Breast Cancer Classification: Comparative Performance Analysis of Image Shape-Based Features and Microarray Gene Expression Data

Ahmed Fawzi Otoom[1], Emad E. Abdallah[2] and Maen Hammad [3]

[1,2,3]*Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology*
The Hashemite University, Zarqa, Jordan
{[1]aotoom,[2]emad,[3]mhammad}@hu.edu.jo,

Abstract

Recently, there has been greater attention to the use of classifier systems in medical diagnosis. Medical diagnostic tools provide automated procedures for objective decisions by making use of quantitative measures and machine learning techniques. These tools are effective and helpful for medical experts to diagnose diseases. One of such diseases is breast cancer which is the second largest cause of cancer deaths among women. To build an intelligent tool, it is very important to have an effective set of features. Two types of feature sets have been commonly implemented for the purpose of breast cancer diagnosis: image shape-based features and microarray gene expression data. Both types of feature sets have been widely implemented; however, there has been no work that directly compared the classification performance of these two feature sets. In this paper, we intensively review related works that used both types of feature sets and we also review the implemented machine learning algorithms. Moreover, we run extensive experiments to compare the classification performance of the aforementioned feature sets. Our results show that the image shape-based features are more discriminative for breast cancer classification when tested with ten-fold cross validation. To check the robustness of the best performing feature set, we further examine it with five-fold cross validation and with a variety of generative classification algorithms.

Keywords: Breast cancer, tumor image shape-based features, microarray gene expression data, generative classifiers, discriminative classifiers, machine learning.

1. Introduction and Related Work

Cancer is a general name for a group of more than 100 diseases. It starts when cells in a part of the body start to grow out of control. A widely spread type of cancer is breast cancer. According to the American Cancer Society, in 2013, around 39,620 women were expected to die from it, in the US alone, coming in the second place after lung cancer for cancer deaths in women [5]. Breast cancer is a *malignant* tumor that starts in the cells of the breast. Malignant tumor is a group of cancer cells that can invade surrounding tissues or can spread to distant areas of body. It is important to note that not all tumors are cancer. Tumors that are not cancer are called *benign* [5]. Early detection of breast cancer can save thousands of lives each year and can help in treating it successfully. This was the major motive for us to study intelligent systems that can provide early diagnosis of the disease. Recently, there has been growing attention to the areas of medical informatics and machine learning. A lot of research is conducted towards the development of diagnostic tools that are designed to support medical experts (E.g. [2], [9-11], [13-14], [18-19]). Machine learning techniques are used to provide second view in analysis of medical data with the incorporation of knowledge-based approaches [11]. It can afford help for medical experts to examine medical data in a shorter time. Moreover, it can provide early diagnosis of medical diseases and prompt an early action to be taken to save

patients' lives. Early detection of the disease is an important issue in many medical problems and, in specific, breast cancer as early discovery of the disease means improving the chance of successful treatment.

In this work, we target the problem of breast cancer diagnosis which is a classification problem. In any classification problem, there are two main stages: feature set extraction and classification.Two commonly applied types of feature sets for breast cancer classification are: image shape-based features and microarray gene expression features. Generally, image-based features are extracted from images based on edge information, local regions, or based on measurements of pixel intensities of the object(s) in the image [3]. For breast cancer data sets, features are usually computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. FNA is a diagnostic procedure used to investigate lumps or masses under the skin. It involves fluid extraction from a breast mass using a small needle and then it is visually inspected under the microscope [11]. These features are related to the shape of the cell nuclei present in the image. A commonly source for these features is the Wisconsin Diagnostic Breast Cancer (WDBC) data set which is publicly available from the UCI repository [1]. In this data set, thirty features are computed for each cell nucleus in which they provide a description for the characteristics of the cell nucleus. These features are calculated using active contour model or snakes framework. An example of these features is radius which is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. Another example is the compactness feature which is calculated by dividing the square of the perimeter to the area. The total distance between consecutive snake points represents the nucleus perimeter and the area is measured by counting the number of pixels on the interior of the snake and adding one-half of the pixels on the parameter [10]. WDBC has been widely used in the pattern recognition community [e.g. [8-11]].

Another commonly applied type of features is gene expression features [e.g. [2], [14], [16], [18]]. Gene expression patterns are captured using cDNA microarrays. cDNA microarrays are novel biotechnologies increasingly used in the cancer research. They allow the monitoring of expression levels of thousands of genes concurrently. Therefore, they can lead to more understanding of the molecular variations among tumors and lead to a more accurate classification [11]. cDNA microarrays consist of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic arrayer [11]. The matrix consists of n rows and m columns, which is known as gene expression profile. The rows represent genes and the columns represent samples. Comparisons can be conducted between the rows (genes) or the columns (samples) of the matrix. If two rows are similar, it can be hypothesized that the two genes are co-regulated and possibly functionally related. Due to complexity of gene expression data, machine learning techniques are implemented to understand and classify or characterize gene expression data [2].

Methods like clustering are usually applied to cluster genes on the basis of similarity in the pattern where this expression varies over samples. Gene expression patterns are grossly divided into normal or malignant human breast tissues [7]. The researchers that implement this type of features argue that tumors show great variation in their patterns of gene expression. Moreover, the variation is multidimensional; where many independent sets of genes show independent patterns of variation. In addition, these patterns have a persistent order reflecting relationships among the genes and relations among tumors and a connection between certain genes and certain tumors [7]. Thus, these features can be discriminative for breast cancer classification.

For any classification problem, it is important to have a strong learning algorithm in combination with an effective feature set. Learning algorithms can be split into two main groups: discriminative classifiers and generative classifiers. Discriminative classifiers emphasize the modeling of class boundaries without attempting to model the entire

underlying class density [3]. Examples of this model include Support Vector Machines (SVM), Neural Networks (NN), Decision Trees (DT), and Bagging.

SVM maps pattern vectors to a higher dimensional feature space where a maximal separating hyperplane is constructed. For a two class problem, two parallel hyperplanes (canonical) are constructed on each side of the separating hyperplane that separates the data. The points that lie on the separating hyperplane are called support vectors. The distance between canonical hyperplanes and the separating hyperplan is called margin. The main idea is to maximize the margin between the classes by selecting a minimum number of support vectors [3, 6].

A great attention from researchers worldwide has been paid to another discriminative approach which is neural networks. They are inspired by the way human brain processes information. Neural networks are computational systems that map complex relationships between inputs and outputs. They are made up of interconnected nodes or neurons and weighted connections [15]. Two commonly applied algorithms of neural networks are: multilayer perceptron (MLP) and radial basis function (RBF). Another widely applied discriminative classification approach is based on decision trees. Its basic structure consists of one root node, a number of internal nodes and a set of terminal nodes. Classification is attained with this algorithm by recursively dividing the data down the decision tree according to the defined classification rule [12]. Two popular classification algorithms of decision trees are: J48 and random forests. In addition to the aforementioned methods, bagging or bootstrap aggregation has been commonly applied. It is an ensemble method for improving base classification algorithms and reducing its variance [17].

On the other hand, generative classification algorithms have been widely implemented. Given a set of class labels $\{C_k\}_{k=1..K}$, generative classifiers model the class likelihood $P(y|C_k)$, and estimate the priors, $P(C_k)$, for each of the C_k individually. Once class densities are learnt, classification is attained by computing the likelihood of a new observation and assigning it to the class label providing the maximum posterior value, $P(C_k|y) \propto p(y|C_k)p(C_k)$. Bayesian approach is a common example of generative approaches. It is a probabilistic classifier that is based on applying Bayes' theorem. This theorem relates the conditional and marginal probabilities of random variables. Another commonly applied generative approach is Probabilistic Principle Component Analysis (PPCA). It describes a probabilistic framework for PCA, a popular dimensionality reduction method, by considering it as a maximum likelihood solution for a latent variable method. PPCA is applied for data from each class to directly model its density with maximum likelihood. In PPCA, a P-dimensional observed data vector y can be described in terms of a D-dimensional latent vector x as:

$$y = Wx + \mu + \epsilon \qquad (1)$$

Where W is a $P * D$ matrix describing a linear transformation and ϵ is an independent Gaussian noise with a spherical covariance matrix $\sigma^2 I$. The probability of an observed data vector y is:

$$p(y) = (2\pi)^{-P/2} |C|^{-1/2} exp\left(-\frac{1}{2}(y-\mu)^T C^{-1}(y-\mu)\right) \qquad (2)$$

Where C is the model covariance matrix given by:

$$C = WW^T + \sigma^2 I \qquad (3)$$

By assuming a Gaussian prior over the latent variable x , an Expectation Maximization (EM) algorithm can be deployed to find the parameters μ , σ^2, and W. To deal with globally non-linear but locally linear structure of data, finite mixture model of PPCA has been proposed, named mixture of PPCA (MPPCA) [3]. Another popular mixture model is the Gaussian Mixture Model (GMM) which models the probability density function of observed variables using a multivariate Gaussian mixture density. Thus, given a number of input data, it improves the weights of each distribution through EM algorithm [6]. In

[4], the authors proposed MLiT(N), a mixture of Normalized Gaussian distributions under linear transformations. It is a mixture model for dimensionality reduction where at each iteration of an EM process and within each of the mixture components, the class' data are projected by a linear transformation in a way that maximizes the log-likelihood of the data in that component, for that particular class. The transformation matrices are not restricted to be orthogonal. To avoid a singularity problem, normalization is applied to the projection matrices.

Generally, learning approaches have proved successful for the purpose of breast cancer classification. In the work of [10], a comparison between J48, SVM, and other classifiers is carried out. The authors reported 97.7% accuracy with the SVM method. In the work of [11], extensive comparative analysis of classification algorithms for breast cancer diagnosis is presented. The algorithms include SVM, neural networks and Bayes methods. The authors experimented with different kernel functions with the SVM classifier. Neural networks outperformed other classifiers with 97.9 % accuracy with 10-fold cross validation on the WDBC data set. In the work of [8], a clustering approach is proposed and tested on the WDBC data set. The authors achieved 97.2% accuracy with the hold-out validation method. Recently, the authors in [9] ran experiments with ensemble learners in conjunction with feature selection for the purpose of breast cancer diagnosis. The authors achieved an accuracy of 95.7% with 10-fold cross validation on the WDBC data set.

On the other hand, there has been an extensive research work that is based on gene expression data. The researchers that use this approach argue that image shape-based features are not very discriminative between cancer subtypes, and therefore gene expression data can provide richer information for classification [2]. In [2], ensemble machine learning classifiers are applied to a number of cancer data sets that are based on gene expression information. For the breast cancer data set, the authors reported their highest performance accuracy of 89.47 % with bagging and Adaboost approaches. In the work of [16], seven different machine learning techniques have been applied to microarray gene expression data for the purpose of predicting breast cancer and other cancer types. For breast cancer problem, the highest performance results were achieved with SVM and RBF Neural Networks with an accuracy of 97.6 with 10-fold cross validation. Other works such as [2, 7, 13-14, 18] have also targeted the problem of breast cancer classification based on the application of machine learning techniques to gene expression data.

There has been an extensive amount of literature related to the application of machine learning algorithms to breast cancer data sets. The feature sets used in these works are based on image or gene expression data. However, there has not been in the literature any work that directly compares prediction performance of learning algorithms across the two types of feature sets. Thus, in this paper, we compare the performance of machine learning algorithms on the two types of feature sets. Moreover, we provide extensive analysis for the performance results and we choose the best classification performance feature set and test it with a number of mixture models that are based on a generative classification approach.

The rest of this paper is organized as follows: In section 2, we explain the data sets and classification algorithms that are used in the experiments. Section 3 presents the experimental results and analysis. Finally, section 4 concludes the paper and suggests future work.

2. Classification

In this section, we present the data sets used in our experiments. We will also present the implemented classifiers and the performance evaluation mechanism.

2.1 The Data Sets

To evaluate the performance of breast cancer classification based on image shape features and gene expression data, two commonly applied data sets are selected. The first data set is the Wisconsin Diagnostic Breast Cancer (WDBC) which is publicly available from the UCI repository [1]. The second data set is the microarray gene expression (MGE) data set which is available from the University of Stanford [19].

WDBC Data Set

This data set is the result of efforts made at the University of Wisconsin hospital for diagnosis of breast tumor using FNA test. The data set consists of 569 instances where 357 instances are benign and 212 instances are malignant. Each instance is represented with a 30-dimensional feature vector describing computations related with ten real-values features. The features are computed from a digitized image of an FNA of a breast tumor and describe the cell nuclei present in the image. These ten real values are the following: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

MGE Data Set

This data set is a result of efforts made by the authors in [7]. It characterizes gene expression profiles of 84 samples composed of 65 tumors and 19 cell lines, using cDNA microarray, representing 8102 human genes. A filtering process is applied resulting in 1753 genes that are related to the 84 samples of normal vs. tumor subtypes. The data set is available from Stanford Breast Cancer Microarray Project from the University of Stanford [19].

2.2 Classifiers

The classifiers that are used in our experiments are the following: support vector machines, radial basis function neural networks (RBF NN), MLP neural networks (MLP NN), Bayes, J48 decision tree, RF decision tree and bagging. Moreover, we run experiments with a number of mixture models that include: Gaussian mixture model (GMM), mixture of probabilistic PCA (MPPCA) and Mixture of normalized linear transformations (MLiT (N)). In section 1, we provided an overview about all the aforementioned techniques.

The performance of the classifier is evaluated in terms of classification accuracy. Classification accuracy is calculated as the proportion of the number of correctly classified instances against the total number of tested instances.

3. Experimental Results and Analysis

Experiments are conducted in order to evaluate the classification performance of the two types of feature sets for the aim of breast cancer classification. For this purpose, we carried out three major experiments. The first experiment is implemented with seven classifiers on

Table 1. Performance Results Across Different Data Sets with 10-fold Cross Validation

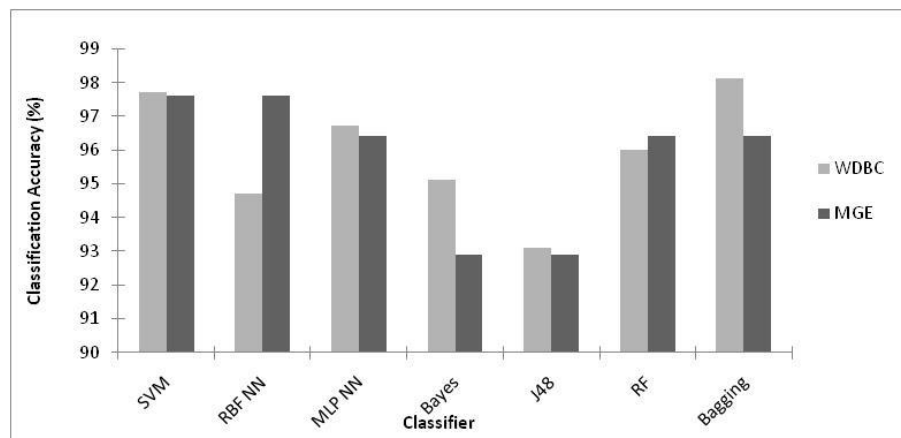| Classification Algorithm	WDBC (%)	MGE (%)
SVM	97.7	97.6
RBF NN	94.7	97.6
MLP NN	96.7	96.4
Bayes	95.1	92.9
J48	93.1	92.9
RF	96.0	96.4
Bagging	98.1	96.4

Figure 1. Classfifcation Accuracy Across Data Sets with 10-fold Cross Validation
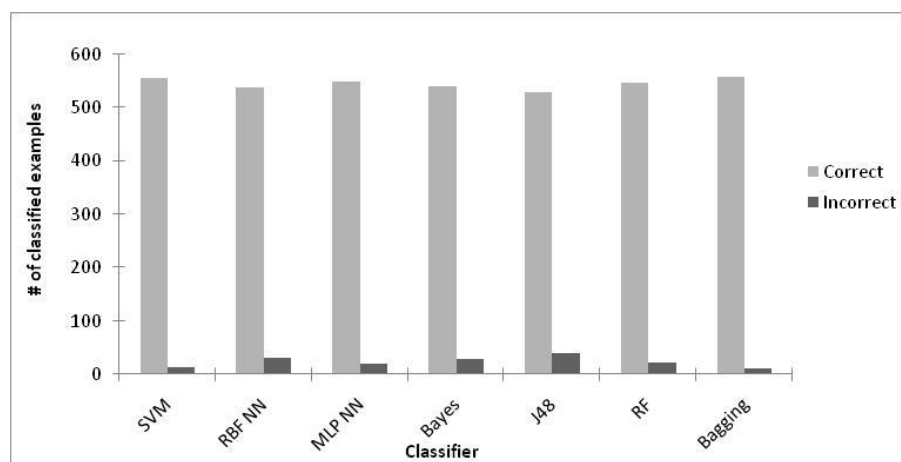

Figure 2. Number of Correctly and Incorrectly Classified Sample on WDBC Data Set with 10-fold Cross Validation.

the WDBC data sets and the results are compared with that of the performance of the same seven classifiers on the MGE data set. The results are validated with ten-fold cross

validation. In the second experiment, we select the best performance data set and test it with five-fold cross validation for further examination of its accuracy robustness. This same data set is further examined with generative mixture models and the results are presented in the third experiment.

3.1. Experiment 1: A Comparison of Classification Performance between Image Shape-based Features and Microarray Gene Expression Features with Ten-fold Cross Validation

Our main aim is compare the breast cancer classification performance of the WDBC data set and the MGE data set. For the MGE data set, we use the accuracy results of a recently published work by a group of scientists working the medical research field [16]. In this work, the authors examined MGE data set with the following classification algorithms: SVM, RBF neural networks, MLP neural networks, Bayesian, J48 decision trees, id3 and bagging. Ten-fold cross validation was implemented in the aforementioned work and classification accuracy was used as a measure of performance. We chose the same classification algorithms and excluded the id3 algorithm as it works only with nominal attributes. These seven chosen algorithms were examined on the WDBC using ten-fold cross validation and classification accuracy was calculated.

Table 1. and Figure 1. show the performance classification of this experiment. It is clear from Table 1. that the highest classification performance results across the two data sets are achieved with the WDBC data set using the bagging approach with an accuracy of 98.1%. The best results achieved on the gene expression data set is with SVM and RBF neural networks with a classification performance of 97.6%. Across five of the implemented classifiers, the image shape-based features outperformed the classification performance of that of the gene expression features. Generally, image shape-based features seem to be more discriminative for breast cancer classification. However, there is no dramatic difference between the performance of the two feature sets with accuracy results differences are in the range of 0.1% to 2.9%. In Figure 2, we show in details the number of correctly and incorrectly classified examples across the different classifier with the WDBC data set.

3.2. Experiment 2: Classification Performance with Five-fold Cross Validation on the WDBC Data Set

As the WDBC data set provided the best performance results, we decided to further examine its classification performance with five-fold cross validation. In this test, all the 569 instances of the data set are used and divided into five disjoint groups instead of ten disjoint data sets as with the ten-fold cross validation. Four of the disjoint data set are used for training and the 5th one is used for testing. The algorithm runs for five times and the average accuracy across all the folds is calculated. Figure 3 shows the number of correctly and incorrectly classified instances across different classifiers using five-fold cross validation on the WDBC data set. Moreover, Table 2 shows the accuracy results of this method. It is clear from this table that the bagging algorithm again provides the highest accuracy results with an accuracy performance of 97.5%. The same accuracy is achieved with SVM. It is encouraging results considering the size of the testing set which is bigger than the one used in the first experiment. Moreover, we can note from this table that the accuracy results across all the classifiers, except j48, have slightly decreased. Usually, with the cross-validation test the results decrease as the number of folds decrease and the size of the testing set increases.
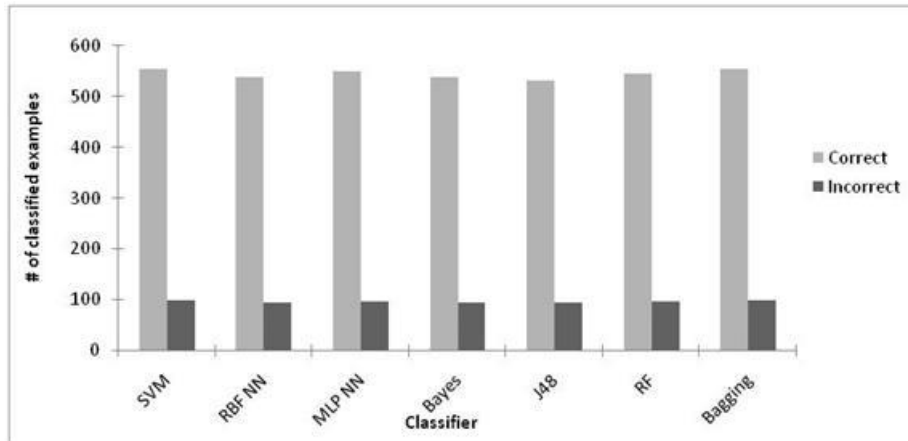
Figure 3. Number of Correctly and Incorrectly Classified Sample on WDBC Data Set with 5-fold Cross Validation

Table 2. Performance Results on WDBCData Set with 5-fold Cross Validation

Classification Algorithm	WDBC (%)
SVM	97.5
RBF NN	94.6
MLP NN	96.5
Bayes	94.6
J48	93.3
RF	95.8
Bagging	97.5

Table 3. Five-fold Performance Results on WDBC Data Set with Generative Classifiers

Classification Algorithm	WDBC (%)
GMM	95.9
MPPCA	94.7
MLiT (N)	96.1
Bayes	94.6

However, the results remain stable with slight decreases and there are no dramatic changes between the two validation tests.

3.3. Experiment 3: Performance Evaluation with Generative Classifiers on the WDBC Data Set

In this experiment, we further examine the performance of generative classifiers on the best performing data set; WDBC. For this purpose, we chose widely applied

generative classifiers and compare their performance with that of discriminative classifiers. The tested classifiers are: GMM, MPPCA, MLiT (N) and the Bayes approach. Table 3 shows the classification performance of the selected classifiers on the WDBC data set with five-fold cross validation.

It is clear from this figure that the best performance classifier is MLiT (N) with an accuracy of 96.1%. This accuracy performance is less than that of the best discriminative classifier, SVM and Bagging, with a decrease of 1.4%. However, the generative classifiers seem to be performing well on the WDBC data set with slight decreases from that of discriminative classifiers.

4. Conclusions and Future Work

With recent improvements in machine learning tools, they are being applied into a variety of technological domains and one of these domains is medical diagnosis. The problem of breast cancer diagnosis is an interesting problem as there is high number of women that are diagnosed with it worldwide. Early diagnosis means a better chance for treatment. In this paper, we targeted this problem with an aim of examining the classification performance of machine learning techniques with different types of feature sets.

For this purpose, two widely applied data sets are selected: the first data set is based on tumor image shape-based features and the second one is based on gene expression microarray data. From our experiments, we can conclude that the best performing feature set is the image shape-based features where the majority of the classification algorithms achieved higher accuracy results with this type of features using ten-fold cross validation. To further examine the performance of the aforementioned feature set, we examined it with five-fold cross validation and the data set retained its accuracy with slight decreases as the size of the testing data increased. Moreover, we studied the classification performance of generative classifiers with a focus on mixture models. The results were encouraging but the overall performance is less than that of discriminative classifiers. In the future, we plan to shift the focus from the feature sets to the learning algorithms for an aim of an improved performance.

5. References

[1] A. Asuncion and D. Newman, "UCI machine learning repository", **(2007)**.
[2] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification", Applied Bioinformatics, vol.2, no.3, **(2003)**, pp. 75-83.
[3] A.F. Otoom, "Effective features set and dimensionality reduction for object classification", University of Technology, Sydney (UTS), **(2010)**.
[4] A. F. Otoom, H. Gunes, O.P. Concha, and M. Piccardi, "MLiT: mixtures of Gaussians under linear transformations", Pattern Analysis and Application, vol. 14, no. 2, **(2011)**, pp. 193-205.
[5] American Cancer Society, http://www.cancer.org/, **(2014)**.
[6] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, **(2006)**.
[7] C. M. Perou, T. Sørlie, M. B. Eisen, et al., "Molecular portraits of human breast tumours", Nature, vol. 406, no. 17, **(2000)**, pp. 747-752.
[8] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial Bee Colony (ABC) algorithm", Applied Soft Computing, vol. 1, no.1, **(2011),** pp. 652-657.
[9] D. Lavanya and Dr. K. U. Rani, "Ensemble Decision Tree Classifier for Breast Cancer Data", International Journal of Information Technology Convergence and Services (IJITCS), vol .2, no. 1, **(2012)**, pp. 17-24.
[10] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers", International Journal of Computer and Information Technology, vol.1, no.1, **(2012)**, pp. 36-43.
[11] I. Maglogiannis, E. Zafiropoulos, and I. Anagnostopoulos, "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers", Applied intelligence, vol. 30, no.1, **(2009),** pp. 24-36.

[12] J. R. Otukei and T. Blaschke, "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms", International Journal of Applied Earth Observation and Geoinformation, vol. 12, **(2010)**, pp. 27-31.

[13] M. Burton, M. Thomassen,Q. Tan and A. Kruse, "Gene expression profiles for predicting metastasis in breast cancer: a cross-study comparison of classification methods", The Scientific World Journal, vol. 2012, **(2012)**.

[14] M. Dettling and P. Bühlmann, "Boosting for tumor classification with gene expression data", Bioinformatics, vol. 19, no.9, **(2003)**, pp. 1061-1069.

[15] M. M. Khan, A. M. Ahmad, G. M. Khan, G., and J. F. Miller, "Fast learning neural networks using Cartesian genetic programming", Neurocomputing, vol. 121, **(2013)**, pp. 274-289.

[16] M. Pirooznia,J. Y. Yang, M. Q. Yang and, Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data", BMC genomics, vol. 9, no.1, **(2008)**.

[17] P. Bühlmann, "Bagging, boosting and ensemble methods", Handbook of Computational Statistics, Springer Berlin Heidelberg, **(2012)**, pp. 985-1022.

[18] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data", Journal of the American statistical association, vol. 97, no 457, **(2002)**, pp. 77-87.

[19] Stanford Breast Cancer Microarray Project, http://genome-www.stanford.edu/sbcmp/, **(2014)**.

Authors

Ahmed Fawzi Otoom is currently working as an assistant professor at the Software Engineering department at Hashemite University, Jordan. He has a PhD degree in computer science from the University of Technology, Sydney (UTS), Australia, 2010. He received his BS in Computer Science from Jordan University of Science and Technology, Jordan, and MS in Software Engineering from the University Western Sydney, Australia, in 2002 and 2003, respectively His main research interests include computer vision and pattern recognition techniques for image and video analysis with a focus on realistic scenarios within video surveillance. He has also research interests in the bioinformatics area.

Emad E. Abdallah received his PhD in Computer Science from Concordia University in 2008, where he worked on multimedia security, pattern recognition and 3D object recognition. He received his BS in Computer Science from Yarmouk University, Jordan, and MS in Computer Science from the University of Jordan in 2000 and 2004, respectively. He is currently an Assistant Professor in the Department of Computer Information Systems at the Hashemite University (HU), Jordan. Prior to joining HU, he was a Software Developer at SAP Labs Montreal. His current research interests include computer graphics, multi-media security, pattern recognition, and computer networks.

Maen Hammad is an Assistant Professor in Software Engineering Department at The Hashemite University, Jordan. He completed his Ph.D. in computer science at Kent State University, USA in 2010. He received his Master in computer science from Al-Yarmouk University—Jordan and his B.S. in computer science from The Hashemite University—Jordan. His research interest is Software Engineering with focus on software evolution and maintenance, program comprehension and mining software repositories.