Available online at www.sciencedirect.com

ScienceDirect

Journal of Electrical Systems and Information Technology 3 (2016) 348–359

www.elsevier.com/locate/jesit

# Fully automatic classification of breast cancer microarray images

Nastaran Dehghan Khalilabad [a], Hamid Hassanpour [a,*], Mohammad Reza Abbaszadegan [b]

[a] *Shahrood University of Technology, Shahrood, Iran*
[b] *Mashhad University of Medical Sciences, Mashhad, Iran*

Received 22 October 2015; received in revised form 12 June 2016; accepted 17 June 2016
Available online 11 August 2016

## Abstract

A microarray image is used as an accurate method for diagnosis of cancerous diseases. The aim of this research is to provide an approach for detection of breast cancer type. First, raw data is extracted from microarray images. Determining the exact location of each gene is carried out using image processing techniques. Then, by the sum of the pixels associated with each gene, the amount of “genes expression” is extracted as raw data. To identify more effective genes, information gain method on the set of raw data is used. Finally, the type of cancer can be recognized via analyzing the obtained data using a decision tree. The proposed approach has an accuracy of 95.23% in diagnosing the breast cancer types.

*Keywords:* Microarray; Gene expression; Information gain; Breast cancer; Decision tree

## 1. Introduction

According to the World Health Organization (WHO), breast cancer is the top cancer among women in both developed and developing countries. Early detection and survival are important issues to control breast cancer (Shulman et al., 2010).

One of the most important and accurate methodologies to diagnose the disease is DNA analysis of individuals. DNA microarrays are an important technology to study gene expression (Cano et al., 2009). A microarray is a glass slide on which single-stranded DNA molecules are attached at fixed locations called spots. There can be thousands of spots on a single microarray chip, each spot on a microarray chip contains multiple identical strands of DNA which identify one gene.

The microarray images contain multiple blocks (referred to sub-grid) and these blocks consist of many spots, situated in rows and columns. Each spot is an area in the image which represents the level of the hybridization between a single probe and the samples.

[*] Corresponding author.
*E-mail address:* h.hassanpour@shahroodut.ac.ir (H. Hassanpour).
Peer review under the responsibility of Electronics Research Institute (ERI).

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jesit.2016.06.001

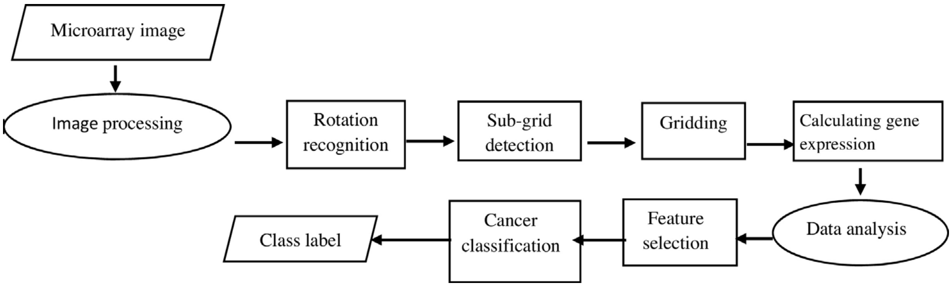

Fig. 1. Using microarray images and image processing techniques for breast cancer classification.

Hybridization experiments with two samples consist of the following steps: firstly, the total mRNA (Messenger RNA) from cells in two different conditions (for example, healthy and cancerous cells) is dyed with two different fluorescent labels (commonly cy3 and cy5). Secondly, the labeled samples are washed over the microarray. These labeled gene products are hybridized to their complementary sequences in the spots (Stekel, 2003). The intensity of each spot signifies the degree of hybridization of the sample to a known gene, thereby indicating the expression level of the particular gene.

Gene expression profiling by microarray method has been came out for classification and diagnostic prediction of tumor. Images are transformed into gene expression matrices in which rows correspond to genes, and the columns represent samples or trial conditions (Gunavathi and Premalatha, 2014).

Key steps in microarray analysis include image processing and data analysis (Fig. 1). Image processing is an important step in microarray experiments which increases accuracy of data microarray analyses including clustering.

Within the image processing step, the quantification of gene expression levels from microarray images are performed in four steps, including rotation recognition, sub-grid detection, gridding and calculating gene expression. Sub-grid detection involves partitioning the blocks into distinct cells in the image, then assigning coordinates to each blocks and in each sub-grid various spots are isolated in the image. Gene expression data is calculated by using the log ratio between the two intensities of each dye. Considering that, error at an early stage can lead to incorrect results, hence, gridding is the most important step for subsequent tasks.

Within the data analysis steps, the two steps for predicting breast cancer include the following: feature selection and tumors classification methods. Feature selection or gene selection is the process by which great amounts of data are analyzed and summarized into useful information. This information can be used to cut costs. The general feature selection and classification application tools are used for classification tasks.

In this paper, a most widely used gene selection method, information gain (IG), is adopted for extracting useful knowledge from microarray data. IG method can quickly reject a large number of non-critical noise and irrelevant genes, hence, refine search area of the optimal subset of genes (Cho and Won, 2003; Wang et al., 2006). Also in our proposed system, decision trees are developed to diagnose two subclass tumor.

The structure of this paper is organized as follows: in Section 2, the gridding and feature selection and classification methods are reviewed. Section 3 contains a short description of the breast cancer data set. Methods for image processing and data analysis are presented in Section 4. Section 5 presents the conclusion of this research.

## 2. Related work

The main goal of this section is to provide a brief review on different basic steps of the works related to gridding, gene selection and tumor classification using microarray image and microarray gene expression data.

It is essential to efficiently analyze DNA microarray data because the number of features is much larger than the number of samples. Many image processing, machine learning and data mining methods have been employed to overcome this difficulty.

### 2.1. Gridding

Gridding is the most important task in the initial stage, if done correctly, substantially improves the efficiency of subsequent stages. During the recent years, there have been methods and software packages available dealing with one or a few problems, the most important is that they all require several parameters to be set by the user; for examples software such as ScanAlyze, GenePix Pro6 and ImaGene (Bariamis et al., 2010; Fouad et al., 2014).

Angulo and Serra (2003) proposed a non-supervised set of algorithms for a fast and accurate spot data extraction from DNA microarrays using morphological operators. In this method, there are drawbacks that have to be resolved before fully automatic gridding can take place. As an example, this model requires that grid rows and columns are strictly aligned with the $x$- and $y$-axes.

An approach be utilized in Katzer et al. (2003) for automatic grid segmentation of the raw fluorescence microarray images by Markov random field (MRF) techniques. This method is not suitable for microarray gridding since a complete search is not feasible for the optimal MRF configuration.

Zacharia and Maroulis (2008) utilized a support vector machine (SVM) to grid microarray images. The method uses a set of soft-margin SVMs to forecast the lines of DNA microarray grid by maximizing the margin between lines and spots.

Microarray image gridding method based on image projection transformation and power spectral analysis is discussed in Feng et al. (2015). Firstly in this paper is transformed 2D microarray image into vertical and horizontal 1D projection sequences, secondly utilized signal processing methods of low pass filtering, zero mean, FFT, and power spectral estimation by periodogram method to get spots array row and column span information, and finally realized the microarray image gridding according to the local maxima and span information of spots array.

### 2.2. *Gene selection*

The major limitation of the gene expression data is its high dimension which contains thousands of genes and a very few samples (Gunavathi and Premalatha, 2014). A number of researchers have turned to data mining technologies and machine learning approaches for predicting breast cancer.

In 1995 (Wolberg et al., 1995) data mining and machine learning approaches were embedded into a computer-aided system for diagnosing breast cancer. Recent researches (Wong and Hsu, 2008) have shown that a small number of genes are sufficient for accurate diagnosis of the most diseases, even though the number of genes vary greatly between different diseases. Thus gene selection plays a major role in the proposed system (Horng et al., 2009).

Some existing methods proposed a hybrid system to obtain a small set of highly discriminative genes; for example the work performed by Min Xu (Xu and Setiono, 2003) have designed hybrid method from the univariate maximum likelihood method (LIK) and the multivariate recursive feature elimination (RFE) method.

A most widely used gene selection method, information gain (Cho and Won, 2003; Wang et al., 2006), is adopted for the purpose of this work. Information gain measures the goodness of gene using the presence and absence within the corresponding class.

### 2.3. *Classification*

Classification is the task of finding the common properties among a set of objects in a database and classifying them into different classes (Chen et al., 1996). In the literature, statistical approaches like boosting, and self-organizing map (Golub et al., 1999), K-nearest-neighbor classification (Li et al., 2001), discrimination methods (Dudoit et al., 2002), decision tree, multi-layer perceptron (Khan et al., 2001; Xu et al., 2002), least square and logistic regression (Fort and Lambert-Lacroix, 2005), Naive Bayes approach (Fan et al., 2009) and active learning using fuzzy K-nearest-neighbor for cancer classification (Halder et al., 2015) were used to generate the classifier model for gene expression data. Hybrid methods have been recently tested on microarray data (Lee and Leu, 2011) obtaining high classification accuracies. For example Mahmoud and Basma (2014) the proposed approach is integrated with two machine learning classifiers; K-nearest neighbor (KNN) and support vector machine (SVM); to classify microarray gene expression.

Decision trees have been considered as the classifier in this paper. The performance of the proposed approach is evaluated using gene expression data set viz., Type 2 breast cancer (Kao et al., 2009; Bergamaschi et al., 2006).

## 3. Data set

The data set used for evaluation of the proposed method consists of 71 DNA microarray images related to breast cancer, from the Stanford Microarray Database (Stanford, 2015). The images are stored in TIFF files with 16-bit gray

Table 1
Accuracy of the proposed method on the rotated images.

| Angle | Accuracy rotation adjustment | |
|---|---|---|
| | Image 51769 | Image 51865 |
| 1° | 99% | 100% |
| 0.75° | 99% | 98.67% |
| 0.5° | 99% | 98% |
| 0.25° | 98% | 98% |
| −1° | 100% | 99% |
| −0.75° | 99.3% | 98% |
| −0.5° | 99% | 100% |
| −0.25° | 98.6% | 98% |

level depth. Fluorescent separated image for each sample is divided into two channels, ch1 (cy5) and ch2 (cy3). The images include 48 blocks of about 870 spots, a total of 41,760 spots in the image.

There are 22 samples related to primary breast tumors which cells were grown to 70–80% confluence, then harvested for total RNA and genomic (Kao et al., 2009) DNA and the remaining related to breast tumor specimens were derived from 49 patients with locally advanced (T3/T4 and/or N2) breast cancer receiving either doxorubicin or fluorouracil-mitomycin based neoadjuvant chemotherapy (Bergamaschi et al., 2006).

The breast cancer data was already divided into training and testing sets. There are 50 training samples and 21 testing samples. The 21 testing set contains five class-1 and sixteen class-2 samples. Also, the 50 training set contains seventeen class-1 and thirty three class-2 samples.

## 4. Systems and methods

### 4.1. Rotation adjustment

A microarray image is often skewed during the scanning process, most likely because the chip microarray is not correctly placed inside the microarray scanner. Rotations of the images are seen in two different direction clockwise, with respect to the x axis. Radon transform is applied to find angles of rotation in the image (Rueda and Rezaeian, 2011). The Radon transform $g(\rho, \tau)$, is the integral of a continuous 2D function $f(x,y)$ over a collection of slanted lines (Smith, 1995):

$$g(\rho, \tau) = \int_{-\infty}^{+\infty} f(x, \rho x + \tau) dx \tag{1}$$

The Radon transform is linear – the Radon transform of a weighted sum of functions is the same weighted sum of the individually Radon transformed functions. This is an important property which can be used to approximate the transform by means of a computer program. Also, the Radon transform of a rotated, scaled or translated image can be obtained knowing the Radon transform of the original image and the affine transformation parameters. Fig. 2 shows the detection of rotated angles 1.5° and −1.5°, and their corrections.

Although the dataset only includes microarray images with rotation to a few degrees. Therefore, to test the effect of the Radon transform, the first two images are selected with skew angle 0°. Then, we rotated images with 0.25, 0.5, 0.75 and 1 degrees in both clockwise and counter-clockwise directions. Table 1 shows the accuracy of the proposed method on two of the rotated images. This shows the effectiveness and robustness of the proposed method in significantly rotated images. The algorithm accuracy is calculated as shown in Eq. (2):

$$\text{Accuracy} = \frac{|\theta_{real} - \theta_{est}|}{\theta_{real}} \times 100 \tag{2}$$

### 4.2. Sub-grid detection

Image thresholding is one of the most widely-used techniques that has many applications in image processing, including segmentation, classification and object recognition. The proposed sub-grid method refers to the process of

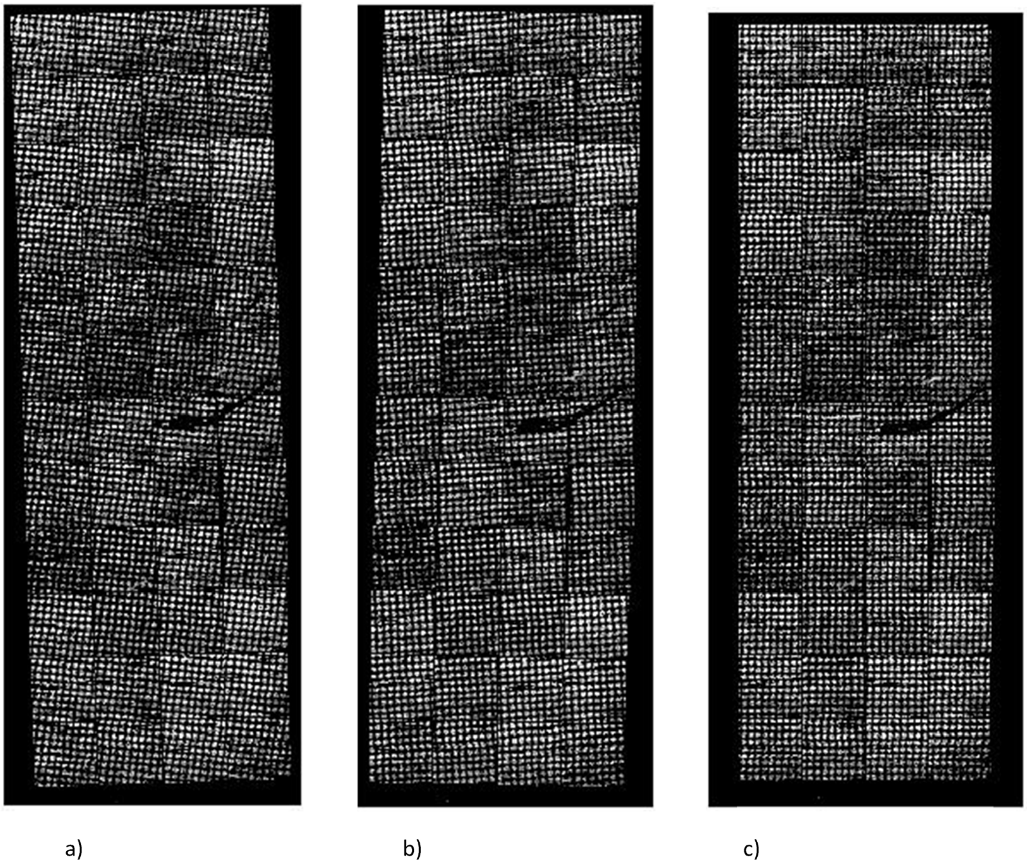

Fig. 2. (a) Detection of the rotated angle = 1.5°. (b) Detection of the rotated angle = −1.5°. (c) Correcting result.

locating each sub-array within a microarray image (Rueda and Rezaeian, 2011). The global parameters required for accurately locating sub-arrays are width and height of each sub-array as well as spacing between them.

According to sub-grid, we computed the row or column means of pixel intensities, and a one-dimensional function is obtained. The graph one-dimensional is plotted. Its domain reflects the position of the rows/columns of pixels. In this work, that function is considered as a histogram in which each bin represents one column (or row). Finally, the space between the blocks is calculated by the local minimum. This method can deal with various types of noisy images. For example, images with black regions mean that some of the spots have been lost during the scanning. Fig. 3 shows the steps of gridding a microarray image containing 12 × 4 sub-grids, along with the corresponding row or column sums.

### 4.3. Gridding

Gridding is an important step which indicates coordinates for the individual spots. The proposed gridding method tries initially to remove the large flare noise in a microarray image (Fouad et al., 2013).

Microarray image gridding is complex, as it contains noise and features low intensity and weak contrast. Therefore, we proposed the preprocessing step by applying histogram equalization function to produce the high contrast between the foreground (spots) and the background. But this method may increase the contrast of background noise, so we resorted to Wiener filtering to eliminate it (Acharya and Ray, 2005).

To remove the large flare noise at first, the erosion operator with a structuring element is applied to remove foreground spots. A less noisy image is obtained by subtracting resulted background from the original image. This is mainly due to removal of large flare noise. After that, morphological opening with a structuring element is applied to remove small spikes in the image.

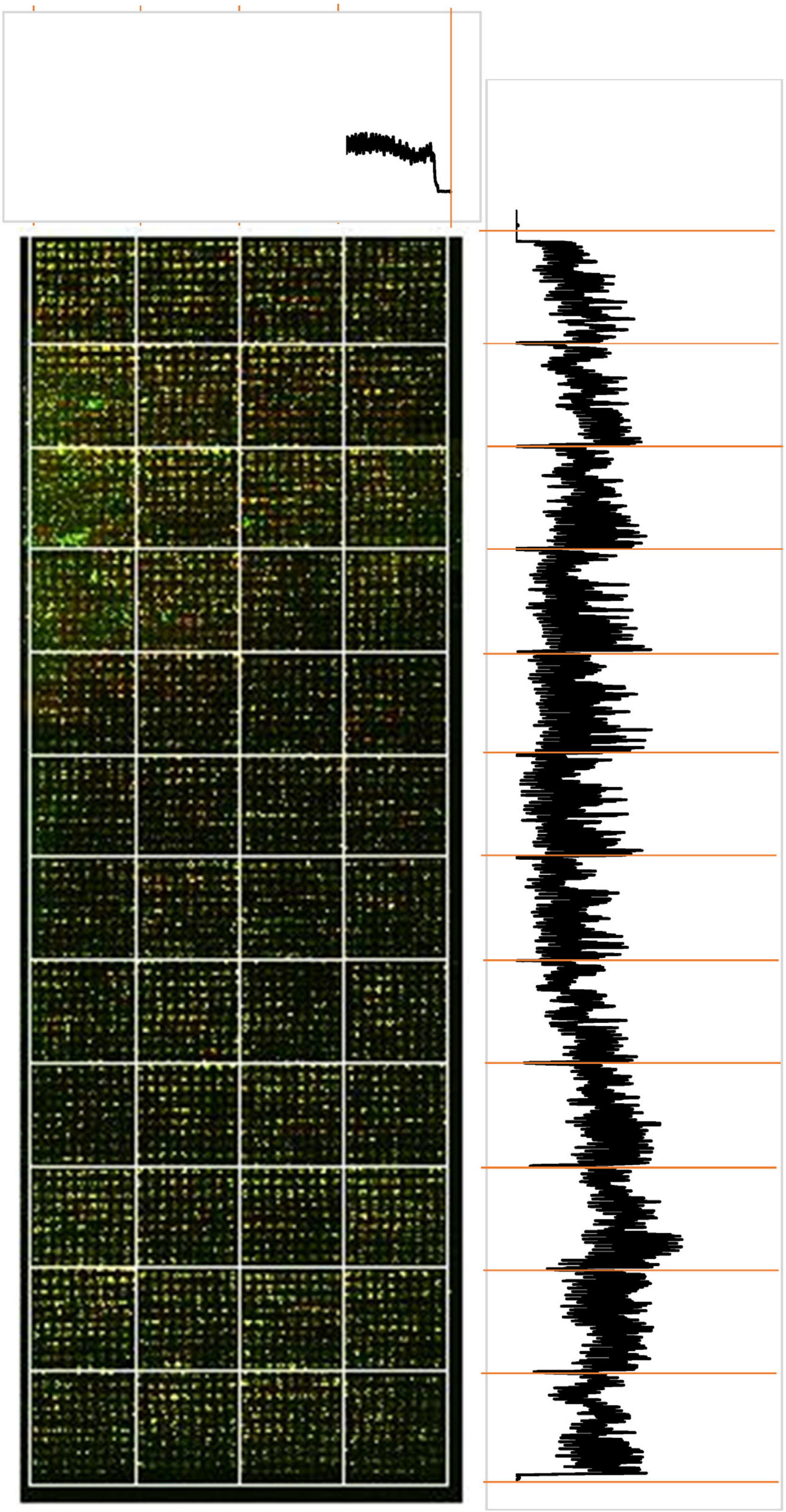

Fig. 3. Sub-grid detection in a microarray image from the breast cancer dataset.

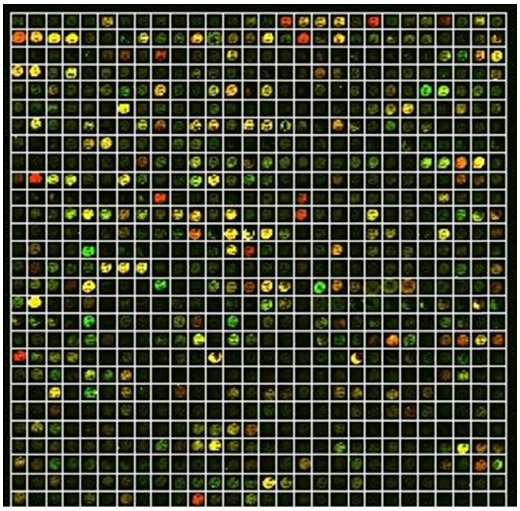

Fig. 4. Spot detection in a sub-grid with removing flare noise.

Table 2
Accuracy of the proposed method for gridding image.

| Image | Accuracy |
|---|---|
| Image 51786 | 100% |
| Image 51771 | 98.83% |
| Image 54846 | 99.76% |

We searched a regular grid of spots, so we started by considering the mean intensity for each column of the image. After that, we computed the mean intensity for each column of the image (Labib et al., 2012). To obtain the mean horizontal intensity profile $MH(y)$ of the image $f(x, y)$ (dimensions $x$ and $y$), the formula is defined as follows:

$$MH(y) = \frac{1}{x} * \sum_{x=0}^{x=X-1} f(x, y) \quad (3)$$

We can use the spacing estimate to help designing a filter to remove the background noise from the intensity profile. Hence, autocorrelation is used to enhance the self-similarity of the profile. The result promotes estimation of spot spacing. Then we extract the centroids of the peaks. These correspond to the horizontal centers of the spots. The midpoints between adjacent peaks provide grid point locations. These parameters are used to determine the locations of the vertical grids and draw them on the image. We simply transpose the image and repeat all the steps used above to get the horizontal grids on the image. Microarray spot gridding is shown in Fig. 4.

The proposed gridding method was implemented on noisy microarray images of breast cancer. Results obtained in this work are shown in Table 2.

The accuracy of the applied gridding method on a specified input image can be calculated as follows:

$$\text{Accuracy} = \left(\frac{N(\text{corret spot})}{N(\text{total spot})}\right) * 100 \quad (4)$$

### 4.4. Measuring gene expression

There are major types of application of DNA microarrays in medicine. The first involves finding differences in expression levels between predefined groups of samples. This is called a “class comparison” experiment. Once the microarrays have been hybridized, the resulting images are used to generate a dataset. This dataset needs to be “preprocessed” prior to the analysis and interpretation of the results. Preprocessing is a step that extracts or enhances

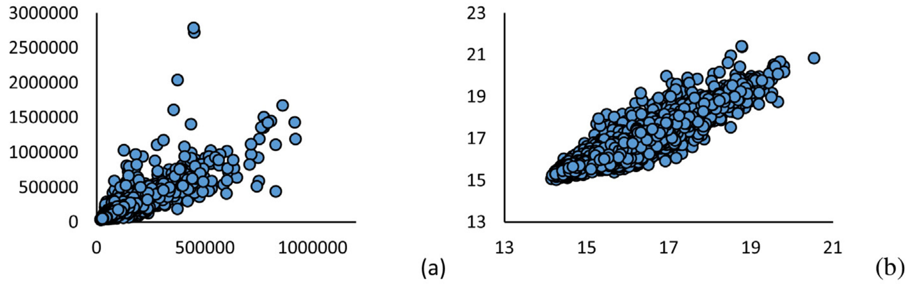

Fig. 5. Display of expression profiling data obtained from one cDNA array (two channel technology). (a) Spot raw intensity. (b) Spots log transformation in two channels microarray experiments using Cy3 (green) and Cy5 (red).

meaningful data characteristics and prepares the dataset for the application of data analysis methods. A typical example of preprocessing is taken the logarithm of the raw intensity values (Tarca et al., 2006).

The total amount of hybridization for a spot is proportional to the total fluorescence at the spot in two channel microarray experiments using Cy3 (green) and Cy5 (red). Therefore, spot intensity is equal to sum of pixel intensities. Then, the data is generally log-transformed. The log transformation improves the characteristics of the data distribution and allows the use of classical parametric statistics for analysis as shown in Fig. 5. Hence, reasons for working with log-transformed intensities and ratios include: (1) spreading features more evenly across intensity range, (2) making variability more constant across intensity range, and (3) resulting in close to normal distribution of intensities and experimental errors.

With two-channel arrays, the intensity values of the two competing samples are expressed as ratios and then log-transformed. The log ratio between the two intensities of each dye is used as the gene expression data (Labib et al., 2012):

$$\text{Gene expression} = \log 2\left(\frac{\text{int}\,(cy5)}{\text{int}\,(cy3)}\right) \tag{5}$$

where *int*(*cy5*) and *int*(*cy3*) are the intensities of the red and green colors (channel 1, 2 in image).

Since at least hundreds of genes are put on the DNA microarray, it is so helpful that we can investigate the genome-wide information in short time.

## 5. Gene selection

As previously mentioned, the number of features is much larger than the number of samples. Most genes are not related to the performance of the classification. In fact, a large collection of gene expression features will have higher computational cost and slower learning process. There are studies suggesting that only a few genes are sufficient. Thus, it is crucial to recognize whether a small number of genes can suffice for good classification (Li and Yang, 2002).

Information gain technique uses the concept of Shannon entropy. Given entropy $E$ as a measure of impurity in a set of training samples, it is possible to define a measure of the effectiveness of a feature/gene in classifying the training data. This measure is simply the expected reduction in entropy caused by partitioning the data according to this feature, so-called information gain (Wang et al., 2006).

Assume a given set of microarray gene expression data $M$, the information gain of a gene $i$ is defined as:

$$IG(M,\ i) = E(M) - \sum_{v \in V(i)} \frac{M_v}{M} E(M_v), \tag{6}$$

where $V(i)$ is the set of all possible values of feature $i$, $M_v$ is the subset of $M$ for which feature $i$ has value $v$, $E(M)$ is the entropy of the entire set, and $E(M_v)$ is the entropy of the subset $M_v$. The entropy function $E$ is defined by

$$E = \sum_{J=1}^{C} - \frac{|Cj|}{\left|\sum c\right|} \log 2 \frac{|Cj|}{\left|\sum c\right|}, \tag{7}$$

where $|Cj|$ is the number of samples in class $Cj$.

Table 3
Genes selected through IG for Breast cancer.

| Gen-no | Gene ID | IG value | Rank | Gene name |
|---|---|---|---|---|
| 30216 | ASCL1::PAH | 0.753 | 1 | achaete-scute complex homolog 1 (Drosophila)::Phenylalanine hydroxylase |
| 697 | SETBP1::SETBP1 | 0.722 | 2 | SET binding protein 1::SET binding protein 1 |
| 28275 | PSCA::PSCA | 0.707 | 3 | prostate stem cell antigen::Prostate stem cell antigen |
| 8737 | PRKAR2A::PRKAR2A | 0.707 | 4 | Protein kinase, cAMP-dependent, regulatory, type II, alpha::protein kinase, cAMP-dependent, regulatory, type II, alpha |
| 1369 | ELF4::ELF4 | 0.696 | 5 | E74-like factor 4 (ets domain transcription factor)::E74-like factor 4 (ets domain transcription factor) |
| 8728 | – | 0.688 | 6 | Transcribed locus |
| 21867 | MBP::MBP | 0.682 | 7 | Myelin basic protein::myelin basic protein |
| 20044 | UROS:: UROS | 0.682 | 8 | Uroporphyrinogen III synthase::uroporphyrinogen III synthase |
| 9679 | – | 0.682 | 9 | Full length insert cDNA clone YF46C08 |
| 18945 | RNF138::RNF138 | 0.679 | 10 | Ring finger protein 138::ring finger protein 138 |

The entropy is supposed to give the information required in bits, and is traditionally used to deal with boolean-valued features (hot/cold, true/false, etc.). Fortunately, this method can be extended to handle the data with continuous valued features, for example, microarray gene expression data. This method shows that only a few genes are having important information about the disease and the remaining have very low amount of information. Regarding the study suggesting that only few genes are sufficient for understanding their biological relationship with the target diseases, hundred genes with higher IG value are selected as informative genes. Table 3 gives the detail of the genes selected using IG for breast cancer data set (only the major ten genes were displayed).

## 5.1. Classification

In this paper, the decision tree algorithm is used to classify the data set. The decision tree (Han and Kamber, 2001) is one of the classifying methods, which classifies the labeled trained data into a tree or rules.

One of the main advantages of decision trees is the ability to generate understandable knowledge structures, i.e., hierarchical trees or sets of rules, a low computational cost when the model is being applied to predict or classify new cases, the ability to handle symbolic and numeric input variables, provision of a clear indication of which attributes are most important for prediction or classification (He and Hui, 2009).

In this paper, j48 tree is used. This tree is a slightly modified version of C4.5. This algorithm is a successor to ID3 developed by Quinlan (1986). The default criteria of choosing splitting attributes in C4.5 is information gain ratio. This criteria is used to employ the entropy measure as an impurity measure. In tree C4.5 the attribute values is divided into two groups to handle continuous attributes based on the selected threshold such that all the values above the threshold are considered as one child and the remaining as another child. C4.5 uses Gain Ratio as an attribute selection measure which removes the biasness of information gain when there are many outcome values of a feature. Gain ratio of each feature is calculated and thereafter the root node will be the attribute whose gain ratio is maximum. Removing unnecessary branches using pessimistic pruning in the decision tree is done to improve the accuracy of classification. C4.5 classification tree can be described into two parts, learning using training data and classification using test data.

The structure is in the form of statements including <if>. . .<then>. . .. One advantage of using logical statements is that it is convenient for coding into other programs and therefore the code can be easily reused.

The three concepts of information gain, entropy and gain ratio are described in the following (Patidar et al., 2015; Al Snousy et al., 2011).

Let $c$ denotes the number of classes, and $p(S, j)$ the proportion of instances in $S$ that are assigned to $j$-th class. Therefore, the entropy is calculated as follows:

$$\text{Entropy}\,(S) = -\sum_{j=1}^{c} p(S, j) \times \log p(S, j) \tag{8}$$

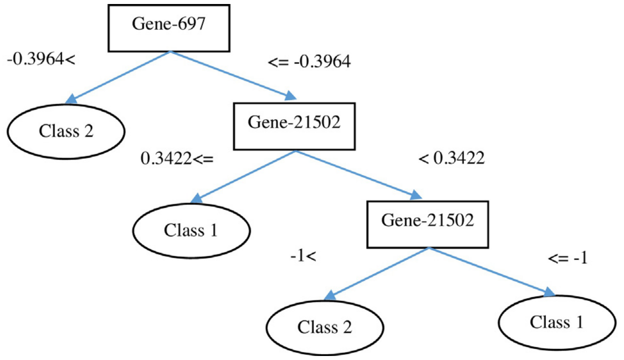

Fig. 6. The decision tree for breast cancer diagnosis.

Accordingly, the information gain by a training dataset $T$ is defined as:

$$\text{Gian}\,(S,T) = \text{Entropy} - \sum\nolimits_{v\in Value(T_s)} \frac{|T_{s,v}|}{|T_s|}\text{Entropy}\,(Sv), \tag{9}$$

where $Value(T_s)$ is the set of values of $S$ in $T$, $T_s$ is the subsets of $T$ induced by $S$ and $T_{s,v}$ is the subset of $T$ in which attribute $S$ has a value of $v$.

Therefore, the information gain ratio of attribute $S$ is defined as:

$$GainRatio\,(S,T) = \frac{Gain\,(S,T)}{SplitInfo\,(S,T)} \tag{10}$$

where $SplitInfo(S, T)$ is calculated as:

$$SplitInfo\,(S,T) = -\sum_{(v=Value(T_s))}\left(\frac{|T_{s,v}|}{|T_s|}\times\log\frac{|T_{s,v}|}{|T_s|}\right) \tag{11}$$

In the decision tree method, a great deal of statistical information is supplied, including true positives ($TP$), false positives ($FP$), true negatives ($TN$), false negatives together ($FN$). With this amount, the classification accuracy is calculated using the Eq. (12):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

Fig. 6 shows the decision tree for subclass diagnosis of breast cancer. The decision tree method with three genes gives 100% classification accuracy in the diagnosis phase of breast cancer type.

## 6. Conclusion

In this paper, we presented a system for extracting data from image microarray and classifying raw data related to breast cancer, which consists generally of six steps. The first step aims to rectify any detected rotations within the images. Next, image thresholding technique was used to detect the correct number of individual sub-grids containing 12 × 4 sub-grids. In the third step, large flare noise using morphological operators were removed initially. Then, the mean row or column of intensities computed and autocorrelation was implemented to enhance the self-similarity of the profile. Lastly, separated lines were drawn for each of the spots. The fourth step explains how gene expression values were calculated by using log ratio between the two intensities of each dye as the gene expression data. Within the fifth step, information gain method was employed to reduce the number of genes when small number of genes was sufficient for accurate diagnosis of most cancers. As a next step, decision tree approach was implemented to classify breast cancer microarray data.

Overall, the advantage of the proposed system is that gridding performance has been successful on the images when following conditions are met: low intensity, poor quality spots, noise and artifacts, as well as rotation. The effects of noise artifacts are diminished by morphology operator. Furthermore, the generalization performance of this method allows determination of the grid lines in the presence of weakly expressed spots. While in more than 98% of the cases, the spots reside completely inside their respective grid cells, merely in a few images with noticeable noise and defects which results accuracy lower than 98%.

In the proposed system, image's raw data stored in an Excel file, and each row and column represented genes and samples, respectively. In the data analysis step, the effectiveness of the proposed system has been demonstrated using breast cancer image data sets. System with three genes and 95.23% classification accuracy was evolved.

The proposed method leads to a high classification accuracy system for breast cancer data set. This approach can be used as a tool to overcome microarray gene expression data classification limitations.

One of the most important advantages of this method is to extract data from images, while the previous methods have been implemented using published dataset to classify various types of microarray data.

## References

Acharya, T., Ray, A.K., 2005. Image Processing: Principles and Applications. John Wiley and Sons (Chapter 6).

Al Snousy, M.B., El-Deeb, H.M., Badran, K., Al Khlil, I.A., 2011. Suite of decision tree-based classification algorithms on cancer gene expression data. Egypt. Inform. J. 12 (2), 73–82.

Angulo, J., Serra, J., 2003. Automatic analysis of DNA microarray images using mathematical morphology. Bioinformatics 19 (5), 553–562.

Bariamis, D., Maroulis, D., Iakovidis, D.K., 2010. Unsupervised SVM-based gridding for DNA microarray images. Comput. Med. Imaging Graph. 34 (6), 418–425.

Bergamaschi, A., Kim, Y.H., Wang, P., Sørlie, T., Hernandez-Boussard, T., Lonning, P.E., Tibshirani, Børresen-Dale, A.L., Pollack, J.R., 2006. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer Genes. Chromosomes Cancer 45 (11), 1033–1040.

Cano, C., Garcia, F., Lopez, F.J., Blanco, A., 2009. Intelligent system for the analysis of microarray data using principal components and estimation of distribution algorithms. Expert Syst. Appl. 36 (3), 4654–4663.

Chen, M.S., Han, J., Yu, P.S., 1996. Data mining: an overview from a database perspective. IEEE Trans. Knowl. Data Eng. 8 (6), 866–883.

Cho, S.B., Won, H.H., 2003. Machine learning in DNA microarray analysis for cancer classification. In: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics, Adelaide, Australia, pp. 189–198.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. 97 (457), 77–87.

Fan, L., Poh, K.L., Zhou, P., 2009. A sequential feature extraction approach for naive Bayes classification of microarray data. Expert Syst. Appl. 36 (6), 9919–9923.

Feng, Y., Song, K., Liu, J., 2015. A microarray image gridding method based on projection transformation and power spectral analysis. In: 2015 International Symposium on Computers and Informatics, Beijing, China, pp. 44–51.

Fort, G., Lambert-Lacroix, S., 2005. Classification using partial least squares with penalized logistic regression. Bioinformatics 21 (7), 1104–1111.

Fouad, I.A., Mabrouk, M.S., Sharawy, A.A., 2013. A new method to grid noisy cDNA microarray images utilizing denoising techniques. Int. J. Comput. Appl. 63 (9), 36–44.

Fouad, I., Mabrouk, M., Sharawy, A., 2014. Automaticand accurate segmentation of gridded cDNA microarray images using different methods. Adv. Comput. 4 (2), 41–54.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286 (5439), 531–537.

Gunavathi, C., Premalatha, K., 2014. Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification. Int. J. Comput. Electr. Autom. Control Inform. Eng. 8 (8), 1490–1497.

Halder, A., Dey, S., Kumar, A., 2015. Active learning using fuzzy k-NN for cancer classification from microarray gene expression data. In: Advances in Communication and Computing, Springer, India, pp. 103–113.

Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Press (Chapter 9).

He, Y., Hui, S.C., 2009. Exploring ant-based algorithms for gene expression data analysis. Artif. Intell. Med. 47 (2), 105–119.

Horng, J.T., Wu, L.C., Liu, B.J., Kuo, J.L., Kuo, W.H., Zhang, J.J., 2009. An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Syst. Appl. 36 (5), 9072–9081.

Kao, J., Salari, K., Bocanegra, M., Choi, Y.L., Girard, L., Gandhi, J., Kwei, K.L., Hernandez-Boussard, T., Wang, P., Gazdar, A.F., Minna, J.D., Pollack, J.R., 2009. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. PLoS One 4 (7).

Katzer, M., Kummert, F., Sagerer, G., 2003. A Markov random field model of microarray gridding. In: 18th ACM Symposium on Applied Computing, pp. 72–77.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7 (6), 673–679.

Labib, F.E.Z., Fouad, I., Mabrouk, M., Sharawy, A., 2012. An efficient fully automated method for gridding microarray images. Am. J. Biomed. Eng. 2 (3), 115–119.

Lee, C.P., Leu, Y., 2011. A novel hybrid feature selection method for microarray data analysis. Appl. Soft Comput. 11 (1), 208–213.

Li, W., Yang, Y., 2002. How many genes are needed for a discriminant microarray data analysis. In: Methods of Microarray Data Analysis. Springer US, pp. 137–149.

Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G., 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17 (12), 1131–1142.

Mahmoud, A.M., Basma, A.M., 2014. A hybrid reduction approach for enhancing cancer classification of microarray data. Int. J. Adv. Res. Artif. Intell. 3 (10).

Patidar, P., Dangra, J., Rawar, M.K., 2015. Decision Tree C4.5 algorithm and its enhanced approach for Educational Data Mining. Int. J. Futuristic Trends Eng. Technol. 2 (2), 14–24.

Quinlan, J.R., 1986. Introduction of decision trees. J. Mach. Learn. 1 (1), 81–106.

Rueda, L., Rezaeian, I., 2011. A fully automatic gridding method for cDNA microarray images. BMC Bioinform. 12 (1), 113.

Shulman, L.N., Willett, W., Sievers, A., Knaul, F.M., 2010. Breast cancer in developing countries: opportunities for improved survival. J. Oncol. 2010, 1–6.

Smith, R., 1995. A simple and efficient skew detection algorithm via text row accumulation. In: Proceedings of the Third International Conference on Document Analysis and Recognition, IEEE, vol. 2, Montreal, Canada, pp. 1145–1148.

Stanford Microarray Database, http://smd.stanford.edu/ (last accessed December 2015).

Stekel, D., 2003. Microarray Bioinformatics. Cambridge University Press.

Tarca, A.L., Romero, R., Draghici, S., 2006. Analysis of microarray experiments of gene expression profiling. Am. J. Obstet. Gynecol. 195 (2), 373–388.

Wang, Z., Palade, V., Xu, Y., 2006. Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. In: International Symposium on Evolving Fuzzy Systems, IEEE Computational Intelligence Society, Ambleside, UK, pp. 241–246.

Wolberg, W.H., Street, W.N., Mangasarian, O.L., 1995. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Anal. Quant. Cytol. Histol. 17 (2), 77–87.

Wong, T.T., Hsu, C.H., 2008. Two-stage classification methods for microarray data. Expert Syst. Appl. 34 (1), 375–383.

Xu, M., Setiono, R., 2003. Gene selection for cancer classification using a hybrid of univariate and multivariate feature selection methods. Appl. Genomics Proteom. 2, 79–91.

Xu, Y., Selaru, F.M., Yin, J., Zou, T.T., Shustova, V., Mori, Y., Meltzer, S.J., 2002. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Res. 62 (12), 3493–3497.

Zacharia, E., Maroulis, D., 2008. An original genetic approach to the fully automatic gridding of microarray images. IEEE Trans. Med. Imaging 27 (6), 805–813.