RESEARCH ARTICLE

**Editorial Process:** Submission:01/28/2019 Acceptance:08/11/2019

# Comparison Analysis of Linear Discriminant Analysis and Cuckoo-Search Algorithm in the Classification of Breast Cancer from Digital Mammograms

**Sannasi Chakravarthy S R*, Harikumar Rajaguru**

## Abstract

**Objective:** Breast cancer is the most common invasive severity which leads to the second primary cause of death among women. The objective of this paper is to propose a computer-aided approach for the breast cancer classification from the digital mammograms. **Methods:** Designing an effective classification approach will assist in resolving the difficulties in analyzing digital mammograms. The proposed work utilized the Mammogram Image Analysis Society (MIAS) database for the analysis of breast cancer. Five distinct wavelet families are used for extraction of features from the mammograms of MIAS database. These extracted features are statistical in nature and served as input to the Linear Discriminant Analysis (LDA) and Cuckoo-Search Algorithm (CSA) classifiers. **Results:** Error rate, Sensitivity, Specificity and Accuracy are the performance measures used and the obtained results clearly state that the CSA used as a classifier affords an accuracy of 97.5% while compared with the LDA classifier. *Conclusion:* The results of comparative performance analysis show that the CSA classifier outperforms the performance of LDA in terms of breast cancer classification.

**Keywords:** Breast cancer- mammogram- discriminant analysis- cuckoo

*Asian Pac J Cancer Prev,* **20 (8)**, 2333-2337

## Introduction

The increasing risk of breast cancer severity is more for the women in their entire lifetime. The sickness is fairly more among middle-aged than in young-aged women (Henriksen et al., 2019). Based on the information of the American Cancer Society, breast cancer is the second leading cancer next to lung cancer among women and it occurs quite rare for men (DeSantis et al., 2013). The risk of breast cancer initiates in the breast cells of the human body and these affected cells can easily invade its neighbour tissues. The illness effect may be subject to the cancer type, risk level and patients' oldness. In general, the identification of breast cancer is found either by perceiving a lump in the breast or through mammogram screening (Falk et al., 2018). This lump is classified as either benign or malignant tumours; where benign tumours are normal or controllable and malignant tumours are abnormal or vigorous in invading the adjacent tissues. The clinical physicians use microscopic study for the detection and classification of breast cancer conventionally (Documet et al., 2015). But nowadays, machine learning algorithms that utilize the computational techniques pave them an easier way for the analysis and classification of cancer.

The detection at an earlier stage is very significant in the diagnosis of breast cancer. This is attained by the use of mammogram screening; where mammogram gives an x-ray imaging of the patients' breast (Sundaram et al., 2011). The sign or symptom of microcalcification in the breast is straightforwardly identified with the help of mammogram. The microcalcification is typically a tiny deposit of calcium in the breast area which is found as spots (white) on the obtained mammograms images (Peairs et al., 2017).

The study herein proposes a computer-aided diagnosis for the breast cancer classification as either benign class or malignant class from digital mammograms. The digital mammogram images used for the analysis are obtained from the Mammogram Image Analysis Society (MIAS) dataset. The work involves the performance analysis by considering 80 mammogram images from the MIAS database; where the first 40 mammogram images are benign and the rest 40 are malignant. All the works in this paper are carried out using MATLAB 2013a.

## Materials and Methods

*A. Feature Extraction*

The examination of breast cancer in this work starts with the extraction of features from the taken mammogram

*Department of Electronics and Communication Engineering, Anna University (Bannari Amman Institute of Technology), Sathyamangalam, India. *For Correspondence: elektroniqz@gmail.com*

images in the MIAS database as in Figure 1. The feature extraction is intended to transform the input raw mammograms into a set of features that can be provided as input to any classification algorithm (Abirami et al., 2017). The feature extraction is attained using Discrete Wavelet Transform (DWT). Daubechies (DB4), Haar (HAAR), BiorSplines (BIOR4.4), Symlet8 (SYM8) and DMeyer (DMEY) are the five preferred wavelet families used for feature extraction from the raw-input mammogram images at level 4 decomposition (Abirami et al., 2017). This step will offer twelve statistical features as Mean, Median, Mode, Maximum, Minimum, Range, Standard Deviation, Median Absolute Deviation, Mean Absolute Deviation, L1 norm, L2 norm and Max norm from the input mammograms (Amin et al., 2015).

These extracted features are then served as input to the Linear Discriminant Analysis (LDA) and Cuckoo-Search Algorithm (CSA) classifiers which are discussed in the following subsection.

*B. Classification using LDA*

The LDA as a classifier is analogous to the Principal Component Analysis (PCA) method. The PCA method intends to determine the component axes which maximize the variance of given data. As analogous to PCA, the LDA classifier similarly determines the component axes which maximize the variance of given data and additionally it determines the axes which maximize the separation among different output classes. This could be useful in classification problems (Mika et al., 1999). The Bayes classifier gives a highly effective classification result that assigns each input features to the most likely output class given its predictor values. If the terms are properly specified, then the Bayes classifier has the minimum possible error rate than all other classifiers (Tharwat et al., 2017). Thus LDA as a classifier that efforts to approximate the Bayes classifier (Kan et al., 2015). The five general steps of LDA classifier is

i. Calculate the d-dimensional mean vectors for various classes from the dataset.

ii. Calculate the scatter matrices (both within-class and between-class).

iii. Calculate the Eigen vectors ($e_1$,$e_2$,...,$e_d$) and respective Eigen values ($\lambda_1$, $\lambda_2$,...,$\lambda_d$) for the above calculated scatter matrices.

iv. Now sort the Eigen vectors by reducing order of Eigen values and select k Eigen vectors with maximum value of Eigen values to generate a d × k dimensional matrix W.

v. Use this above generated Eigen vector matrix W to renovate the samples onto the different subspace. And it can be represented shortly by matrix multiplication as Y=X×W.

*C. Classification using CSA*

The same set of twelve statistical features are now served as an input to the CSA classifier. The CSA is a popular nature-inspired and metaheuristic optimization technique that depends on the exciting breeding characteristics such as brood parasitism of certain species of cuckoos (Yang et al., 2010). CSA is principally imitates the breeding behaviour of cuckoo, which involves dumping of eggs inside the nests of other host birds and making these host to nurture their chicks. The traditional cuckoo-search algorithm is modified for the proposed problem. Levy flight denotes the random flight characteristics of birds and is performed to obtain the next position $p_i^{(t+1)}$ based on the current position $p_i^{(t)}$ as

$$p_i^{(t+1)} = p_i^{(t)} \oplus \alpha^{Levy\,(\gamma)} \quad (1)$$

where ⊕ and α represent the entry-wise multiplication and step size. In general, α>0 should be taken as scaling factor for step size; here α=1 is considered for the classification problem. The random walk is observed by using levy flight where the random step size is calculated from the levy distribution (Gandomi et al., 2013) as,

$$Levy \sim u = t^{-\gamma} \quad (2)$$

where the value of γ is kept as 0.2 for classification which indicates the infinite variance with infinite mean. Now the cuckoo's successive steps principally provide a random-walk approach and this follows power-law like step length distribution using a heavy tail (Cui et al., 2017). The Mantegna algorithm (Gandomi et al., 2013) is used for a symmetric Levy stable distribution.

The three general rules of cuckoo-search algorithm can be summarized as:

i. Each cuckoo bird lays an egg at a time and then dumps those egg in a host nest randomly.

ii. The best host nests with supreme quality eggs will be conceded over to the subsequent generations.

iii. The amount of host nests available is constant and the cuckoo's egg is revealed by the host one with a probability $p_a \in [0,1]$ . Now the host one can either dispose of the cuckoo's egg away or just remove the nest in order to construct a newer nest in another new location.

The proposed work considers the target value for benign and malignant images as 0.1 and 0.85. The LDA and CSA classifiers are evaluated using performance metrics is discussed in the succeeding section.

## Results

The comparison of performance analysis of the classifiers are done using the standard metrics like Error Rate (ER), Sensitivity (SE), Specificity (SP), Precision (PR), Accuracy (ACC) and Matthews Correlation Coefficient (MCC) (Zhu et al., 2010). These metrics are calculated based on the confusion matrix (Powers, 2011) that is made of True Positives (TP), False Negatives (FN), True Negatives (TN) and False Positives (FP).

The Table 2 gives the comparative analysis of LDA and CSA classifiers used for the classification of mammograms as either benign or malignant and its graphical analysis is shown in Figure 2.

As in Table 1, the LDA classifier along with dmey wavelet misclassifies 10 malignant mammogram images as benign and 16 benign mammogram images

as malignant out of 80 total mammogram images; this misclassification leads to the highest error rate of

Table 1. Confusion Matrix of LDA and CSA Classifiers

| Classifier | Wavelet | TP | FN | TN | FP |
|---|---|---|---|---|---|
| LDA | DB4 | 26 | 14 | 30 | 10 |
| | HAAR | 32 | 8 | 34 | 6 |
| | BIOR4.4 | 34 | 6 | 28 | 12 |
| | SYM8 | 32 | 8 | 36 | 4 |
| | DMEY | 30 | 10 | 24 | 16 |
| CSA | DB4 | 28 | 12 | 30 | 10 |
| | HAAR | 40 | 0 | 38 | 2 |
| | BIOR4.4 | 36 | 4 | 26 | 14 |
| | SYM8 | 36 | 4 | 34 | 6 |
| | DMEY | 34 | 6 | 32 | 8 |

TP, True Positive; FN, False Negative; TN, True Negative; FP, False Positive, LDA, Linear Discriminant Analysis; CSA, Cuckoo-Search Algorithm; DB4, Daubechies; HAAR, Haar; BIOR4.4, Bior Splines; SYM8, Symlet 8; DMEY, DMeyer

Table 2. Performance Comparison of LDA and CSA Classifiers Using Standard Metrics

| Classifier | Wavelet | ER (%) | SE (%) | SP (%) | ACC (%) | PR (%) | MCC (%) |
|---|---|---|---|---|---|---|---|
| LDA | DB4 | 30 | 65 | 75 | 70 | 72.22 | 40.20 |
| | HAAR | 17.5 | 80 | 85 | 82.5 | 84.21 | 65.08 |
| | BIOR4.4 | 22.5 | 85 | 70 | 77.5 | 73.91 | 55.63 |
| | SYM8 | 15 | 80 | 90 | 85 | 88.89 | 70.35 |
| | DMEY | 32.5 | 75 | 60 | 67.5 | 65.22 | 35.40 |
| CSA | DB4 | 27.5 | 70 | 75 | 72.5 | 73.68 | 45.06 |
| | HAAR | 2.5 | 100 | 95 | 97.5 | 95.24 | 95.12 |
| | BIOR4.4 | 22.5 | 90 | 65 | 77.5 | 72 | 56.80 |
| | SYM8 | 12.5 | 90 | 85 | 87.5 | 85.71 | 75.09 |
| | DMEY | 17.5 | 85 | 80 | 82.5 | 80.95 | 65.08 |

ER, Error Rate; SE, Sensitivity; SP, Specificity; ACC, Accuracy; PR, Precision; MCC, Matthews Correlation Coefficient; LDA, Linear Discriminant Analysis; CSA, Cuckoo-Search Algorithm; DB4, Daubechies; HAAR, Haar; BIOR4.4, BiorSplines; SYM8, Symlet8; DMEY, DMeyer

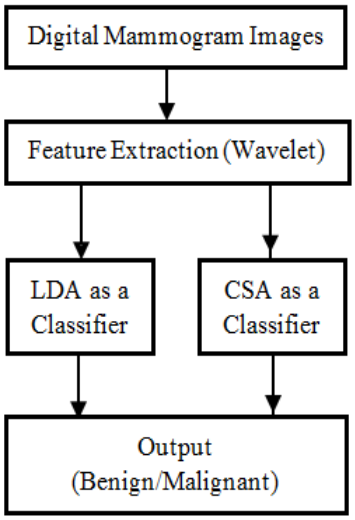


Figure 1. Proposed Method ( LDA, Linear Discriminant Analysis; CSA, Cuckoo-Search Algorithm )

Table 3. Comparison of Proposed Model with Related Works

| References | Methodology | Classification Accuracy (%) |
|---|---|---|
| Srivastava et al., (2014) | Hybrid type of features with k-nearest neighbour Classifier | 87 |
| Saini et al., (2015) | Texture features with artificial neural network. | 87.5 |
| Pawar et al., (2016) | Wavelet coefficient features with genetic fuzzy system. | 89.47 |
| Gardezi et al., (2016) | Curvelet based grey level co-occurrence matrix and geometric invariant shift transform with support vector machine classifier. | 92.39 |
| Vaidehi et al., (2017) | Texture features with sparse representation classifier. | 93.75 |
| Harefa et al., (2017) | Texture features with support vector machine classifier. | 93.88 |
| Pratiwi et al., (2015) | Texture features with radial basis function neural network. | 93.98 |
| Gautam et al., (2018) | Texture features with back propagation neural network. | 96.3 |
| Proposed work | Statistical features with cuckoo-search algorithm | 97.5 |

32.5 and lowest MCC value of 35.40 as shown in Table 2 and Figure 2 when compared with all wavelets used in LDA classifier. The lowest error rate as 15 with the highest value of MCC as 70.35 for LDA classifier is obtained for the sym8 wavelet and so it misclassifies only 8 malignant mammogram images as benign and only 4 benign mammogram images as malignant. Similarly for CSA classifier, the highest misclassification or error rate with lowest MCC value is found for dB4 wavelet as in Table 2. However the haar wavelet with CSA classifier gives the lowest error rate of 2.5 and the highest MCC value of 95.12 as in Figure 2. This is because the CSA classifier using haar wavelet correctly classifies almost

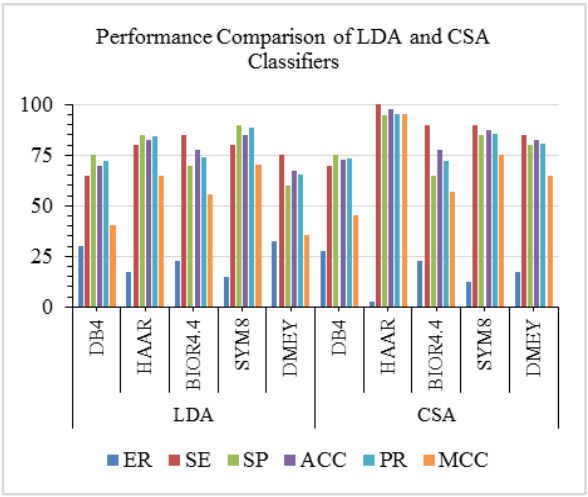


Figure 2. Comparative Analysis of LDA and CSA Classifiers. ER, Error Rate; SE, Sensitivity; SP, Specificity; ACC, Accuracy; PR, Precision; MCC, Matthews Correlation Coefficient; DB4, Daubechies, HAAR, Haar; BIOR4.4, BiorSplines; SYM8, Symlet8; DMEY, DMeyer; LDA, Linear Discriminant Analysis; CSA, Cuckoo-Search Algorithm

all 40 taken malignant mammogram images as malignant and falsely classifies only 2 benign mammogram images as malignant images. This impact will leads to the highest accuracy of 97.5% for the CSA classifier with haar wavelet and the LDA classifier with sym8 wavelet will give the highest accuracy of 85% over other wavelet families. The above discussion noticeably evident that the performance of nature-inspired CSA classifier using haar wavelet outperforms the performance of LDA classifier in classifying the mammogram images.

The comparison of classification accuracy of other related existing works (Srivastava et al., 2014; Gautam et al., 2018) with the proposed model is discussed in Table III. As in the table, the comparison is based on the methodology with different feature extraction approaches using various classification algorithms. The nature-inspired based proposed classifier with haar wavelet gives higher classification accuracy than the other related systems. This is because of the simplicity and proficiency of CSA in tackling highly non-linear problems.

## Discussion

The work intends to propose a system for the classification of mammogram images from MIAS data corpus into either benign class or malignant class. The work utilizes five individual wavelet families for the extraction of statistical features from the input mammograms. And these features are served as input to the LDA and CSA classifiers respectively. The smoothness nature of haar wavelet along with CSA classifier provides an improved accuracy of 12.5% when compared with LDA classifier. The future work of the study is to implement a dynamic class-label classification for the identification of breast cancer with some other datasets.

## Acknowledgements

The authors received no financial support for the research, authorship, and/or publication of this article. There is no conflicts of interest.

## References

Abirami C, Harikumar R, Chakravarthy SS (2016). Performance analysis and detection of micro calcification in digital mammograms using wavelet features. In 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) IEEE, pp 2327-31.

Amin HU, Malik AS, Ahmad RF, et al (2015). Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques. *Australas Phys Eng S*, **38**, 139-49.

Cui Z, Sun B, Wang G, Xue Y, Chen J (2017). A novel oriented cuckoo search algorithm to improve DV-Hop performance for cyber–physical systems. *J Parallel Distr Com*, **103**, 42-52.

DeSantis C, Ma J, Bryan L, Jemal A (2014). Breast cancer statistics, 2013. *CA Cancer J Clin*, **64**, 52-62.

Documet P, Bear TM, Flatt JD, et al (2015). The association of social support and education with breast and cervical cancer screening. *Health Edu Behav*, **42**, 55-64.

Falk D, Cubbin C, Jones B, et al (2018). Increasing breast and cervical cancer screening in rural and border Texas with friend to friend plus patient navigation. *J Cancer Edu*, **33**, 798-05.

Gandomi AH, Yang XS, Alavi AH (2013). Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Eng Comput*, **29**, 17-35.

Gardezi SJS, Faye I, Adjed F, Kamel N, Eltoukhy MM (2016). Mammogram classification using curvelet GLCM texture features and GIST features. In International Conference on Advanced Intelligent Systems and Informatics Springer Cham, pp 705-13.

Gautam A, Bhateja V, Tiwari A, Satapathy SC (2018). An improved mammogram classification approach using back propagation neural network. In Data Engineering and Intelligent Computing Springer, Singapore, pp 369-76.

Harefa J, Alexander A, Pratiwi M (2017). Comparison classifier: support vector machine (SVM) and K-nearest neighbor (K-NN) in digital mammogram images. *Jurnal Informatika dan Sistem Informasi*, **2**, 35-40.

Henriksen EL, Carlsen JF, Vejborg IM, Nielsen MB, Lauridsen CA (2019). The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. *Acta Radiol*, **60**, 13-8.

Kan M, Shan S, Zhang H, Lao S, Chen X (2015). Multi-view discriminant analysis. *IEEE Trans Pattern Anal Mach Intell*, **38**, 188-94.

Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR (1999). Fisher discriminant analysis with kernels. In Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop IEEE, pp 41-8.

Pawar MM, Talbar SN (2016). Genetic fuzzy system (GFS) based wavelet co-occurrence feature selection in mammogram classification for breast cancer diagnosis. *Perspect Sci*, **8**, 247-50.

Peairs KS, Choi Y, Stewart RW, Sateia HF (2017). Screening for breast cancer. In Seminars in oncology WB Saunders, 44, pp 60-72.

Powers DM (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol*, **2**, 37-63.

Pratiwi M, Harefa J, Nanda S (2015). Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. *Procedia Comput Sci*, **59**, 83-91.

Saini S, Vijay R (2015). Mammogram analysis using feed-forward back propagation and cascade-forward back propagation artificial neural network. In 2015 Fifth International Conference on Communication Systems and Network Technologies IEEE, pp 1177-80.

Srivastava Z, Sharma N, Singh SK, Srivastava R (2014). Quantitative analysis of a general framework of a CAD tool for breast cancer detection from mammograms. *J Med Imaging Health Inform*, **4**, 654-74.

Sundaram M, Ramar K, Arumugam N, Prabin G (2011). Histogram modified local contrast enhancement for mammogram images. *Appl Soft Comput*, **11**, 5809-16.

Tharwat A, Gaber T, Ibrahim A, Hassanien AE (2017). Linear discriminant analysis: A detailed tutorial. *AI Commun*, **30**, 169-90.

Vaidehi K, Subashini TS (2015). Automatic characterization of benign and malignant masses in mammography. *Procedia Comput Sci*, **46**, 1762-9.

Yang XS, Deb S (2010). Engineering optimisation by cuckoo search. arXiv preprint arXiv:1005.2908.

Zhu W, Zeng N, Wang N (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis

with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland, **19**, pp 67.