International Journal of Integrated Engineering:
Special Issue 2018: Data Information Engineering, Vol. 10 No. 6 (2018) p. 209-213.
 DOI: https://doi.org/10.30880/ijie.2018.10.06.030

An Enhancement of Multi Classifiers Voting Method for Mammogram Image based on Image Histogram Equalization

Ashraf Osman Ibrahim[1,2], Ali Ahmed [3]*, Anik Hanifatul Azizah[4], Saima Anwar Lashari[5], Mohamed Alhaj Alobeed[6], Shahreen Kasim[5], Mohd Arfian Ismail[7]

[1]Faculty of Computer and Technology, Alzaiem Alazhari University, Khartoum, Sudan
[2]Arab Open University, Khartoum, Sudan
[3]Faculty of Computer Science and Information Technology, Karary University, Omdurman, 12305, Sudan
[4]School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia
[5,7]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
[6]Information technology, Shendi University, Shendi, Sudan
[7]Faculty of Computer System & Software Engineering, University of Malaysia Pahang, 26200 Gambang, Malaysia

Received 28 June 2018; accepted 5 August 2018, available online 24 August 2018

Abstract: Breast cancer is one the most curable cancer types if it can be diagnosed early. Research efforts have reported with increasing confirmation that the computation methods have greater accurate diagnosis ability. An enhancement of multi classifiers voting technique based on histogram equalization as a preprocessing stage proposed in this paper. The methodology is based on five phases starting by mammogram images collection, preprocessing (histogram equalization and image cropping based region of interest (ROI)), features extracting, classification and last evaluating the classification results. An experimental conducted on different training-testing partitions of the dataset. The numerical results demonstrate that the proposed scheme achieves an accuracy rate of 81.25% and outperformed the accuracy of voting method without using histogram equalization.

Keywords: Multi classifiers, Voting, Histogram Equalization, mammogram Image

1. Introduction

Recently, data mining (DM) received an attention in most of the industrial and social applications. As well known the data is grown very fast due to the technology that used in different side in our daily use [1]. DM tasks usually divided into two part; first is descriptive describe the belongings of the data storage. Other part is predictive task which is execute inference to make predictions [2].

One of the main issues in DM is classification, it means the learning process of the model which can defines dissimilar classes of data and each class is determined. In other word, we can call this process is supervised learning. After building the model successfully, the model can used to classify new data.

For classification task used two part of data, training and testing data. In the first stage of the model use training part to classify the dataset based on the attribute or label, which shows which class belongs to which record. Usually, the model that created is appeared as decision tree or a set of rules.

As important issues of the model and the algorithm that yields the model ability of the model to predict the correct class label of new data. The computational cost associated with the algorithm, and the scalability of the algorithm[3].

The classification of the medical images is done by extracting features and describe the important class of each. Mammography is a good method used for early cancer detection. Breast Cancer is a disease that threat women in the whole world especially in the developed country.

Mammogram images actually is an X-ray images for breast cancer. It plays the main role to detect the breast cancer disease in early stage.

All mammogram images of breast cancer need pre-processing step to enhance the quality and performance screening of the image in computer aided diagnosis (CAD),the CAD process can mainly help in decrease the number of errors in diagnosis phase. In CAD process, the classification algorithms can help in searching or help in classifying lesions in benign or malignant types. Actually, the CAD system contains of numerous modules, for instance, pre-processing, segmentation and classification process of pathological cases [4].

*Corresponding author: alikarary@gmail.com

Table 1: Statistical functions for features extraction of a mammogram images.

Function	Formula	Description
Mean	$L-1$	A measure of average intensity
Standard	$MMMMMMMM = \sum Z_{ii} * P(Z_{ii})$ $ii=0$ $ll-1$	Secondmoment
Deviation	$ssssss = \sum (z_{ii} - m)^2 * p(z_{ii})$ $ii=0$ ll	about the mean
Skewness	-11 $SSSSMMSSMMMMssss = \sum zz_{ii} - m^{33} * pp(zz_{ii})$ $ii=0$ ll	Third moment about the mean
Kurtosis	-11 $KKKKKKssKKssiiss = \sum zz_{ii} - m^{44} * p(zz_{ii})$ $ii=00$ $ll-1$	Fourth moment about the mean
Contrast	$\text{Contrast} = \sum z - m^2 * p(z)$ $ii=0$	Standard deviation of pixel intensities
Smoothness	$s = 1 - \frac{1}{(1+\sigma)}$	Measures the relative intensity variations in a region

From table 1, zi is a random variable indicating intensity, p(zi) is the histogram of the intensity levels in a region, L is the number of possible intensity levels, is the standard deviation[3].

Feature extraction is an important stage, it can make analysis of the objects and images due to extract the most protuberant features that are correspondence of different objects in class. Six statistical features used in this study and their formulas are described in Table 1, these features are commonly used and suggested by many related studies [5].

Ensemble classification algorithms defined as learning algorithms that construct a set of classifiers that can be used to classify a new set of data either by equal weights such as in Bootstrap aggregating or Bagging method, or by different weights such as in Boosting method [6]. This paper propose multi-classifier approach based on SVM, Bayes Naïve and K-nearest Neighbors classifiers and voting method for mammogram images. The voting proposed method here is depends on the three different classifiers, mentioned above, rather than performs a group or a set of classification experiments based single classifier algorithm as in the known previous Ensemble methods.

In this paper we make an Enhancement of multi classifiers voting technique for mammogram image based on image histogram equalization

This paper is organized into several sections: section 2, the methods are described. In addition, the proposed methodology is introduced. The experiments are shown in section 3. The results and discussion are presented in section 4 and the paper is concluded in section 5.

2.1 Dataset Description

The data sets used for this study taken from Mammogram Image Analysis Society (MIAS) all images are 1024 × 1024 pixels. The data base has Benign and Malignant breast cancer images, 119 images is a total patients. All images contain the positions of abnormalities [14]. Table 2 shows the dataset description.

Table 1: MIAS Database Details.

Column No	Description	Details
1st	MIAS database reference number	No details
2nd	Character of background tissue	No details
3rd	Class of abnormality present	
4th	Severity of abnormality	
5th 6th	X,Y image-coordinates of center of Abnormality.	
7th	Approximate	

2. Methods

In this section, we describe the dataset that used, in addition the methods that used for enhancing the multi classifier and voting method and proposed method as well.

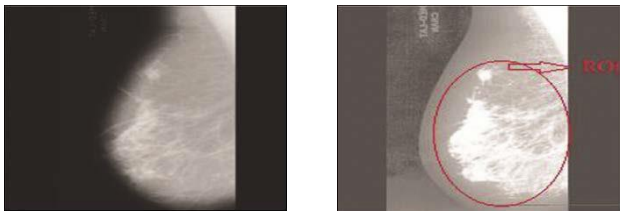
radius (in pixels) of a Circle enclosing the abnormality.

F - Fatty
G - Fatty-glandular
D - Dense-glandular

CALC – Calcification CIRC-Well- defined/circumscribed
SPIC-Spiculated masses
MISC-Other,ill-defined masses
ARCH-Architectural distortion
ASYM-Asymmetry
NORM-Normal

B – Benign
M – Malignant

No details



a. Image Cropping b. ROI detected

Fig. 1 A Sample images of MIAS Data set

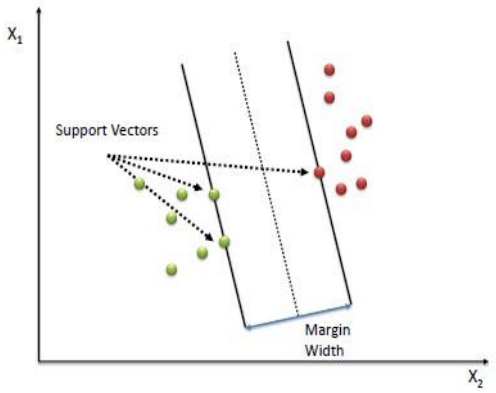


Fig. 2 maximizes the margin by pick the separating hyper plane

2.2 Histogram Equalization

Histogram equalization is an entirely automatic procedure used to stretch out the gray levels to make the appearance of image is better. The main idea of this process to calculate the new values for the gray levels

according to the following formula:.

$$\frac{(M_1 + MM_2 + \cdots + M_{ii})}{M} * LL \quad (1)$$

where L is the different gray levels with values start from 0, 1, 2, . . . , L – 1.The gray level i happens *ni* times in the image. This proposed preprocess was applied before feature extraction stage.

2.3 Individual Classifier Algorithms

The classification process using individual classification approaches or methods involves four major steps namely image collection and preprocessing if any, image cropping, feature extraction, apply classification algorithm. Image Cropping was applied to the images based on given X,Y and R values defined in MIAS dataset, Figure 1 shows an example of this image. The following paragraphs give basic ideas about our three individual classification algorithm propose to use in this study; those are NBC, KNN and SVM, these methods are equations, the other probability form is used for nominal data.

$$\mu\mu = \frac{1}{M} \sum_{ii=1}^{MM} x_{ii}$$

$$\sigma\sigma = \frac{1}{MM - 1} \sum_{ii=1}^{MM} (x_{ii} - \mu)^2 \Big)^{0.5}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} M^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

where : μ : mean , σ : standard deviation , σ^2: variance

2.5 K-nearest-neighbor

K-nearest-neighbor method (KNN) is one of the non-parametric techniques. It is well known and simple method, it used in pattern recognition field and used for classification and regression purpose. KNN can calculate the distance between a query scenario and a set of scenarios in the data set by using distance function d(X, Y). Both of X and Y are where scenarios consisting of N attributes. The X values is ($X = \{x_1 \ldots x_n\}$) and Y values is ($Y=\{y_1 \ldots y_n\}$). The major two common distance functions used with these classification algorithms are Absolute Distance measuring and Euclidean Distance measuring. In this study we use the second distance measure as in the following formula as shown in equation 3.

$$s_{EE}(x, y) = \sum_{ii=1}^{NN} x_{ii}^{2 - yy^2} \quad (3)$$

used in many similar studies [7-9].

2.4 Individual Classifier Algorithms

Naïve Bayes classifier (NBC) used in this study is based on the equation 2 below, which is found to apply for the continuous data and it illustrated in the following

2.6 Support Vector Machine

Support Vector Machine (SVM) is one of the machine learning techniques. It is a supervised learning technique. In addition, it is a discriminative classification method realized by a separating hyper-plane and maximizes the margin between the two classes. The hyper-plane or a line which is separating a plane in two parts.

The basic idea of this classifier is shown in Figure 2 below and the following algorithm.

The three main ideas of the algorithm:

1. Identifythe optimal hyper-plane is maximize margin.
2. Separable problems: it have a penalty expression for misclassifications.
3. It is easier to classify with linear decision surfaces: reformulate problem so that data is mapped implicitly to this space.

2.7 Multi Classifier Voting Algorithm

This section includes experiments that led to the results of applying simple voting method on the three classifiers shown in the Fig 3.

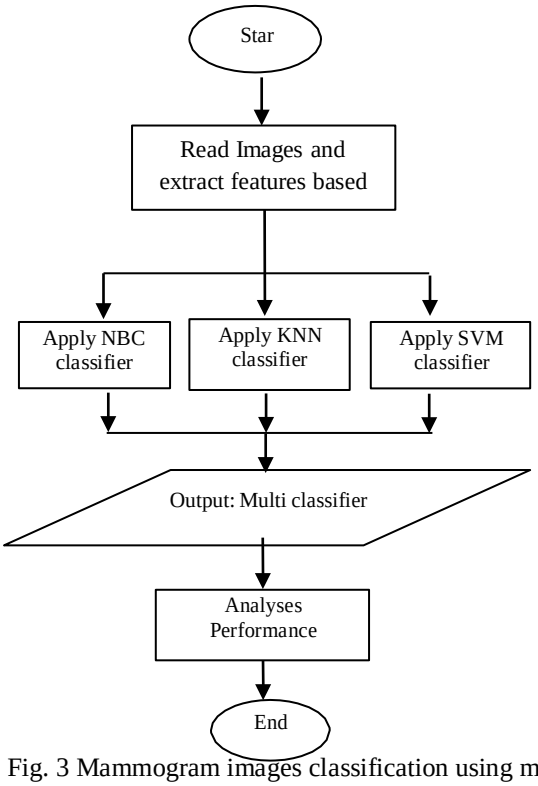


Fig. 3 Mammogram images classification using multi voting classifiers

The multi-classifier method used in this study based on the individual results obtained by each single classifier discussed above. The concept of our proposed approach depends on the voting method as described in Fig 4.:

```
Begin
    Read class labels predict by NBC classifier
    Read class labels predict by SVM classifier
    Read class labels predict by KNN classifier
    s1 = size of test instances
    For counter=1 to s1
        sum=summation of three predicted instances
        If sum >=2 Then
                Set voting_output =1
        Else
                Set voting_output =0
        End if
    End for
End
```

Fig. 4The Multi Classifier Voting Algorithm

3. Experiments

The dataset used in this study is Mammographic Image Analysis Society (MIAS) [10]. It has a total of 119 instances of breast cancer patients with each, either having malignant or benign type of cancer, this data set used by many latest related studies [11-13].

After we apply the required preprocessing stages which include the histogram process and image cropping (based on X, Y and R values) and feature extraction (using the six function described earlier in Table 1) we divide the data set into three different sizes of training and testing which are 60%, 70%, 85% for training and 40%, 30%, 15% for testing purpose, the overall accuracy is calculated using the equation defined in equation 4.

$$CCCC \frac{(TTPP+TTNN)}{FFPP} \quad (4)$$

where: TP : True Positive , FP: False Positive , TN: True Negative, FN : False Negative

4. Results and Discussions

The first round of experiment is carried using the three individual algorithms, while the second round used the predicted class labels as an input and implemented the previous proposed algorithm. The two experiments was done after histogram process. All the results of our experiments are shown in Table 2 and 3 below.

Table 3: Classification accuracy using NBC, KNN, SVM and Voting Methods before histogram equalization

Data set Percentages	NBC	KNN	SVM	Voting
60-40	0.5532	0.5319	0.5106	0.5957
70-30	0.6000	0.5429	0.5714	0.6571
85-15	0.5882	0.6471	0.6471	0.7647

Table 4: Classification accuracy using NBC, KNN, SVM and Voting Methods after histogram equalization

Data set Percentages	NBC	KNN	SVM	Voting
60-40	0.5735	0.5520	0.5316	0.6197
70-30	0.6233	0.5639	0.5824	0.6773
85-15	0.5902	0.6773	0.7379	0.8125

As we have seen in the above table 4, the highest data set percentage that gives good accuracy is 85%, and it was 0.8125, while in 70% the accuracy was 0.6773 and in 60% the accuracy was 0.6197. Generally we found that the accuracy was increased after applying the histogram equalization and voting in the three different percentages of data set.

5. Results and Discussions

An enhancement of multi classifiers voting approach for mammogram image based on image histogram equalization is addressed in this paper to improve the classification accuracy of the breast cancer.

The methods implement voting classification method based on three classifiers which were; K-nearest-neighbor, support vector machine and naïve bayes classifier. The proposed technique implemented on MIAS medical images data, the data divided into two parts; training part that used 60, 70, 85 percentage of data set for training and the rest were use for testing purpose. For the future work will try to use more than three classifiers to fulfil better results in terms of accuracy [15-16].

Acknowledgement

The first author gratefully acknowledges Arab Open University for supporting research and development. The second author would like to give a great thanks to the Faculty of Post Graduate Studies and Scientific Researches at Karary University, for their support.

References

[1] Ramos, N. M., J. M. Delgado, R. M. Almeida, M. L. Simões and S. Manuel (2016). Data Mining Techniques. Application of Data Mining Techniques in the Analysis of Indoor Hygrothermal Conditions, Springer: 13-30.

[2] Freitas, A.A., Data mining and knowledge discovery with evolutionary algorithms. 2013: Springer Science & Business Media.

[3] Zhao, Y. (2015). "Data mining techniques."

[4] Beura, S., B. Majhi, and R. Dash, Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing, 2015. 154: p. 1-14.

[5] Aarthi, R., K. Divya, N. Komala and S. Kavitha (2011). Application of Feature Extraction and clustering in mammogram classification using Support Vector Machine. Advanced Computing (ICoAC), 2011 Third International Conference on, IEEE.

[6] Ren, Y., P. Suganthan and N. Srikanth (2015). "Ensemble methods for wind and solar power forecasting—A state-of-the-art review." Renewable and Sustainable Energy Reviews 50: 82-91.

[7] Krishnaveni, S., R. Bhanumathi and T. Pugazharasan (2014). "Study of Mammogram Microcalcification to aid tumour detection using Naive Bayes Classifier." International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 3(3).

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

[9] Shirazi, F. and E. Rashedi (2016). Detection of cancer tumors in mammography images using support vector machine and mixed gravitational search algorithm. Swarm Intelligence and Evolutionary Computation (CSIEC), 2016 1st Conference on, IEEE.

[10] The mini-MIAS database of mammograms, "http://peipa.essex.ac.uk/info/mias.html", on March 2017.

[11] Wajid, S. K. and A. Hussain (2015). "Local energy-based shape histogram feature extraction technique for breast cancer diagnosis." Expert Systems with Applications 42(20): 6990-6999.

[12] Gedik, N., A. Atasoy and Y. Sevim (2016). "Investigation of wave atom transform by using the classification of mammograms." Applied Soft Computing 43: 546-552.

[13] Lashari, S. A., R. Ibrahim and N. Senan (2015). "Fuzzy Soft Set based Classification for Mammogram Images." International Journal of Computer Information Systems and Industrial Management Applications 7: 66-7.

[14] http://peipa.essex.ac.uk/info/mias.html 2-12-2016 At 2:15 am.

[15] Azmi A., Khaman K.H., Ibrahim S., Md Khairi M.T., Faramarzi M., Abdul Rahim R., Md Yunus M. A. (2017). "Artificial Neural Network and Wavelet Features Extraction Applications in Nitrate and Sulphate Water Contamination Estimation". nternational Journal of Integrated Engineering. 9(4): 64-75.

[16] Ismail, M. A., Mezhuyey, V., Deris, S., Mohamad, M. S., Kasim, S., and Saedudin, R. R. (2017). "Multi-objective Optimization of Biochemical System Production Using an Improve Newton Competitive Differential Evolution Method", International Journal on Advanced Science, Engineering and Information Technology, 7: 1535-1542.