ISSN: 1813 – 1662 (Print)
E-ISSN: 2415 – 1726 (On Line)

Feature Selection based on Genetic Algorithm for Classification of Mammogram Using K-means, k-NN and Euclidean Distance

Kameran Adil Ibrahim
College of Education Tuzkhurmatu , Tikrit University , Tikrit , Iraq

Abstract

There have been several supervised classification attempts for mammograms in the recent times, but very few research works have focused on unsupervised classification to explore its potentialities and weaknesses. I have in this paper attempted to utilize unsupervised clusters to classify malignant, and benign mammograms samples. MiniMIAS database has total 322 mammogram images out which 64 are benign and 51 are malignant. I used 115 images for my experimentation i.e. 64 benign and 51 malignant. Out of these 115, 60% were used for training and 40% for testing. Therefore from 64 benign cases 39 images were used for training and rest for testing, and out of 51 malignant cases 31 images were used for training and rest for testing., the classifications was done on the bases of the features selected using genetic algorithm. Attempts have also been made to study the performance of each feature selected by Genetic Algorithm (GA) in classification. The initially identified clusters using K-means are used to classify 60 unknown samples using k-NN. The proposed work got reasonably good results with 96.23% accuracy for malignant samples, 95.37% for benign. The proposed work can help the radiologists and oncologist as second opinion during screening sessions for early detection.

Keywords: K-means Clustering, Unsupervised Classification, K-NN, Euclidean Distance, Genetic Algorithm .

Introduction

Amongst all forms of known malignant cancers, breast cancer is the one concerning women globally since it is one of the most prominent source of cancer related female deaths [1]. There is significant drop of mortality rate in economically rich countries [2] whereas there is still a growing life treat of the breast carcinoma in economically average and poor countries especially in Eastern Mediterranean Region (EMR) [3]. EMR is grouped having twenty one member nations from Middle East, central Asia and North America, this grouping is done by World Health Organization (WHO). Iraq is one of the Middle Eastern nations included in EMR. International Agency for Research on Cancer (IARC) indicated that just in the year 2012, there were 292,677 newly diagnosed cases in EMR countries [4]. Since breast cancer is a very common malignancy found amongst females of all nations of EMR, therefore within the next one and half decade, WHO indicates that largest rise in the cancer cases is expected to be in the EMR region [5]. To tackle these indications from WHO, in the year 2009 a “National Breast Cancer Research Program-NBCRP” was established in Iraq in association with International Agency for Research on Cancer (IARC) and WHO [5].

In medical image processing texture analysis plays a significant role [6] mainly in three ways i.e. either for feature extraction or pattern detection or segmentation based on the contents of the image. For texture analysis of any image there are three possible approaches one being statistical, while structural and spectral being the other two. In this work I have utilized statistical approach for studying the texture of mammograms and obtained four statistical features and twenty-two texture features totaling twenty-six features. Amongst these twenty six features which are potentially strong to give better classification results are not known. Therefore I have used genetic algorithm to select features that result in better classification of malignant and benign tissues. The features suggested by genetic algorithm were tested in unsupervised classification using k-means clustering algorithm. The clusters obtained by k-means were utilized for classification of unknown samples using k-NN classifier.

Genetic Algorithm (GA)

Genetic algorithm (GA) is popular due its capability of proficiently searching large data sets [7]. GA is considered to be an excellent choice for texture classification due to relative insensitivity towards noisy data [8]. It differs significantly from the other existing wrapper algorithms owing to these reasons:

1. Traditional methods search from a single population point unlike GA which searches from parallel population points [9]. Hence GA is not trapped in local optimally best solution.
2. Other wrapper algorithms like Statistical Dependency (SD) are deterministic type whereas GA is based on probabilistic approach [10].

Apart from these major differences, there are several other points of consideration for selection of GA for this work. The main reason is that GA has several local optimal solutions to the problem.

Proposed Methodology: The methodology and the logical sequence of the steps are made up of four major steps: Pre-Processing of the mammograms (preprocessing includes Artifact and pectoral Muscle Removal), Features Extraction, Feature selection and Benign/Malignant Classification. All the steps involved in the proposed methodology in depicted in the workflow shown in the Figure 1.

Each block of the workflow shown in Figure 1 represents the steps involved in the proposed methodology. The details are presented in the following sections

1. Database: A UK based organization named Mammographic Image Analysis Society (MIAS) has made freely available a subset of their database with 322 mammogram images called mini MIAS. I randomly selected 20 samples each from Normal, Benign & Malignant cases from minMIAS database for the experiments. The images in the database are 8-bit grayscale images and size of each image is 1024 x 1024. miniMIAS includes both side breast mammograms of 161 subjects carefully selected to present all possible categories of abnormalities found in breast cancer patients. The images are stored in PGM format (Portable Gray MAP) and are accompanied with annotation files describing the case for correlation purpose shown in Table 1.

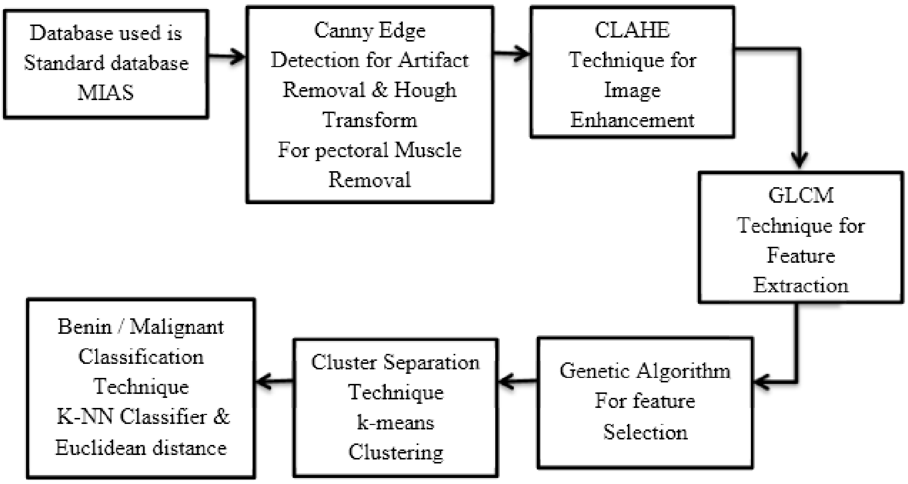

Database used is Standard database MIAS
Canny Edge Detection for Artifact Removal & Hough Transform For pectoral Muscle Removal
CLAHE Technique for Image Enhancement
GLCM Technique for Feature Extraction
Genetic Algorithm For feature Selection
Cluster Separation Technique k-means Clustering
Benin / Malignant Classification Technique K-NN Classifier & Euclidean distance

Figure 1: Experimental Workflow Diagram

Table 1: Cases in Database

Class	Benign	Malignant	Total
Micro calcification	12	13	25
Circumscribed Masses	19	4	23
ill-defined Masses	7	7	14
Speculated Masses	11	8	19
Architectural distortion	9	10	19
Asymmetry lesion	6	9	15
Normal tissue	-	-	**207**
Total	64	51	322

2. Artifact Removal: Raw mammogram images contain wedges and labels, which are markings for the radiologists and oncologist to aid in their diagnostic process. But these wedges and labels become hindrance for the image processing algorithms, therefore procedure of artifact removal from the mammogram before of processing is applied. Canny edge detection binary mask is generated. By using 8-adjecency to connect between two end points of any opened connected component as shown in table. The process starts from generating a mask of size 3 × 3 pixels and placing the center of the mask at one end of the edge. Then any pixel that are 8-adjecency is assumed to be another end and they are connected together to form a closed boundary. The results of artifact removal process can be seen in Table 2.

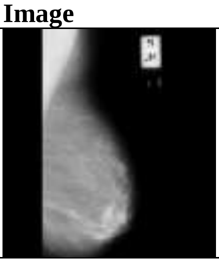
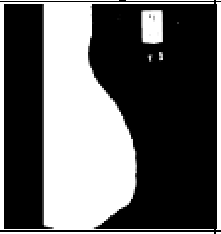
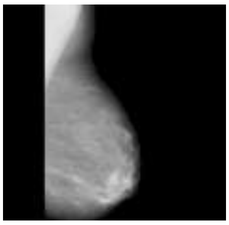
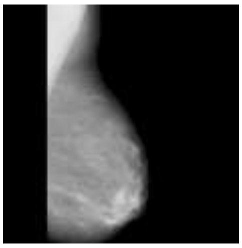
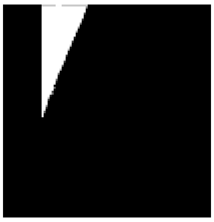
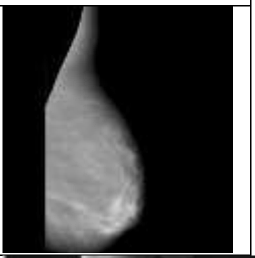
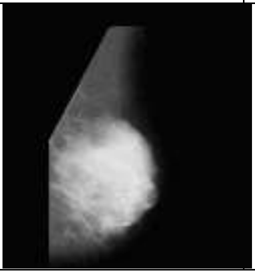
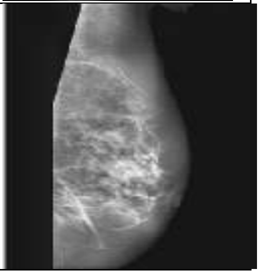
3. Pectoral Muscle Removal: By converting pixels to Hough space information about the line orientation and distance as well as line length are got. Standard Hough transform can be expressed as

$P = x \cos(\theta) + y \sin(\theta)$ ………….(I)

Where ρ indicates the distance between the origin of the line and θ is the angle of inclination of the normal line from the x-axis of almost vertical orientation of the breast in the observed image we can easily search for maximums in Hough space which correspond to vertical lines. θ can have values in range $[-\pi/2, \pi/2]$ and can easily limit our region of interest in Hough space to search for maximum around $\pi/2$. This approach will give us the distance and therefore the position of the longest vertical line which is detected using canny filter from the morphologically opened binary image. The results of pectoral muscle removal process can be seen in Table 2.

4. Image Enhancement: The Contrast Limited Adaptive Histogram Equalization (CLAHE) operates on small regions in the image called tiles rather than the entire image. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches the uniform distribution or Rayleigh distribution or exponential distribution. The neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The results of mammogram image enhancement process can be seen in Table 2.

Table 2: Preprocessing of Mammograms

Technique	Purpose	Original Image	Application of Technique	Result After Applying technique
Canny Edge Detection	Artefact Removal			
Hough Transformation	Pectoral Muscle Removal			
CLAHE	Mammogram Enhancement		No Intermediate Result	

5. Feature Extraction: Feature extraction is a method of capturing visual content of an image. Gray-Level Co-occurrence Matrix (GLCM) and statistical features were extracted separately from all mammograms images. GLCM considers the relation between two neighboring pixels in one offset, as the second order texture, where the first pixel is called reference and the second one the neighbor pixel. GLCM is the two dimensional matrix of joint probabilities $P_{d\theta}(i, j)$ between pairs of pixels, separated by a distance d in a given direction θ [4]. Haralick [16] defined 14 statistical features from GLCM for texture classification. In this work twenty two features from second order of GLCM approach [17] namely contrast, correlation, Variance, homogeneity, information measure correlation, difference variance etc.

6. Feature Selection: Usually feature extraction techniques yield number of features which may or may not contribute in proper classification results. These features at times may be redundant or non-informative. Presence of such features significantly degrades the classification performance. Therefore selecting a proper feature subset from the original features set becomes a vital step before any classification attempt. Genetic algorithm (GA) is popular due its capability of proficiently searching large data sets [7]. GA is considered to be an excellent choice for texture classification due to relative insensitivity towards noisy data [8]. It differs significantly from the other existing wrapper algorithms owing to these reasons: Traditional methods search from a single population point unlike GA which searches from parallel population points [9]. Hence GA is not trapped in local optimally best solution. Other wrapper algorithms like statistical dependency (SD) are deterministic type whereas GA is based on probabilistic approach [10]. Apart from these major differences, there are several other points of consideration for selection of GA for this work. The main reason is that GA has several local optimal solutions to the problem.

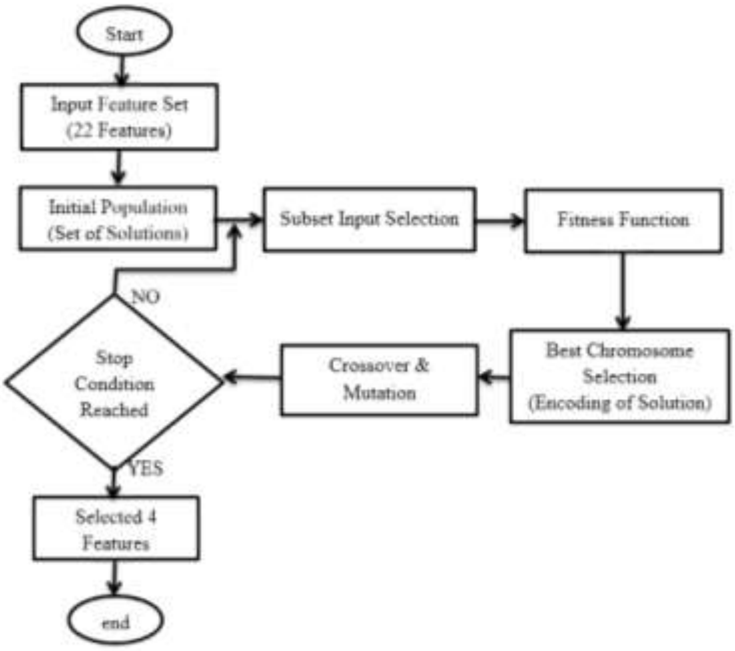


Figure 2: Genetic Algorithm flow Diagram

The flowchart shown in Figure 2 depicts the steps involved in the execution of genetic algorithm. The algorithm takes as input a subset of features usually called as initial population. After completion of every iterations algorithm adds features to its input and the performance of these selected feature subsets is evaluated using a fitness function using the k-NN classifier. Population size was set to 50 and the number of generations was set 100 and the selection of features was set to rank based. At the end of all iterations the GA algorithm selected 4 effective features shown in Table 3 and Figure 3 shows the screenshot of execution of genetic algorithm.

Table 3: Feature Selected GA

Sr. No.	Feature Name
1	Contrast
2	Correlation
3	Difference variance
4	Information measure of correlation2

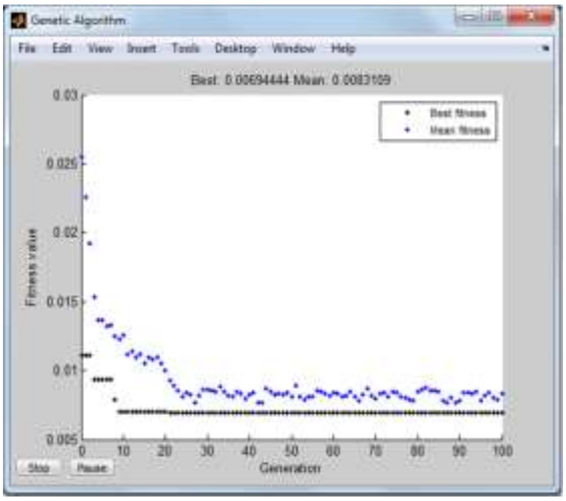


Figure 3: Screenshot of Experimental Output of Genetic Algorithm

7. Cluster Separation: One of the important unsupervised learning approaches is clustering, which intends to identify a set from unlabeled data with intrinsic similarity. Therefore a cluster can be called as group of objects with some similarity amongst themselves which are dissimilar to other objects in a different cluster. The k-means algorithm attempts to minimize the objective function. These functions are the used to decide the aberrations in the patterns from their cluster centroids. The K-means algorithm partitions a dataset into k predefined number of clusters that will try to minimize the intra-cluster distance based on Euclidean distance. The k-mean algorithm is very fast and simple algorithm. In unsupervised machine learning, k-means clustering is a method of cluster analysis which aims to partition ‘n’ observations in to ‘k’ clusters in which each observation belongs to the cluster with the nearest mean. For a given set of observation (x1, x2,...xn) , where each observation is a d-dimensional real vector, then k-means clustering aims to partition the ‘n’ observations in to ‘k’ sets (k<n), {S= S1, S2,...Sn} so as to minimize the within cluster sum of squares (WCSS) in eqn (1).

Arg min s Σ i=1 k Σxi ∈Si || xj – μ i ||2 . . . (1)

Where, μ i is the mean of Si. The number of cluster k is assumed to be fixed in k-means clustering.

8. Classification: - This method of classification, classifies object based on closest training examples in the feature space. We used k-Nearest neighbors algorithm for classification of the selected four features. k-NN is instance-based learning and it is simplest algorithm among all the other learning algorithms. K-NN classifier internally uses Euclidean distance measure for finding the difference between to samples. The concept of distance between two samples or between two variables is fundamental in multivariate analysis. The Euclidean distance between any two given samples can be computed using eqn. 2.

$$d = \sqrt{\sum_{i=1}^{v} (p_{1i} \quad p_{2i})^2} \quad ….. \ (2)$$

9. Results and Discussions: A feature vector containing the predictors i.e. the four top ranked features selected by GA are given as feature set to k-means clustering algorithm [18, 19, and 20]. The algorithm tries to divide the data three classes i.e. normal, malignant and benign as shown in Figure 4. Three centroids are obtained each representing the central point of each cluster. These centroids are used to train a k-NN classifier [21]. This trained classifier is further utilized in classifying the unknown samples and resulting into a randomized set of clusters as shown in Figure 5. The values of centroids are adjusted to arithmetic mean of the cluster it represents. The procedure of classification followed by centroid modification is repeated till the centroids values are not stabilized. The final stabilized centroids values are utilized in classification of the unknown test samples. Out of 322 mammogram images present in MIAS database 64 samples are benign while 51 are malignant and 207 cases are normal. We used first twenty samples of each class, 10 to train the classifier and rest 10 to tested the trained classifier to verify its performance and unknown samples. For benign class 90.38% classification accuracy was obtained while for malignant class 96.23% was obtained and for normal class results were little better than the other two 97.72% as seen in Table 4. The results are obtained by calculating confusion matrix as shown in Figure 5.

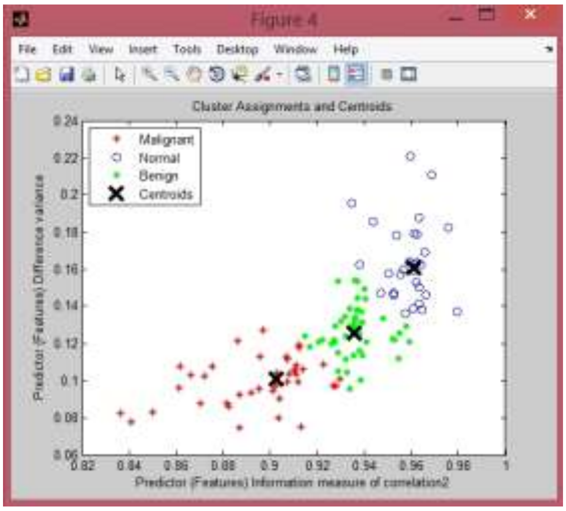


Figure 4: Data divided into three clusters malignant, normal and benign along with centroids of each cluster.

Table 4: Classification Results

Class	Classified	Misclassified
Benign	90.3846	9.6154
Malignant	97.7273	2.2727

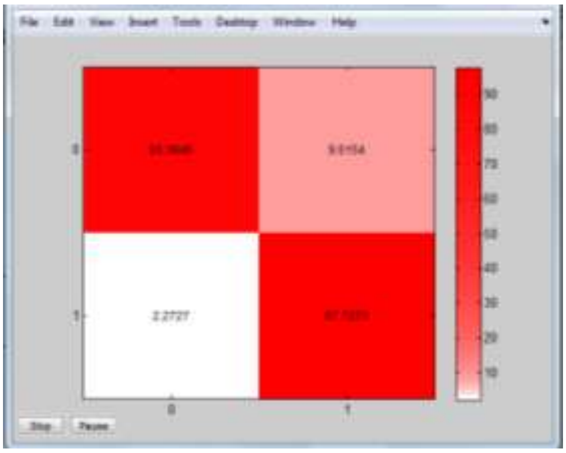


Figure 5: Heat Map of the confusion obtained as Classification Results.

10. Conclusion: The proposed work experimented on 60 images out of 322 mammogram images present in the MIAS database, Mammogram images were preprocessed using the canny edge detection for artefact removal then applied Hough transformation technique to remove to pectoral muscle. For improving the readability of image contents all mammogram images were enhanced using CLAHE technique. I successfully extracted twenty-two texture features using GLCM technique and four features from statistical approach, in all twenty-six features were extracted to be used in classification. Since some of the features were redundant and non-informative, genetic algorithm was applied to select potentially strong features that would help in better classification performance. The four selected features by GA were fed as input to k-means clustering algorithm and three clusters were obtained representing normal, benign and malignant groups along with centroids for each cluster. The centroids of these clusters were further used to classify unknown samples from two groups i.e. benign and malignant using k-NN classifier. The proposed work got reasonably good results with 97.72% accuracy for malignant samples, while 90.38% for benign samples. The proposed work can help the radiologists and oncologist as second opinion during screening sessions for early detection. There is always a room for improvisation of accuracy either by adding more cases or testing with other classifiers.

References

(1) Tavassoli, F. Devilee, P. *Pathology And Genetics Of Tumours Of The Breast And Female Genital Organs*; 1st ed.; IAPS Press: Lyon, 2003.

(2) Jemal, A.; Center, M.; DeSantis, C.; Ward, E. Global Patterns Of Cancer Incidence And Mortality Rates And Trends. *Cancer Epidemiology Biomarkers & Prevention* 2010, *19*, 1893-1907.

(3) Omar, S., N. H. M. Alieldin, and O. M. N. Khatib. "Cancer magnitude, challenges and control in the Eastern Mediterranean region." (2007).

(4) Alwan, Nada AS. "commentaries Breast Cancer Among Iraqi Women: Preliminary Findings From a Regional Comparative Breast Cancer Research Project." (2016).

(5) Alwan, Nada. "Iraqi initiative of a regional comparative breast cancer research project in the Middle East." *J Cancer Biol Res* 2.1 (2014): 1016.

(6) Tuceryan, Mihran, and Anil K. Jain. "Texture analysis." *Handbook of pattern recognition and computer vision* 2 (1993): 207-248.

(7) Man, Kim-Fung, Kit-Sang Tang, and Sam Kwong. "Genetic algorithms: concepts and applications." *IEEE transactions on Industrial Electronics* 43.5 (1996): 519-534.

(8) Osowski, Stanislaw, et al. "Application of support vector machine and genetic algorithm for improved blood cell recognition." *IEEE Transactions on Instrumentation and Measurement* 58.7 (2009): 2159-2168.

(9) Akhter, Nazneen, et al. "Feature Selection for Heart Rate Variability Based Biometric Recognition Using Genetic Algorithm." Intelligent Systems Technologies and Applications. Springer International Publishing, 2016. 91-101.

(10) Dy, Jennifer G. "Unsupervised feature selection." *Computational methods of feature selection* (2008): 19-39.

(11) Cohen, Alexander L., et al. "Defining functional areas in individual human brains using resting functional connectivity MRI." *Neuroimage* 41.1 (2008): 45-57.

(12) Raba, David, et al. "Breast segmentation with pectoral muscle suppression on digital mammograms." *Iberian Conference on Pattern Recognition and Image Analysis*. Springer Berlin Heidelberg, 2005.

(13) Zimmerman, John B., et al. "An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement." *IEEE Transactions on Medical Imaging* 7.4 (1988): 304-312.

(14) Srinivasan, G. N., and G. Shobha. "Statistical texture analysis." *Proceedings of world academy of science, engineering and technology*. Vol. 36. 2008.

(15) Bharati, Manish H., J. Jay Liu, and John F. MacGregor. "Image texture analysis: methods and comparisons. "*Chemometrics and intelligent laboratory systems* 72.1 (2004): 57-71.

(16) Haralick, Robert M. "Statistical and structural approaches to texture." *Proceedings of the IEEE* 67.5 (1979): 786-804.

(17) Gaike, Vrushali, et al. "Application of higher order GLCM features on mammograms." *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*. IEEE, 2015.

(18) Gaike, Vrushali, et al. "Clustering of breast cancer tumor using third order GLCM feature." *Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on*. IEEE, 2015.

(19) Shaikh, Shazia, Hanumant Gite, Ramesh R. Manza, K. V. Kale, and Nazneen Akhter. "Segmentation of Thermal Images Using Thresholding-Based Methods for Detection of Malignant Tumours." In *The International Symposium on Intelligent Systems Technologies and Applications*, pp. 131-146. Springer International Publishing, 2016.

(20) Shaikh, Shazia, Nazneen Akhter, and Ramesh R. Manza. "Application of Image Processing Techniques for Characterization of Skin Cancer Lesions using Thermal Images." *Indian Journal of Science and Technology* 9.15 (2016).

(21) Akhter, Nazneen, et al. "Heart-Based Biometrics and Possible Use of Heart Rate Variability in Biometric Recognition Systems." *Advanced Computing and Systems for Security*. Springer India, 2016. 15-29.

ميزة الاختيار على أساس الخوارزمية الوراثية لتصنيف تصوير الثدي بالأشعة مستخدما الخوارزمية التصنيفية (K-means) , وخوارزمية اقرب جار (k-NN) وخوارزمية المسافة الاقليدية (Euclidean Distance)

كامران عادل ابراهيم

كلية التربية طوزخورماتو ، جامعة تكريت ، تكريت ، العراق

الملخص

كانت هناك عدة محاولات تصنيف رقابية لتصوير الثدي بالأشعة السينية في الآونة الأخيرة، ولكن عدد قليل جدا من الاعمال البحثية ركزت على تصنيف غير خاضع للرقابة لاستكشاف إمكانيتها ونقاط ضعفها. لقد حاولت في هذا البحث الاستفادة من المجموعات الغير الخاضعة للرقابة لتصنيف الورم الخبيث والحميد، جمعية تحليل تصوير الثدي بالأشعة (MiniMIAS) لديها 322 تصوير للثدي بالأشعة، منها 64 حميدة و51 خبيثة. استخدمت 115 صورة للتجربة اي 64 حميد و51 خبيث من بين هذه 115 صورة تم استخدام 60% للتدريب و 40% للاختبار. لذلك من 64 حالة حميدة تم استخدام 39 صورة للتدريب والبقية للاختبار, ومن اصل 51 حالة خبيثة تم استخدام 31 صورة للتدريب والبقية الاختبار. التصنيف كان على اساس الميزات المختارة باستخدام الخوارزمية الجينية. كما اجريت تجارب لدراسة اداء كل ميزة مختارة من قبل الخوارزمية الجينية في التصنيف. المجموعات التي تم تحديدها في البداية استخدمت (K-means) لتصنيف 60 عينة غير معروفة مستخدما K-NN. لقد حصل العمل المقترح على نتائج عالية من الدقة 96.23% للعينات الورمية الخبيثة، 95.37% للحميدة. هذا العمل المقترح يُساعد أطباء الأشعة والأورام كمقترح ثان خلال جلسات الفرز للكشف المبكر.

الكلمات المفتاحية: تجميع اوساط–k ، التصنيف الغير المسيطر، خوارزمية اقرب جار (K-NN)، المسافة الإقليدية ، الخوارزميات الجينية .