A hybrid artificial neural network classifier based on feature selection using binary dragonfly optimization for breast cancer detection

S Parvathavarthini\textsuperscript{1*}, D Deepa\textsuperscript{2}
\textsuperscript{1}Assistant Professor (Sr. Grade), School of Computer Technology and Applications, Kongu Engineering College, Erode, Tamilnadu, INDIA
\textsuperscript{2}Data Scientist, Rubixe Technologies private Limited, Karnataka

Abstract. Medical image analysis has become a challenging task as it contributes to disease diagnosis. Breast cancer has been the prominent reason for death among women. While analysing mammogram images, there is a need for clear differentiation of between benign and malignant tissues. Also, early detection of breast masses lead to prediction of breast cancer at the initial stage and minimizes risk of death. In this work, the image is preprocessed using Median filter and is segmented using Fuzzy C Means clustering. Fuzzy C-Means clustering algorithm helps in extracting the region of interest by allocating pixels with similar characteristics into a single group. A pixel may be present in various clusters with different membership values. The belongingness of a pixel to a cluster is decided by the highest membership value. Then the statistical, texture and shape features are extracted from the image. Since there may be many features that are less relevant for classification process, prominent features are selected with the help of Binary Dragonfly Optimization Algorithm and the selected features are fed into a Feed Forward Neural Network trained with Back Propagation Learning to classify the mass as benign or malignant. Experiments are conducted over 320 images from mini-MIAS database out of which 200 ROIs are used in training and 120 ROIs are used in testing phase. The region of interest from given mammogram images are extracted successfully and classified with an accuracy of 98.75%.

Keywords: Binary dragonfly optimization based feature selection, breast cancer detection, feed forward neural network, first order statistics, fuzzy C means clustering, texture and shape features.

1. Introduction
Breast cancer is one of the dreadful diseases affecting women. But the consoling fact is that it can be easily cured if it is detected at its earlier stage. However, the expertise of the medical professionals needs to be assisted by proper Computer Aided Detection Systems (CAD). By utilizing the trending techniques like machine learning, diagnosis can be more accurate and effective. Digital image processing is employed to process the digital image and extract some useful information based on its pixel values. This method proposes an efficient approach to detect initial stage masses and helps in alleviating the risks associated with breast cancer.
Breast cancer detection through CAD systems has grabbed the attention of many researchers. [1] detected malignant tissues with higher intensity levels that differ them from the background. An averaging filter and thresholding operations are performed and Max-mean and Least-variance techniques are applied. Finally morphological closing operation and image gradient techniques are utilized to find the region boundary.

[2] devised a CAD system with the ability to detect malignancy from histopathological images using Radial Basis Function Networks (RBFN). [3] proved that texture features are prominent in identifying the variations between benign and malignant lesions. Radio Local Ternary Pattern (RTLP) is used to find the direction of edge patterns in the Region Of Interest (ROI) with respect to the center of masses. Artificial Neural Network (ANN), SVM and Random forest classifiers are used and features are extracted using RTLP, LTP, Gray Level Co-occurrence Matrix (GLCM) and wavelet. The combination of RTLP features fed into the three classifiers yielded higher results that the other feature extraction methods.

[4] constructed a fuzzy rule based system and a genetic algorithm based learning process learns the knowledge from the rule base. Wavelet Co-occurrence Feature (WCF) is used for feature selection from four levels of decomposition to choose the most optimal features to be fed into genetic fuzzy system for classification. The system produced 89.47% accuracy.

[5] investigated on the use of Adaptive Artificial Immune Networks (A2INET) for diagnosing bilateral asymmetry in mammogram images which is highly correlated with the risk of breast cancer. The accuracy of the image reached upto 90%. [6] proposed a self-adaptive dragonfly optimization based thresholding method for segmenting digital images. The results are evaluated using Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSI) and Standard Deviation Index (SDI) and found to be highly impressing. [7] utilized continuous wavelet transform to extract the mammogram image features which are given to four different classifiers like Support Vector Machines (SVM), K Nearest Neighbor, NaiveBayes and Fast correlation based filter. It is used to evaluate how dimensionality reduction helps for better classification. [8] proposed an optimized region growing technique for segmenting breast masses. The threshold value for region growing algorithm is fixed with the Dragonfly optimization algorithm. Texture features obtained from segmented image are given as input to a feed forward neural network for classification as normal or cancerous.

[9] reviewed the deep learning methods available in the literature for breast cancer detection from mammogram images. [10] extracted ROI using intuitionistic fuzzy clustering with neighbourhood attraction from mammogram images and evaluated the segmentation accuracy using Jaccard and Dice indices. [11] utilized conventional neural networks for efficient classification of lesions in mammogram images. Since gathering a large sample of medical images for training the CNN is tedious, transfer learning of non-medical images is applied and three different networks pre-trained with Imagenet dataset are used. The features extracted from CNN are fed into SVM for further classification and also some handcrafted features are introduced to get effective classification.

[12] proposed an efficient technique for retrieval of images from databases. The images are divided into rings. Features are extracted from these multiple links and ranked to retrieve relevant images from database. [13] proved that Binary Dragonfly Algorithm (BDA) is efficient for feature selection and provides exhaustive search within the search space leading to the identification of useful features of classification. The experiments are conducted over 18 datasets from UCI repository and results are compared with Particle Swarm Optimization and Genetic Algorithm to show its efficacy.

A discrete wavelet transform based on multiscale surrounding region detection is performed by [14] to detect breast cancer. [15] classified mammographic masses by extracting the ROI using Intuitionistic Fuzzy C-Means and decision tree method.

However, there are still numerous research gaps to be addressed. Real time medical images are hard to obtain and only less samples could be collected. Therefore, suitable methods that assist experts in diagnosing the disease with a higher accuracy need to be devised.
2. Materials and methods

The proposed methodology contains various steps. To start with, the images are collected from MIAS database. Then the images are preprocessed using Median filter. The noise removed image is segmented using Fuzzy C Means (FCM) algorithm. The region of interest is extracted and in order to analyse it, several texture, shape and color features are extracted from the image. These features extracted are numerous and selection of the prominent features is done using Binary Dragonfly Algorithm. The selected features are fed into a feed forward neural network using back propagation learning. These steps are explained in detail in the following sections and are represented in figure 1.

![Figure 1. Workflow of Proposed work.](image)

2.1. Data acquisition and preprocessing

While developing new algorithms, researchers have to test the performance against a standard test database. In this work the images are taken from the standard mammogram image repository named Mammographic Image Analysis Society (MIAS). The mini-MIAS database contains 322 images of normal, benign and malignant categories.

Preprocessing is necessary to preserve the essential intensity information naturally present in the image while removing unwanted noise from it. The image is resized and unwanted distortions are suppressed to get a clear image. The background information and pectoral muscles are to be removed while preprocessing the image. The key idea behind using median filter is that it is capable of preserving the edges while removing noise. It is more robust and the median value is one of the pixel values in the neighborhood. Thus it will not create any new pixel value that is not in the original image. Also, there is no boundary shift and no reduction in contrast levels of the image.

2.2. Segmentation using FCM

There is a need for segregating the data into meaningful non-overlapping parts and to extract the specific Region of Interest (ROI) from the vast set of images. Image segmentation is carried out to do this. It can be done using various techniques like feature thresholding, template matching, edge detection, histogram based methods, compression based methods, clustering and region based techniques. In this work, FCM is used for segmenting the image and extracting the ROI. FCM partitions the image into meaningful parts each with similar traits or properties. Each of the pixels in a segmented region appears similar with respect to some characteristics like color, intensity or texture.
Fuzzy C Means is a popular soft clustering algorithm capable of allocating objects to more than one cluster. Based on the membership values, the belongingness of the object to the cluster is decided. The object goes to the cluster with the highest degree of membership. The FCM algorithm is iterative and produces ‘C’ optimal partitions of the image by minimizing the weighted distance within group sum of squared error. This will serve as the objective function of FCM.

2.3. Feature extraction
Image feature gives much more detailed information about the image under observation. In order to process the image, information like statistical, shape and texture features are extracted. The first order statistics like mean, standard deviation, variance, kurtosis and skewness are extracted. Then, the second order statistics such as GLCM is used to identify the occurrence of pair of pixels at specific locations in the image. GLCM is constructed at various angles like 0, 45, 90 and 135. The texture features extracted involve contrast, energy, entropy, correlation, maximum probability, dissimilarity, homogeneity, inverse different moment and sum average. Nine features in four different angles constitute a total of 36 texture features. Shape features such as area, centroid, bounding box, solidity, extent, regional area, compactness, eccentricity and Euler number are extracted. Based on these 50 features, the information about the specific ROI is gathered. Extracting large number of features from images is not directly proportional to classification accuracy. Some irrelevant or noisy information may be present in the features. Apart from that, there might be some features prominent to the process of classification. They are selected using BDA.

2.4. Feature selection using binary dragonfly optimization algorithm and classification by FFANN
Feature selection results in either selection or rejection of a feature. It improves the performance of classification algorithm by removing redundant or irrelevant features. As it contains two states, BDA is very much suitable for feature selection. A selected feature is indicated by 1 whereas the other non-selected features are marked with zero. Since 50 features are considered in this work, the solution vector will be of size 50. Dragonfly algorithm is designed by mimicking the swarm behaviour of dragonflies. The static swarming behavior contributes to diversification while the dynamic swarming behaviour contributes to intensification. The five parameters involved in DA are separation, alignment, cohesion, attraction and distraction.

Separation indicates how the search agents separate themselves from each other. It is calculated as

\[ S_i = -\sum_{j=1}^{N} X - X_j \]  

(1)

Alignment is the process of adjusting the velocity with respect to other search agents. It is given by

\[ A_i = \frac{\sum_{j=1}^{N} V_j}{N} \]  

(2)

Cohesion measures the inclination of members towards the centre of the swarm and is denoted as

\[ C_i = \frac{\sum_{j=1}^{N} X_j}{N} - X \]  

(3)

Attraction is the tendency of moving towards food source and is given by

\[ F_i = F_{pos} - X \]  

(4)

where \( F_{pos} \) is the position of food source

Distraction is the intention of keeping distance from enemies which is calculated as

\[ E_i = E_{pos} + X \]  

(5)

where \( E_{pos} \) is the position of the enemy.

The position of dragonflies is updated using step vector and position vector. Step vector is calculated as

\[ \Delta X_{it+1} = (sS_{it} + aA_{it} + cC_{it} + fF_{it} + eE_{it}) + wX_{it} \]  

(6)
Position vector is calculated as

$$X_{it+1} = X_{it} + \Delta X_{it+1}$$  \hspace{1cm} (7)

where the variable $it$ denotes the iteration.

The DA algorithm produces continuous solutions whereas BDA produces binary solutions. The transfer function to convert continuous solutions to binary value generates a probability of changing a position’s element to zero or one which is impacted by the value of step vector.

$$T\left(v^d \right) = \frac{|v^d|}{\sqrt{1+(v^d)^2}}$$ \hspace{1cm} (8)

This transfer function is used to convert the $i$th element of position vector to zero or one as follows

$$X(it + 1) = \begin{cases} X_{it} & r < v^d(it) \\ X_{it} & r \geq v^d(it) \end{cases}$$ \hspace{1cm} (9)

where $R$ is a random number in the $[0, 1]$ interval.

While performing feature selection certain trade-off is to be considered between classification accuracy and reduction rate. So the fitness function is derived such that a balance between these two factors is considered.

$$J(X) = \alpha \gamma(X) + \beta \left(1 - \frac{|S|}{|T|}\right)$$ \hspace{1cm} (10)

In this equation, $\alpha$ ranges from 0 to 1, $\gamma(X)$ denotes classification accuracy while using a subset of features $X$, $\beta$ is calculated as $1 - \alpha$, $S$ is the number of features selected and $T$ is the total number of features considered for evaluation.

2.5. Pseudocode of BDA-based feature selection algorithm

Initialize the extracted features $X(i = 1, 2, \ldots, n)$

Initialize $\Delta X(i = 1, 2, \ldots, n)$

While (stop condition is not reached) do

Evaluate all dragonflies based on equation (10)

Update food source and enemy

Update the parameters $w, s, a, c, f$ and $e$

Calculate $S, A, C, F$ and $E$ based on equations (1 to 5)

Update step vectors $(\Delta X_{it+1})$ by equation (6)

Calculate transfer function $T(\Delta X)$ using equation (8)

Update $X_{it+1}$ by equation (9)

End while

Return the optimal set of selected features

The steps in BDA based feature selection and classification are narrated in figure 2. BDA algorithm returns a vector whose size is exactly equal to the number of features extracted from the image. The value for each feature that is selected will be one and the features not selected will have a value of zero. Eventhough some new technique like Convolutional Neural Networks (CNN) has arrived, ANN has its own advantage that even with limited number of samples, it works well. In contrast, CNN needs a lot of samples and in the case of medical data it is very tough to collect more samples and it is a wearisome task to train CNN.
The feature selection based on optimization proceeds such that the steps are carried out combining the Binary dragonfly optimization and Artificial Neural Network (ANN) classifier. ANN is a computational model based on the structure and function of biological neural networks. It helps to model the complex relationships between inputs and outputs. It considers the data samples rather than entire datasets to find patterns or solutions that exist in data.

Figure 2. Design of Binary Dragonfly optimization based Feature Selection and Classification.
This work utilizes a Feed Forward Back Propagation Network which uses 200 images as training set. The network has 3 layers- input, hidden and output layers. A total of 50 features are extracted. So, the number of nodes in input layer is equivalent to the number of features extracted. This is a two-class classification problem. So output layer has a single node indicating 0 for normal and 1 for malignant classes. To start with, the weights are initialized randomly. Then the output is calculated and the error is found as the difference between actual output and target output. Different set of input features are taken as dragonflies. The inputs are multiplied with weights and bias and the activation function used here is given by

\[ A_{V_b} = \sum W_{a,b} y_a + B_b \]  

Where \( W_{a,b} \) is the weight between \( a^{th} \) and \( b^{th} \) neuron and \( B_b \) is the bias of \( b^{th} \) neuron.

The type of ANN used in this work is a Multilayer Perceptron which is applied with Backpropagation learning in which the error is propagated back to the previous nodes so that adjustments on weight can be made and the error between actual and target output is reduced. The activation function used here is a summation of input weights and bias. The input weights are given in the range -1 to +1. Figure 3 shows the structure of the ANN used in this work. Mean Squared Deviation is used to compute the error at the output layer and is given by

\[ \text{MSDev} = \frac{1}{p} \sum_{i=1}^{p} (\text{Tout}_i - \text{Aout}_i) \]  

Where \( \text{Tout} \) is the target output, \( \text{Aout} \) is the actual output and \( p \) is the number of predictions. Based on the error value, the weights at the previous layers are updated until there is a significant reduction in the error. Once the system is trained, test data can be applied.

2.6. Results and discussion

The proposed algorithm is implemented in Matlab. The result of the proposed system is evaluated using various measures that identify the effectiveness of both segmentation and classification.

2.6.1. Evaluation metrics

Dice index

Dice index [17] measures the existence of spatial overlap between two binary images. If there is a higher value, it indicates the better agreement between pixel intensities of the two images.

\[ D(A, B) = \frac{2|A \cap B|}{|A| + |B|} \]  

Jaccard index

Jaccard Index [18] measures how far there is a coincidence between the ground truth images and the segmented images and also the deviation is measured. It is calculated as

\[ J(A, B) = \frac{|A \cap B|}{|A \cup B|} \]
Sensitivity
It is the ability to identify the true positive that is the persons who are affected by the disease really and is identified the model also.

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad (15)
\]

Specificity
It is the ability to identify the true negative that is the persons who are categorized as not to have the disease both by the expert and the proposed model.

\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (16)
\]

Accuracy
Accuracy [19] is a measure of how perfect the classification matches the actual class label

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)
\]

Precision
Precision determines how valid the classification result is. It is calculated as the ratio of true positives to the total number of positives detected. The result lies between 0 and 1. Zero indicates no precision and 1 indicates perfect precision.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (18)
\]

F-score
F-Score maintains a balance between precision and recall. It is denoted as the harmonic mean of precision and recall.

\[
\text{F-score} = \frac{2TP}{2TP + FP + FN} \quad (19)
\]

Mathews correlation coefficient
Mathews Correlation Coefficient considers true class and predicted class as two variables and computes the correlation coefficient between these binary variables. If there is a higher correlation between these two, then the classification is said to be successful. The value ranges from -1 to +1.

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (20)
\]

Kappa statistics
Kappa statistics measures the interrater reliability. It compares the score given by the model and the actual class to the data. It is given by

\[
\text{kappa} = \frac{(\text{observed agreement} - \text{expected agreement})}{1 - \text{expected agreement}} \quad (21)
\]

Figure 4 shows the results of ROI extracted from the image. It is evident that the region boundaries are clearly shown. The input image is noise removed and median filtered during preprocessing. The preprocessed image is segmented using FCM and the ROI is extracted. The step by step changes that the images undergo are shown in the figure 4.

The accuracy of the classification algorithm greatly depends on the proper segmentation of ROI. That’s why segmentation accuracy is measured with the help of Dice and Jaccard index. Also, MIAS provides ground truth data for the images. Thus, the location of abnormality specified in MAS is compared to that of the segmentation results. FCM enables the representation of finer details from the image and thus yielded effective segmentation of ROI from the surrounding unwanted regions.

The total number of dragonfly population is 50 and the algorithm is run for 100 iterations. The values for \(\alpha\) and \(\beta\) are taken as 0.99 and 0.01 respectively. Out of the five histogram features, 4 have been selected, 21 out of 36 GLCM features are selected and 6 out of 9 shape features are selected. So, a total of 31 optimal features are selected by BDA. The results are shown in table 1 and are compared with the results produced by [8] and [16].

Table 1 infers that it is not necessary that when the number of features extracted is high, it directly contributes to higher classification accuracy. Thus, the role of feature selection algorithms gains importance in selecting the relevant and necessary features for classification. Region growing techniques are always time consuming and there is an inability to distinguish the shading of the real
images. Even though wrapper methods like BDA consume more resources, the results tend to be more promising than those worked out without feature selection. CART consumes too much of time and space when the number of decision variables increase.

| Preprocessing | Segmentation | Feature Extraction | Classifications | Dataset | Accuracy | Sensitivity | Specificity | Jaccard Index |
|---------------|--------------|-------------------|-----------------|---------|----------|------------|------------|--------------|
| Punitha et al. | Sheba et al. | Proposed work     |                 |         |          |            |            |              |
| Gaussian filtering | Median filtering | Median filtering | Multithresholding based on Otsu’s method | Fuzzy C-Means |
| Dragonfly based region growing | Multithresholding based on Otsu’s method | First order statistics(5 features, GLCM(9 features in 4 angles), Shape features (9 features) |            |              |
| GLCM(10 features in 4 angles), GLRLM (5 features) | First order statistics(6 features, GLCM(13 features in 4 angles), GLRLM (11 features in 4 angles), Shape features (15 features) | First order statistics(5 features, GLCM(9 features in 4 angles), Shape features (9 features) |            |              |
| 45 | 117 | 50 |          |              |          |            |            |              |
| -- | CART | Binary Dragon fly based feature selection | Feed Forward ANN with Back Propagation MIAS | MIAS | 98% | 96% | 98.8% | 98.75% |
| MIAS | 98% | 96% | 98.8% | 98.75% | 98.4% | 90% | -- | 95.6% |

Table 1. Comparative study with state-of-the-art methods.

Figure 4. Region of Interest Extraction.
Dice index and Jaccard index are mainly computed to evaluate the performance of segmentation. Since FCM is capable of representing minute details about the data, it is effective in segmenting the image. This could be proved by the values of Dice and Jaccard indices. Jaccard index reached up to 95.6% and Dice index showed 98.2%. Mathew correlation coefficient measured 0.9562 and kappa statistics showed 0.9784. When it comes to binary classification problem, Mathew correlation coefficient is an effective measure as it considers all four quadrants of confusion matrix while other measures like sensitivity and specificity consider only three quadrants of confusion matrix. Figure 5 shows the comparison of Accuracy, Sensitivity and Specificity of various algorithms taken for comparison and it is found that the proposed method outperforms all other methods in all the three aspects.

![Figure 5. Comparison of Accuracy, Sensitivity and Specificity.](image)

2.7. CONCLUSION

The proposed work serves as a guide to radiologists in case of critical situations. Pre-processing using median filters and removing noise facilitates the clear segmentation of images. Since Fuzzy C-Means is an effective and fast converging technique for segmentation, the ROI is extracted such that the boundaries of mass are clearly separated from other normal regions. The performance of segmentation is found quite impressive when evaluated with Dice and Jaccard indices. Finally, the looping of feature selection and classification continues until the optimal set of features leading to minimum error of the neural network is found. Extraction of texture and shape features contributes to increase in classification accuracy which is further increased by utilizing the optimal features selected by BDA. Out of the 320 images considered for evaluation, 316 images are correctly classified leading to an accuracy of 98.75%. The proposed method performs well in all aspects.

In future, several feature selection algorithms can be employed and the best method can be evaluated. The parameters used in BDA and neural network can be tuned using some soft computing methods.

References
[1] Singh AK and Gupta B 2015 A novel approach for breast cancer detection and segmentation in a mammogram Procedia comput. sci. 54 676-82.
[2] Kanojia MG and Abraham S 2016 Breast cancer detection using RBF neural network 2nd International Conference on Contemporary Computing and Informatics (IC3I) 363-68.
[3] Král P and Lenc L 2016 LBP features for breast cancer detection *IEEE International Conference on Image Processing (ICIP)* 2643-47.

[4] Pawar MM and 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.

[5] Magna G, Casti P, Jayaraman SV, Salmeri M, Mencattini A, Martinelli E and Di Natale C 2016 Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system *Knowl-based syst.* 101 60-70.

[6] Sambandam RK and Jayaraman S 2018 Self-adaptive dragonfly based optimal thresholding for multilevel segmentation of digital images. *J. King Saud Univ., Comp. & Info. Sci.* 30 449-61.

[7] Ebrahimpour MK, Mirvaziri H and Sattari-Naeini V 2018 Improving breast cancer classification by dimensional reduction on mammograms *Comput. Methods. Biomech. Biomed Eng. Imaging Vis.* 6 618-28.

[8] Punitha S, Amuthan A and Joseph KS 2018 Benign and malignant breast cancer segmentation using optimized region growing technique. *Future computing inform. j.* 3 348-58.

[9] Gardezi SJ, Awais M, Faye I and Meriaudeau F 2017 Mammogram classification using deep learning features. *IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* 485-88.

[10] Parvathavarthini S, Karthikeyani Visalakshi N and Shanthi S 2019 Breast cancer detection using crow search optimization based Intuitionistic fuzzy clustering with neighborhood attraction. *Asian pac. j cancer.* p. 20 157-65.

[11] Perre AC, Alexandre LA and Freire LC 2018 Lesion classification in mammograms using convolutional neural networks and transfer learning. *Comput. Methods. Biomech Biomed Eng. Imaging Vis.*

[12] Rayen SJ and Subhashini R An efficient mammographic image retrieval using ring-based classification. *Int. J. Emerg. Technol.,* 10 403-07

[13] Mafarja M, Alijrah I, Heidari AA, Faris H, Fournier-Viger P, Li X and Mirjalili S 2018 Binary dragonfly optimization for feature selection using time-varying transfer functions. *Knowl-based syst.* 161:185-204.

[14] Shanthi S and Bhaskaran VM 2013 A novel approach for detecting and classifying breast cancer in mammogram images. *Int. J. Intell. Info. Technol.* 9 21-39.

[15] Shanthi S and Bhaskaran VM 2011 Intuitionistic fuzzy C-means and decision tree approach for breast cancer detection and classification *Eur. J. Sci. Res.* 66 345-51.

[16] Sheba KU and Gladston Raj S 2018 An approach for automatic lesion detection in mammograms Cogent Eng. 5 1444320.

[17] Dice LR 1945 Measures of the amount of ecologic association between species *Ecology* 26 297-02.

[18] Cheetham AH and Hazel JE 1969 Binary (presenceabsence) similarity coefficients. *J Paleontol* 43 1130-36.

[19] Metz CE 1978 Basic principles of ROC analysis *Semin Nucl Med* 8 283-98.