American Journal of Data Mining and Knowledge Discovery
2023; 8(2): 18-22
http://www.sciencepublishinggroup.com/j/ajdmkd
doi: 10.11648/j.ajdmkd.20230802.11
ISSN: 2578-7810 (Print); ISSN: 2578-7837 (Online)

An Efficient Prediction of Breast Cancer Diagnosis Using Data MiningTechnique in Tumor Therapy and Cancer Research Center Shendi

Mohamed Alhag Alobed[1], Namareg Mohamed Ibrahim Ahmed[2]

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

[2]College of Computer Science, Sinnar University, Sinnar, Sudan

Email address:

mohamedelhaj123@hotmail.com (Mohamed Alhag Alobed), namareg29012@gmail.com (Namareg Mohamed Ibrahim Ahmed)

To cite this article:

Mohamed Alhag Alobed, Namareg Mohamed Ibrahim Ahmed. (2023). An Efficient Prediction of Breast Cancer Diagnosis Using Data MiningTechnique in Tumor Therapy and Cancer Research Center Shendi. *American Journal of Data Mining and Knowledge Discovery*, *8*(2), 18-22. https://doi.org/10.11648/j.ajdmkd.20230802.11

Received: September 6, 2023; **Accepted**: October 13, 2023; **Published**: November 24, 2023

Abstract: Breast cancer is the disease that most common malignancy affects female. It has been considered as a second most common leading cause of cancer death among other type of cancer, specifically in developing countries. Most of the previous researches in mammogram images achieved low classification accuracy that because of either inaccurate features or improper classifier methods. Mammography is the most effective method for detection of early breast cancer to increase the survival rate. The aim of this research is to Enhancement of Mammogram Images Classification Accuracy Using Data mining technique (decision tree classifier) for medical datasets classification that can aid the physician in a mammogram image classification as benign or malignant. The study the study methodology focuses on six phases starting with image collection, pre-processing (cutting images of the area of interest), feature extraction, feature selection, classification, and ending with testing and evaluation. Experimental results using a mammogram analysis dataset from Tumor therapy and Cancer Research Center, Shendi Sudan, showed that this approach achieves an accuracy of 97.04%.

Keywords: Breast Cancer, Mammogram Images, Data Mining Technique, Decision Tree

1. Introduction

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically [1]. Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [2]. Data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions [3].

The knowledge discovery process comprises six phases: data selection, data cleaning, enrichment, data transformation or encoding, data mining, and the reporting and display of the discovered information. Data mining is typically carried out with some end goals or applications. Broadly speaking, these goals fall into the following classes: prediction, identification, classification, and optimization [4].

The existence of breast cancer in women has increased significantly in the recent years. In various types of cancer breast cancer is one of the important causes of death among middle and old aged women [5]. Throughout their lifetime, more than 80% of women suffered from this disease. In worldwide breast cancer caused 458,503 deaths (13.7% of cancer deaths in women). In the year 2008 physical examination is one of the ways to detect the breast cancer however the effectiveness of this method is limited by the subjective ability of doctors [6]. Another way to detect breast cancer is screening mammography. It is one of the most effective techniques for the early detection of breast cancer. Screening mammography minimized the death rates from

breast cancer among the women aged between 40 - 70 years [7]. To increase the accuracy double reading of mammogram must be used. But this is not effective and also it requires twice the radiologists' reading time.

Mammogram is the process of using low-dose amplitude X-rays to examine the human breast and is used as a diagnostic and a screening tool. A diagnostic mammogram is used to diagnose breast disease in women who have breast symptoms or an abnormal result on a screening mammogram [8]. Screening mammograms are used to look for breast disease in women who are asymptomatic that is those who appear to have no breast problems. But both screening and diagnostic mammograms depends on the radiologist accuracy in reading the mammograms [9].

Hence efforts have been made to develop a breast cancer classification system [10] to help radiologist in the analysis of mammograms in hospitals which increase the accuracy of diagnosis as well as to improve the uniformity of interpretation of images by the use of the computer's results as a reference.

With the help of Computer Aided Diagnosis (CAD) breast cancer detection in mammogram images is made easier nowadays. Most of the clustering and classification techniques used for diagnosis of breast cancer consider only processing of clinical features given by the radiologist report or the statistical features extracted from the mammogram [11].

The effectiveness of digital mammography in detection of breast cancer is currently under investigation. A variety of algorithms have been developed by independent investigators for use with digital mammograms. The use of computers in processing and analyzing biomedical images allows more accurate diagnosis by a radiologist. Humans are susceptible in committing errors and their analysis is usually subjective and qualitative. Objective and quantitative analysis facilitated by the application of computers to biomedical image analysis leads to a more accurate diagnostic decision by the physician [12].

2. Literature Review

Previous years there are a number of papers that are directly related to classification accuracy of mammograms images. One of these paper Researchers [13] proposed a survey on breast cancer detection techniques, the result of their survey showed that the various medical image classification techniques accuracy is less due to the less number of feature.

Also researchers [14] proposed an automatic mammogram classification technique using decision tree classifier, proposed technique achieve low accuracy and they recommended increasing the accuracy of j48 algorithm.

The aim of the study is to an Enhancement of Mammogram Images Classification Accuracy Using Data mining technique (decision tree classifier) for medical datasets classification.

The remainder of this paper is organized as follows: Literature Review in Section II, Section 3 presents Materials, Methods, Testing, and Evaluation. The experiment is presented, Results and discussions in Section 4. Finally, Section 5 concludes the study.

3. Methodology

The study emphasis on six phases which were described in Figure 1 followed by detailed about each phase.

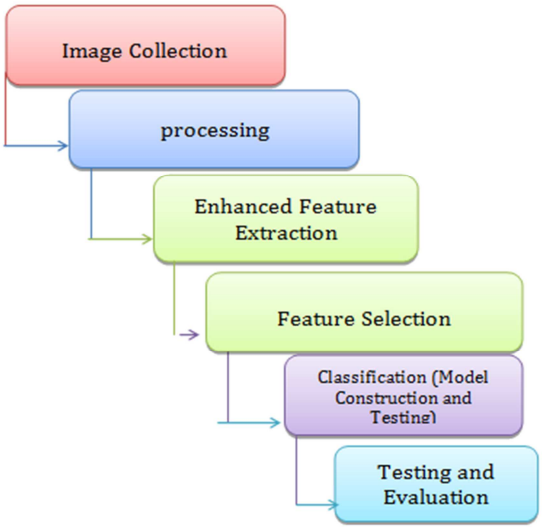

Figure 1. *Research phases.*

Phase 1: Image Collection

In this study we used mammography data from a machine learning repository for the Tumor therapy and Cancer Research Center patients' dataset. A random sample was taken containing 9 features including categories. The data contains two classes of cancer as malignant and fibro adenoma as benign.

Phase 2: Preprocessing

Exactly most of mammograms are noisy and inconsistent, so the enhancement of these mammograms and interpret mammogram images in their raw form is very difficult. So preprocessing techniques are needed to aid in removing outliers, noise and inconsistencies. Also images contain artifacts as labels and noise that need to be removed. In this research this is done by cropping the images as the first phase. Cropping basically removes the unwanted parts of the image. Cropping operation is done automatically by sweeping through the image and finding those areas in the image that had a mean intensity less than a certain threshold and these parts of an image were cut horizontally and vertically. ROI extracted by entering coordinates X, Y and radius in pixels, according to data provided by the MIAS database for each abnormal mammogram image [15].

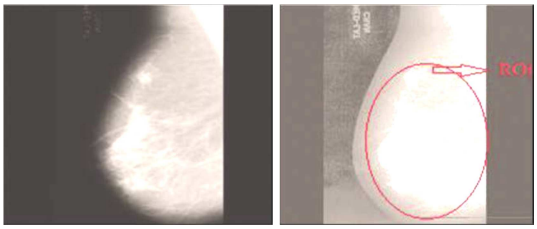

Figure 2. *ROI extracted images.*

Phase 3: Enhanced Feature Extraction

Feature extraction is a process to analyze objects and images to extract the most prominent features that are correspondence of various classes of objects. The purpose of feature extraction process is to characterize raw image into its compressed form to simplify decision-making process such as pattern classification, to acquire high classification ratio. Therefore improving feature extraction process will be likely improving performance of classification algorithm. Six statistical features are used in this study and their formulas are described below, these features are commonly used and suggested by many related studies [15].

Calculating Mean

The Mean is a measure of the average intensity of the neighboring pixels of an image. When taking the mean by measuring the average of intensity that can helps directly in the classification [16].

$$\text{Mean} = \sum_{i=0}^{l-1} z_i * p(z_i) \quad (1)$$

Calculating Standard Deviation:

Standard Deviation is a measure of how spreads out numbers are [16]. To measure the dimension of cropping region (cancer region) the best function is defined as:

$$\text{STD} = \sum_{i=0}^{l-1} (z_i - m)^2 * p(z_i) \quad (2)$$

Calculating Skewness

The Skewness is a measure of symmetry or it is the lack of symmetry. A distribution or data set is symmetric if it looks the same to the left and right of the center point (cancer assumed center point). The Skewness for a normal distribution is zero and any symmetric data should have a Skewness near zero. Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness in dicate data that are skewed right [16]. To calculate the symmetry the effective function is:

$$Skewness = \sum_{i=0}^{l-1} (z_i - m)^3 * p(z_i) \quad (3)$$

Calculating Kurtosis

The Kurtosis is a measure of whether the data are peaked or flat relative to a normal Distribution. That is, data sets (cancer data set) with high kurtosis tend to have a distinct peak near the mean [16]. To make sure that data is peak or flat relative to the normal distribution in our case the best method is used:

$$Kurtosis = \sum_{i=0}^{l-1} (z_i - m)^4 * p(z_i) \quad (4)$$

Calculating Contrast

The Contrast is the difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable. In Visual Perception Of the real world, contrast determined by the difference in the color and brightness of the object and other objects within the same field of view [16]. Measuring that difference in luminance or color (ie. the infected area), that requires using the following method:

$$\text{Contrast } = \sum_{i=0}^{l-1} \sqrt{(z_i - m)^2 * p(z_i)} \quad (5)$$

Calculating Smoothness

In the region of interest it is needed to measures the elative intensity variations [16]. The effective function is:

$$R = 1 - \frac{1}{(1+\sigma 2)} \quad (6)$$

Table 1 below explain the description of six statistical features.

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

Function	Formula	Description
Mean	$\sum_{i=0}^{l-1} z_i * p(z_i)$	A measure of average intensity
Standard Deviation	$\text{STD} = \sum_{i=0}^{l-1} (z_i - m)^2 * p(z_i)$	Second moment about the mean and it is a measure of how spreads out numbers are.
Skewness	$Skewness = \sum_{i=0}^{l-1} (z_i - m)^3 * p(z_i)$	Third moment about the mean and it is a measure of symmetry
Kurtosis	$Kurtosis = \sum_{i=0}^{l-1} (z_i - m)^4 * p(z_i)$	Fourth moment about the mean
Contrast	$\text{Contrast } = \sum_{i=0}^{l-1} \sqrt{(z_i - m)^2 * p(z_i)}$	Standard deviation of pixel intensities and it is ameasure of whether the data are peaked or flat relative to
Smoothness	$R = 1 - \frac{1}{(1+\sigma 2)}$	Measures the relative intensity variations in a region

From Table 1 z_i is a random variable indicating intensity, $p(z_i)$ is the histogram of the intensity levels in a region, L is the number of possible intensity levels, σ is the standard deviation.

Phase 4: Features Selection

Data sets for analysis contain hundreds of attributes, it may be possible for a domain expert to determine the useful attributes, and this can be a difficult and time consuming task, especially when the behavior of the data is not well known. Leaving out relevant attributes or keeping irrelevant attributes may be detrimental; causing confusion for mining algorithm employed. Thus the dimensionality reduction reduces the data size by removing such attributes from it. The method called attribute subset selection is applied to reduce the data size the goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. In this research used Mutual information to feature selection.

Mutual Information

The mutual information (MI) is a measure of the amount of information that one random feature has about another feature [17].

Mutual Information algorithm provides only a score value for each feature to reflect its usefulness. In order to use these feature scoring method for subset determination, additional considerations are needed to determine the size of the subset. The feature selection method (MI) only provides a scoring and associated ranking of features, using different criteria. The size of the feature set selected by using these methods has to be estimated using some additional algorithm. The MI method result in a scoring and ranking of features, according to which chosen number of features having the highest values can be selected.

After applied Mutual Information on all features to rank and select important features; obtained excel files contained the important features.

Phase 5: Classification (Model Construction and Testing)

After cropping the region of interest and extract features of that region through different statistical functions the next step is classifying these features into different classes, here data were being classified into two classes benign and malignant. Data classification is a two-step process. First step consisting of a learning step (where a classification model is constructed). And second step is a classification step (where the model is used to predict class labels for given data).

In the first step a classifier is built describing a determined set of data classes or concepts. This is the learning step or training phase, where a classification algorithm builds the classifier by analyzing or learning from a training set made up of database tulles and their associated class labels.

In the second stage the classification model constructed previously is used to classify unknown class's data which is known as a testing. In this stage individual classifier Decision Tree will be applying.

Decision Tree

During the late 1970s and early 1980s, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomies). This work expanded on earlier work on concept learning systems, described by E. B. Hunt, J. Marin, and P. T. Stone. Quinlan later presented C4.5 (a successor of ID3), which became a benchmark to which newer supervised learning algorithms are often compared [17] In 1984, a group of statisticians (L. Breiman, J. Friedman, R. Olshen, and C. Stone) published the book Classification and Regression Trees (CART), which described the generation of binary decision trees. ID3 and CART were invented independently of one another at around the same time, yet follow a similar approach for learning decision trees from training tuples. These two cornerstone algorithms spawned a flurry of work on decision tree induction [18]. ID3, C4.5, J48 and CART adopt a greedy (ie., no backtracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow a top-down approach, which starts with a training set of tupelos and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built [19].

Phase 6: Testing and Evaluation

Commonly the evaluation measurement in classification problems are defined from matrix with the numbers of correctly and incorrectly classified for each class called confusion matrix. The confusion matrix for a binary classification problem (which has two classes – positive and negative). A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. A confusion matrix illustrates the accuracy of the solution to a classification problem. When evaluating a classifier, there are different ways of measuring its performance. For supervised learning with two possible classes, all measures of performance are based on four numbers obtained from applying the classifier to the test set. These numbers are called true positives TP, false positives FP, true negatives TN, and false negatives FN [20].

Equation (7) shows how Accuracy is computed.

This research focus only on the accuracy measures for evaluation:

$$Ac = \frac{[TN]+[TP]}{[FN]+[FP]+[TN]+[TP]} \quad (7)$$

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

4. Experiment

In this study a simple method with decision tree classifier use continues form of dataset was applied. An experimental conducted on different training-testing partitions of the dataset. It was divided into 60%, 80% training set and 40%, 20%testing set data.

5. Result and Discussion

For classification technique firstly data were being split into training dataset and testing data set. Then data were classified using decision tree classifier. This technique makes a tree to classify a set of input examples according to their class and each branch in the tree represents a decision and each node in the tree refers to a particular attribute. Edges connecting nodes were labeled with attribute values and leaf nodes give a classification that applies to the examples that were reached through that branch.

After applying different size of training and testing data and calculate the overall accuracy which is evaluated according to confusion matrix, the final results are shown below in Table 2.

Table 2. *Experiment Result.*

Training/ testing	Accuracy
60% -40%	97.04%
80%-20%	93.3%

6. Conclusion

In this research an Enhancement of Mammogram Images Classification Accuracy Using Data mining technique (decision tree classifier) for medical datasets classification,

This study emphasis of six phases starting with image Collection, pre-processing, Enhanced Feature Extraction, feature selection, Classification (Model Construction and Testing) and end with testing and evaluating. Six features extracted from each image, these moments are: mean standard deviation, skewness, kurtosis, contrast and smoothness. Features were selected according to mutual information technique. Proposed technique implemented medical images data, the data divided into two parts; training part that used 60, 80 percentage of data set. And the rest were used for testing purpose. Classification accuracy calculated according to confusion matrix. Proposed method achieved accuracy 97.04%. To achieve better performance in terms of accuracy it is recommended to use more than six features for extracting. However, future work will focus in other criteria such as classification speed and cost. In addition, the breast cancer dataset used in this study was taken from the patients' dataset who hesitate on Tumor therapy and Cancer Research Center, Shendi University, Sudan.

References

[1] BURKE, H. B., GOODMAN, P. H., ROSEN, D. B., HENSON, D. E., WEINSTEIN, J. N., HARRELL JR, F. E., MARKS, J. R., WINCHESTER, D. P. & BOSTWICK, D. G. 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79, 857-862.

[2] SUGUNA, N. & THANUSHKODI, K. 2010. An improved k-nearest neighbor classification using genetic algorithm. International Journal of Computer Science Issues, 7, 18-21.

[3] ZAıANE, O. R., ANTONIE, M.-L. & COMAN, A. 2002. Mammography classification by an association rule based classifier. MDM/KDD, 62-69.

[4] MANDELBLATT, J. S., CRONIN, K., M. & PLEVRITIS, S. K. 2009. Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. Annals of internal medicine, 151, 738-747.

[5] SMITHA, P., SHAJI, L. & MINI, G. 2011. "A Review of Medical Image Classification Techniques." International Conference on VLSI, Communication and Instrumentation. International Journal of Computer Application, 34-48.

[6] ARNING, A., AGRAWAL, R. & RAGHAVAN, P. A Linear Method for Deviation Detection in Large Databases. KDD, 1996. 972-981.

[7] RANGAYYAN, R. 2005. Chap. 7, Analysis of texture. Biomedical Image Analysis CRC Press LLC, Boca Raton, FL, 1277-1375.

[8] LASHARI, S. A. & IBRAHIM, R. 2013. A framework for medical images classification using soft set. Procedia Technology, 11, 548-556.

[9] ALI, S. & SMITH, K. A. 2006. On learning algorithm selection for classification. Applied Soft Computing, 6, 119-138.

[10] HAN, J. & KAMBER, M. 2006. Data Mining: Concepts and Techniques, 2nd edition Morgan Kaufmann Publishers. San Francisco, CA, USA.

[11] PAREEK, A. and S. M. ARORA, Breast cancer detection techniques using medical image processing. Breast cancer, 2017. 2 (3).

[12] AARTHI, R., DIVYA, K., KOMALA, N. & KAVITHA, S. Application of Feature Extraction and clustering in mammogram classification using Support Vector Machine. 2011 Third International Conference on Advanced Computing, 2011. IEEE, 62-67.

[13] PAREEK, A. & ARORA, S. M. 2017. Breast cancer detection techniques using medical image processing. Breast cancer, 2.

[14] USHA, S. & ARUMUGAM, S. 2016. Calcification Classification in Mammograms Using Decision Trees. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9, 2127-2131.

[15] IBRAHIM, A. O., AHMED, A., AZIZAH, A. H., LASHARI, S. A., ALOBEED, M. A., KASIM, S. & ISMAIL, M. A. 2018. An enhancement of multi classifiers voting method for mammogram image based on image histogram equalization. International Journal of Integrated Engineering.

[16] MOHAMMED, N. S. M. 2021. Enhancing the Mammogram image Classification using Mutual Information Feature Selection. Sudan University of Science and Technology.

[17] GANESAN, K., ACHARYA, U. R., CHUA, C. K., MIN, L. C. & ABRAHAM, T. K. 2014. Automated diagnosis of mammogram images of breast cancer using discrete wavelet transform and spherical wavelet transform features: a comparative study. Technology in cancer research & treatment, 13, 605-615.

[18] EDDAOUDI, F., REGRAGUI, F., MAHMOUDI, A. & LAMOURI, N. 2011. Masses detection using SVM classifier based on textures analysis. Applied Mathematical Sciences, 5, 367-379.

[19] KHARRAT, A., GASMI, K., MESSAOUD, M. B., BENAMRANE, N. 2010. A hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine. Leonardo journal of sciences, 17, 71-82.

[20] MOHAMED ALHAJ ALOBEED., ALI AHMED, ASHRAF IBRAHIM. Multi-classifier method based on voting technique for mammogram image classification, Journal of Software Engineering & Intelligent Systems issn 2518-8739. 31st December 2017, Volume 2, Issue 3.