American Journal of Data Mining and Knowledge Discovery
2021; 6(2): 31-35
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20210602.13
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

Classification of Breast Cancer Image Using Data Mining Techniques

Mohamed Alhag Alobed[1], Ali Ahmed[2], Ashraf Osman Ibrahim[3]

[1]Tumor Therapy and Cancer Research Center, Shendi University, Shendi, Sudan

[2]Faculty of Computer Science and Information Technology, Karary University, Omdurman, Sudan

[3]Faculty of Computer Science and Information Technology, Alzaiem Alazhari University, Khartoum, Sudan

Email address:
mohamedelhaj123@hotmail.com (M. A. Alobed), alikarary@gmail.com (A. Ahmed), ashrafosman2@gmail.com (A. O. Ibrahim)

To cite this article:
Mohamed Alhag Alobed, Ali Ahmed, Ashraf Osman Ibrahim. Classification of Breast Cancer Image Using Data Mining Techniques. *American Journal of Data Mining and Knowledge Discovery*. Vol. 6, No. 2, 2021, pp. 31-35. doi: 10.11648/j.ajdmkd.20210602.13

Received: October 12, 2021; **Accepted**: November 1, 2021; **Published**: November 25, 2021

Abstract: Breast cancer is the most common malignancy disease that affects female population and the number of affected people is the second most common leading cause of cancer deaths among all cancer types in the developing countries. Mammography is the most effective method for detection of early breast cancer to increase the survival rate. This paper presented the classification method for mammogram Image using the decision tree techniques. Three measures were used to evaluate performance in terms of accuracy, sensitivity, and privacy. The aim of the study is to determine the best decision tree classifier for medical datasets classification. The study emphasizes five phases; starting with collecting images, pre-processing (image cropping of ROI), features extracting, classification and end with testing and evaluating. Experimental results show that Random Forest has a better performance than ID3, J48.

Keywords: Mammograms, Breast Cancer, Decision Tree, Early Detection, Image Classification

1. Introduction

Breast cancer is a life killing disease through the existence of debilitating growths influencing ladies mostly after the age of 30 all over the world [1]. Early diagnosis of the breast cancers by the radiologist reduces the death rate globally. Many techniques are available for the detection of breast cancers among which digital mammography is the familiar and successful technique currently used by the radiologists [2].

Mammograms are collected from patients who are suspected for breast cancers mostly as full field mammograms where the image detection and classification are high due to the high image quality [3]. Mammography cannot stop or decrease breast cancer can be supportive only in detecting the breast cancer at early stages to increase the survival rate [4, 5]. Regular screening can be a successful strategy to identify the early symptoms of breast cancer in mammographic images.

This examination also ensures other pathologies detection suggesting cancer nature as being benign, malignant, or normal. The most important improvement is breast imaging which is possible due to the advancement in digital mammography [6]. Medical images classification can play an important role in diagnostic and teaching purposes in medicine. It is also a form of data analysis that extracts models describing important data classes. Numerous methods have been created to classify masses into benign and malignant categories by using the different classification method [7]. The researchers proposed method aims to apply image mining for breast mammograms to detect and classify the cancerous tissue without any help of radiologist or medical specialist. A total of twenty-six features including GLCM features and histogram intensity features were extracted. A dataset of images consisting of 322 images taken from a MIA's dataset were used in the experiment. Results show that the proposed method has achieved 97.7% accuracy [8]. The researchers performed a comparative study on the performance of binary classifiers. They have used the Wisconsin breast cancer dataset with 10 attributes and not the breast tissue dataset. Moreover, they have not brought out the effect of feature selection in classification. Their experimental study was restricted to four classification algorithms viz. ID3, C4.5, K- Nearest Neighbors (K-NN) and Support Vector Machines (SVM) [9].

Classification methods are one of the most fundamental and important tasks in data mining and machine learning. Many of the researchers performed experiments on medical datasets using decision tree classifiers [10]. The aim of the study is to determine the best decision tree classifier for medical datasets classification.

In [11], researchers analyzed the performance of decision tree classifiers on various medical datasets in terms of accuracy and time complexity which proved that CART is the best. More recent research presented in [12], concerned the identification of breast cancer patients for whom chemotherapy.

Could prolong survival time and is treated here as a data mining problem.

The remainder of this paper is organized as follows: Section 2 introduces the materials and methods and Testing and evaluation. The experiment is given in Section 3. Results and discussions are provided in Section 4. Finally, Section 5 concludes the study.

2. Materials and Methods

This study emphasizes five phases starting with images collection, pre-processing, features extracting, classification of mammogram and end with testing and evaluation followed by detail about each phase Figure 1 shows the five steps research method.

2.1. Mammogram Images Collection

Dataset used in this study is downloaded from the MIAS (Mammographic Image Analysis) database website [13]. This dataset was recently used by many researchers. MIA's dataset is used for experimentation purpose which is a standard and publicly available dataset. The size of each mammogram is 1024 × 1024 pixels and 200 micron resolution. MIAS contains a total of 322 mammograms of both breasts (left and right) of 119 patients.

2.2. Image Cropping Based on ROI

Next step is to extract Regions of Interest (ROI). ROI's are defined as regions containing user-defined objects of interest. Here we applied crop technique to the images; a cropping operation was employed in order to cut the interest parts of the image. Cropping removed the unwanted parts of the image usually peripheral to the regions of interest as shown in Figure 2.

2.3. Feature Extraction

The accurate classification and diagnostic rate mainly depend upon robust features, particularly while dealing with mammograms, after cropping the Region of Interest (ROI) from [x] position to [y] position and [radius] depends on the MIAS dataset. This stage applies the six functions (Mean, Standard Deviation, Skewness, Kurtosis, Contrast, and Smoothness) to extract the feature values from each mammogram image. The following paragraphs give more details about the six functions used to extract features values.

2.4. Classification of Mammograms

The result of the previous three phases converts the data to numeric values. In this stage, we apply three individual classifiers, for different decision trees namely ID3, Random Forest and J48. The process of classifying features into their respective classes, such as normal and abnormal or benign and malignant. We have used the WEKA toolkit classification to experiment with these three algorithms [14]. The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. WEKA version 3.7 was utilized as a data mining tool to evaluate the performance and effectiveness of the breast cancer preliminary prediction models.

Evaluate. In this paper, we presented the classification method for mammogram Image using the decision tree techniques (Decision ID3, Random Forest and J48) to apply on the medical image that is extracted from MIA's data set. In the next paragraphs, we review and present a brief overview of the three classifiers that are used in the classification stage of the mammogram images.

2.5. Random Forest

Random Forest (RF) is an approach which has been proposed by Breiman for classification tasks. It mainly comes from the combination of tree-structured classifiers with the randomness and robustness provided by bagging and random feature selection [15]. The classification is performed by sending a sample down in each tree and assigning it the label of the terminal node it ends up with. At the end, the average vote of all trees is reported as the result of the classification. Random forest is very efficient with large datasets and high dimensional data.

2.6. ID3

The ID3 algorithm is considered as a very simple decision tree algorithm developed by Quinlan in 1986 [16]. ID3 uses information gained as splitting criteria. The growing stops when all instances belong to a single value of target feature or when best information gain is not greater than zero. ID3 does not apply any pruning procedures nor does it handle numeric attributes or missing values. It only accepts categorical attributes in tree building. Also does not support noise data. To remove the noise preprocessing technique has used. ID3 algorithm cannot handle the continuous attributes for that discretization is used to convert continuous attributes to categorical attributes.

2.7. J48

A decision tree is a predictive machine-learning model to decide a new sample's target value (dependent variable) dependent on available data's varied attribute values. The internal nodes of a decision tree denote various attributes; inter-nodal branches reveal attribute's possible values in observed samples, while terminal nodes provide information of the dependent variable's final value [17].

3. Testing and Evaluation

To test and evaluate the performance of the proposed method, different quantitative measures have been used, such as accuracy, sensitivity, specificity and area under The Curve (AUC). These can be calculated by using mathematical equations shown in equations (1), (2) and (3).

Accuracy

It has been used and can be calculated by using mathematical equation:

$$CR = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP is True positive, FP is false positive FN is false negative and TN is true negative.

Specificity

Ability of a classifier to identify the negative results is estimated as specificity, given as: testing phase. The results are presented in the upcoming section. To test the performance of the proposed method, We measure accuracy, sensitivity, and specificity, to show the performance of the proposed method.

$$\text{Specificity} = \frac{TN}{N+FP} \quad (2)$$

Sensitivity

In this study, MIAS data set was used for three decision tree classifiers based on continues data set. The highest precision was given with a good accuracy for Random Forest accuracy 90.4%, sensitivity 88.09%, Specificity 83.5%, while in Decision ID3 accuracy 87.09% sensitivity 80.03%, Specificity 81.04%. and J48 85.00% accuracy, sensitivity 82.00%, Specificity 79.90% Generally, the accuracy, sensitivity, and specificity was Ability of a classifier to identify the positive results quantitatively is evaluated as Sensitivity which is given as:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

4. Experiment

To conduct experiments in the proposed method, MIA's database was used. The MIAS database was created to contain two experimental datasets on the same images. The difference between them is that in the first dataset the images are split in two classes: normal or abnormal. MIA's database is a set of 322 commented images. The abnormal images in this database contain the coordinates and the radius. Matlab 2010 was used to extract all features methods.

WEKA tools were used for images classification with 60-40% percentage split. 60% of the samples are used in the training phase and the remaining samples are used in the increased as shown in Table 1.

5. Results and Discussion

In this study, MIAS data set was used for three decision tree classifiers based on continues data set. The highest precision was given with a good accuracy for Random Forest accuracy 90.4%, sensitivity 88.09%, Specificity 83.5%, while in Decision ID3 accuracy 87.09% sensitivity 80.03%, Specificity 81.04%. And J48 85.00% accuracy, sensitivity 82.00%, Specificity 79.90% Generally, the accuracy, sensitivity, and specificity was increased as shown in Table 1.

After applying three different classifiers for the decision trees, we calculated the overall Accuracy, sensitivity, specificity by using mathematical equations shown earlier, the final results are shown in Table 1 and Figure 3 is the graphical representation of the classification accuracy, sensitivity, and Specificity. It is observed from Table 1 that the best accuracy is achieved by Random Forest. It is observed from the graphs that the Accuracy, sensitivity, specificity is better for decision Random Forest.

We compared three classifiers methods in this experiment: decision trees techniques (Decision ID3, Random Forest, J48). Figure 4 shows the experimental results of the three classifiers of the Decision Tree. The main measurement of comparison is accuracy. In a previous study [18] Researchers proposed an automatic mammogram classification technique using wavelet, consisted of four classifiers based on Decision Tree J48, CART and CFS, Decision stump. Classification accuracy is achieved by Decision stump 80.00%, J48 7.00%, CART 60.00%, Decision stump with CFS 80.00%, J48 with CFS 80.00%, CART with CFS 70.00%. Future work can explore optimizing the classifiers for improving the accuracy.

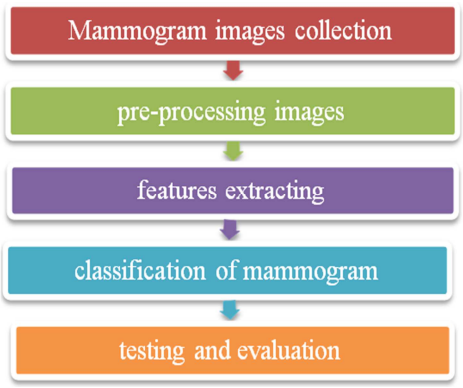

Figure 1. *Research phases.*

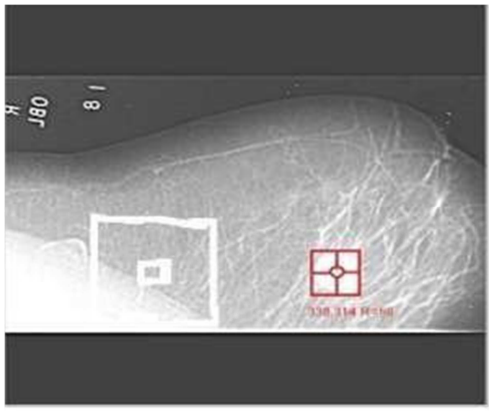

Figure 2. *Full Mammogram with detected region of interest.*

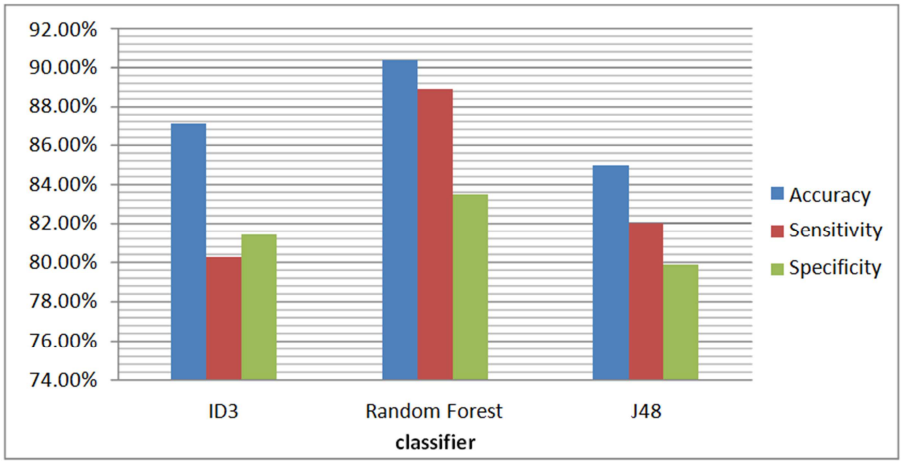

***Figure* 3.** *Result of classification.*

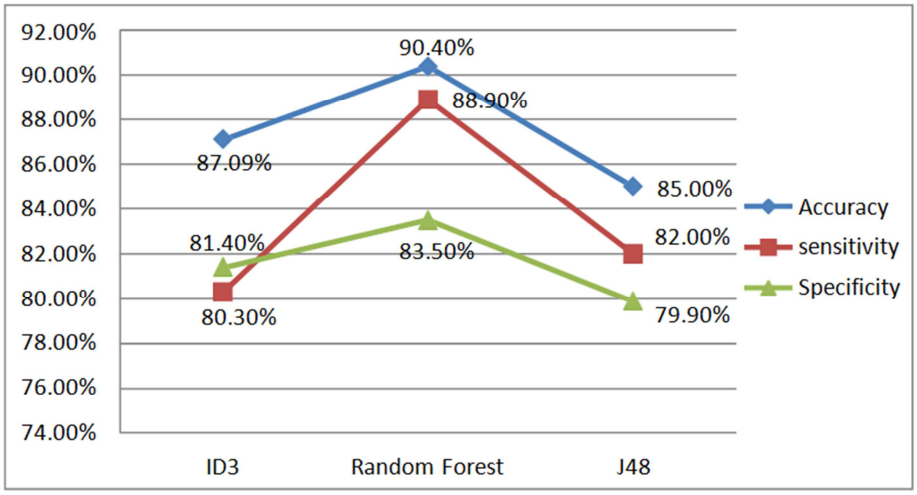

***Figure* 4.** *The compared results.*

Table 1. *Results of the three classifiers.*

classifier	Accuracy	sensitivity	Specificity
ID3	87.09%	80.3%	81.4%
Random Forest	90.4%	88.9%	83.5%
J48	85.00%	82.00%	79.90%

6. Conclusion

This study aimed to determine the best decision tree classifier using ID3 classifiers, Random Forest and J48 that all these decision tree algorithms are applied on medical image that is extracted from MIAS data set. The study contains two main processes; the first one is build the classifier using the 60 percentage from the dataset. The second; building the classifier using the 40 percentage to test the classifier. Classification accuracy is achieved by Decision ID3 87.09% sensitivity 80.03% Specificity 81.04%, Random Forest Classification accuracy 90.4% sensitivity 88.09% Specificity 83.5%, J48 85.00% sensitivity 82.00% Specificity 79.90%. So, in future we shall facus on performing the experiments with ensemble technique on the specified decision tree classifiers for further analysis. It can explore optimizing the classifiers for improving the accuracy.

References

[1] Pareek, A. and S. M. Arora, Breast cancer detection techniques using medical image processing. Breast cancer, 2017. 2 (3).

[2] S. Punitha, S. Ravi, et al., Breast Cancer Detection using Classification Techniques in Digital Mammography: International Science Press, I J C T A, 9 (7), 2016, pp. 3123-3134, ISSN: 0974- 5572.

[3] Curtis, C., et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 2012. 486 (7403): p. 346-352.

[4] Perou, C. M., et al., Molecular portraits of human breast tumours. Nature, 2000. 406 (6797): p. 747-752.

[5] Tang, J., Rangayyan, R. M., Xu, J., El Naqa, I., & Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. Information Technology in Biomedicine, IEEE Transactions on, 13 (2), 23.

[6] Smith, R. A., V. Cokkinides, and H. J. Eyre, American Cancer Society guidelines for the early detection of cancer, 2006. CA: a cancer journal for clinicians, 2006. 56 (1): p. 11-25.

[7] Affar, M. A., Hybrid Texture based Classification of Brea Mammograms using Ad boost Classifier. International Journal of Advanced Computer Science and Applications, 2017. 8 (5).

[8] Choi, J. P., T. H. Han, and R. W. Park, A hybrid bayesian network model for predicting breast cancer prognosis. Journal of Korean Society of Medical Informatics, 2009. 15 (1): p. 49-57.

[9] Aswinikumarmohanty, Sukantakumar swain, Pratapkumarchampati, Sarojkumarlenka, "Image Mining for Mammogram Classification by Association Rule Using Statistical and GLCM features", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, 2011.

[10] S. Aruna, Dr S. P. Rajagopalan and L. V. Nandakishore, 2011 Knowledge Based Analysis Of Various Statistical Tools In Detecting Breast Cancer. 2011.

[11] Y. J. Lee, O. L. M. W. H. W. Survival -Time Classification of Breast Cancer Patients. 2008 [cited 2017; Available from: http://www.cs.wisc.edu/dmi/annrev/rev0601/uj.ppt.

[12] Clark, A. F. The mini-MIAS database of mammograms.

[13] Usha, S. and S. Arumugam, Calcification Classification in Mammograms Using Decision Trees. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2016. 9 (9): p. 2127-2131.

[14] E. Frank, M. Hall, and L. Trigg, "Weka: Waikato environment for knowledge analysis," The University of Waikato, Hamilton, New Zealand, 1999.

[15] Mariana R. Mendoza, Guilherme C. da Fonseca, Guilherme Loss- Morais, Ronnie Alves, RogerioMargis, Ana L. C. Bazzan, "Predicting Human MicroRNA Target Genes with a Random Forest Classifier", plos, 2013.

[16] J. R. Quinlan, "Induction of decision tree". Journal of Machine Learning 1, 1986, Pg. no: 81-106.

[17] Zhang, Y., Zhao, Y., A Comparison of BBN, AD Tree and MLP in separating Quasars from Large Survey Catalogues, ChJAA 7, 289- 296, 2007.

[18] S. Usha, S. Arumugam (2015). "Calcification Classification in Mammograms Using Decision Trees" International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol: 9, No: 9, 201.