the smallest number of features while maintain the high accuracy of classification.

**Table 1**. No. of features selected by various methods

| Dataset/ Method | GA | PSO | ABC | Proposed (MABCFS) |
|---|---|---|---|---|
| MIAS | 50 | 56 | 45 | 42 |
| DDSM | 51 | 56 | 45 | 42 |

**Table 2** Classification Accuracy (%) for various feature selection methods

| Feature Selection Methods | DataSet | |
|---|---|---|
| | MIAS | DDSM |
| Using all feature | 95.96 | 96.67 |
| GA | 95.96 | 97 |
| PSO | 96.27 | 96.83 |
| ABC | 96.27 | 97 |
| Proposed (MABCFS) | 96.89 | 97.17 |

From Table 2 and Table 3 we can see that MABCFS selects the smallest number of features while maintain the high accuracy of classification.

In PSO, a new position vector is calculated using the particle's current and best solution and the swarm's best solution; where as in MABCFS, the new solution vector is calculated using the employed bee's current solution and randomly chosen solution. While GA employ crossover operators to produce new or candidate solutions present one, MABCFS used equation (6) to produce candidate solutions. Furthermore PSO and GA has more control parameters than MABCS. The new search mechanism in MABCFS can balance the exploration and exploitation capabilities very well, which can both maintain the diversity and improve the convergence speed.

The performance of MABCFS is good in terms of local and global optimization due to the selection schemes employed and the neighboring production mechanism used. MABFS balances exploration and exploitation efficiently than ABC.

## IV. Conclusion

. This paper has presented a novel approach for feature selection based on a relatively new swarm intelligence algorithm, namely modified artificial bee colony based feature selection, to solve feature selection problems. The use of feature selection techniques has successfully selected relevant feature subset to improve the classification performance. As reported in Table 2 the highest accuracy is achieved by using the proposed feature selection techniques. Further work will include investigation of the MABSFS performance with large number of real databases in a clinical environment.

.

## REFERENCES


[1] Babaoglu, I., Findik, O. and Ulker, E.(2010) 'A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine', Expert Systems with Applications, vol. 37, pp. 3177–3183.

[2] Hendrawan, Y. and Murase, H. (2011) 'Bio-inspired feature selection to select informative image features for determining water content of cultured Sunagoke moss', Expert Systems with Applications, vol. 38, pp.4321–14335.

[3] Karaboga, D. (2005) 'An idea based on honey bee swarm for numerical optimization', Technical report TR06, Computer Engineering Department, Erciyes University, Turkey, 2005.

[4] Karaboga, D., Gorkemli,B., Ozturk, C., and Karaboga, N., (2012) 'A comprehensive survey: artificial bee colony (ABC) algorithm and applications', Artif Intell Rev, DOI 10.1007/s10462-012-9328-0.

[5] Karaboga, D., Akay, B., (2009) 'A comparative study of Artificial Bee Colony algorithm' Applied Mathematics and Computation, vol. 214, pp.108–132.

[6] Suckling, J. et al.: 'The Mammographic Image Analysis Society digital mammogram database', *Proc. Int. Workshop Dig. Mammography*, 1994, pp. 211–221.

[7] Heath, M., Bowyer,K., Kopans,D., Moor, R., and Kegelmeyer, W,P., (2001). Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed., 2001, pp. 212-218, *Medical Physics Publishing*, ISBN 1-930524-00-5.

[8] Shanthi, S. and Murali Bhaskaran, V. (2011) 'Intuitionistic Fuzzy C-Means and Decision Tree Approach for Breast Cancer Detection and Classification', *European Journal of Scientific Research*, vol. 66, No. 3, pp.345-351.

[9] Shanthi, S. and Murali Bhaskaran, V. (2013) 'A Novel Approach for Detecting and Classifying Breast Cancer in Mammogram Images', *International Journal of Intelligent Information Technologies*, vol. 9, No. 1, pp.20-38.

[10] Shanthi, S. and Murali Bhaskaran, V. (2014) 'A Novel Approach for Classification of Abnormalities in Digitized Mammograms', Sadhana -Academy Proceedings in Engineering Science journal, accepted.

[11] Banik, S., Rangayyan, R M., and Desautels, J E L .(2013) 'Measures of angular spread and entropy for the detection of architectural distortion in prior mammograms' , *International Journal of Computer Assisted Radiology and Surgery*, vol. 8, pp. 121–134.

[12] Rangayyan, R. M., Ferrari, R. J., and Frere, A. F. (2007) 'Analysis of bilateral asymmetry in mammograms using directional, morphological, and density features', SPIE Proceedings vol. 16, *Journal of Electronic Imaging*, vol. 16, No. 01, pp.*013003-1-12.*

[13] Suresh, S, Dong, K & Kim, HJ (2010) 'A sequential learning algorithm for self-adaptive resource allocation network classifier', Neurocomputing, vol. 73, no. 16-18, pp. 3012–3019.


  