GA-based Feature Selection with ANFIS Approach to Breast Cancer Recurrence

Hamza Turabieh

Collage of Computer and Information Technology, Information Technology Department
Taif University, KSA

Abstract

Automatic disease diagnosis systems are important for medical fields. These systems have been used to help doctors to make better diagnosis. Breast cancer is a very common class of cancers among women. In this paper, we focus on breast cancer recurrence problem, hybridizing two methodologies, Genetic Algorithm (GA) and Adaptive Neuro Fuzzy Inference System (ANFIS), to develop a good diagnosis system. GA has been used as a selection algorithm to find the best features, whilst ANFIS has been used as a classifier algorithm. The robustness of the proposed hybrid methodology is examined using classification accuracy, sensitivity, and specificity. The proposed hybrid algorithm has achieved accuracy of 88% for training dataset and 71% for testing. The results demonstrate the effective interpretation and point out the ability to design a good diagnosis system.

Keywords: *Breast cancer; Feature Selection; Genetic Algorithm; ANFIS*

1. Introduction

Breast cancer is the most common form of cancer in women, despite of huge public awareness and scientific research in this field. One in nine women is expected to have breast cancer during her lifetime, and one in 27 will die out of it. High classification accuracy of breast cancer is an important real-world medical problem. After the lung cancer, breast cancer is considered as the second leading cause of cancer death for women in United States, around 40,000 women in US will die from the breast cancer in 2014 [1]. Breast cancer diagnosis (BCD) is done either during a screening examination, before or after symptoms have developed, or when a woman feels a lump. Most masses seen on mammogram and most breast lumps turn out to be benign, which means that the lump is not life-threatening. Generally, when a breast cancer is detected based on clinical exam or breast imaging, microscopic analysis of breast lump is necessary for a definitive diagnosis of the patient status. Moreover, the economic and social values of BCD systems are very high. This area of research has long attracted the attention of the Bioinformatics and Artificial Intelligence communities [2][3][4][5][6][7][8].

Research in the area of machine learning for medical diagnosis has been the center attention for several years like Pattern Recognition, Neural Networks, Evolutionary Algorithms, Naive Bayesian classifier, and Inductive Learning of symbolic rules [9]. However, up to date, none of these methods have been used in clinical diagnosis in term of routine usage or even to replace the radiologist. The reason behind that is the miss understanding that machines could replace the job of the radiologist. A second reason is that we still disbelieve and doubt the abilities of the machine to perform a medical diagnosis. A good computerized diagnostic system should process two characteristics. First the system should have the highest possible performance, i.e. diagnosis the presented case correctly as either benign or malignant. Moreover, the system should not only provide a binary diagnosis (benign or malignant), but also present a numeric value that presents the degree of confidence about the binary diagnosis value. Second, the system should be human-friendly to simplify the physician job.

In this paper, we combine two algorithms, genetic algorithm and adaptive neuro fuzzy inference system for breast cancer recurrence diagnosis. The GA is used as a feature selection to find the best features (inputs) from a pool of features while the ANFIS is used as a classifier algorithm to predict the existence of cancer or not. The major advantage of the proposed system is that the GA exploration process is able to find the most effective features (inputs), while the ANFIS is able to find an accurate rule based system based on its inputs.

The rest of the paper is organized as follows. Section 2 gives background information about breast cancer classification and previous studies. We present the prognosis of breast cancer recurrence is Section 3. The proposed algorithm is presented in Section 4. In Section 5, we give the performance metrics to evaluate any medical system. The experimental results are presented in Section 6. Finally, we conclude this paper in Section 7 with future directions.

2. Background

Breast cancer is a leading cause of death worldwide. As a result, treating breast cancer requires knowing what abnormal behaviors are happening inside the cells. Recently, machine learning tries to help doctors to make a correct decision based on past experiences. Machine learning algorithms take many data from past experience as a training dataset and build an internal model to predict another data which is the testing dataset. Research in the area of neural networks for breast cancer diagnosis has been at the center of attention for several years [2][3][6]. A backpropagation algorithm is applied to learn from 133 instances containing 43 features rated between 0-10 by Michie et al. [10]. The performance of the backpropagation algorithm was found to be competitive to the domain experts. The backpropagation algorithm is also applied by Wu et al. [11] but with different input sets derived from a group of blood tests. However; the data contains 10 features for 104 instances and the algorithm fails to perform well. The time needed to train the backpropagation algorithm is considered as the main drawback of this technique. Moreover, the backpropagation algorithm also requires very large sample sets to train model efficiently. Therefore, an evolutionary programming algorithm is used to train neural network to overcome the weakness of the backpropagation algorithm by Wilding et al. [12]. The proposed algorithm was tested on Wisconsin dataset which is obtainable by anonymous ftp ice.uci.edu [13]. A Support Vector Machine (SVM) combined with feature selection to predict the occurrence of breast cancer is applied by Polat and Güneş [4]. The result of SVM was promising. An automatic diagnosis system based on Association Rules (AR) and neural networks is applied to breast cancer diagnosis by Akay [5]. A hybrid method that combines filters and wrapper algorithms is used to solve the breast cancer problem by Blake and Merz [14]. Artificial Metaplasticity Multilayer Perceptron algorithm (AMMLP) was compared with the backpropagation algorithm and other recent classification methods by Peng et al. [15] and the results were very promising to predict the breast cancer. A supervised fuzzy clustering method is applied by Marcano-Cedeño [16] and the results were good enough compared with other clustering algorithms. Karabatak and Cevdet-Ince [6] compared five different classifiers methods; support vector machine, probabilistic neural network, recurrent neural network, combined neural networks, and multilayer perceptron neural networks; over Wisconsin breast cancer dataset. Abonyi and Szeifert [17] proposed the least square with support vector machine to produce high accuracy rate to the breast cancer diagnosis.

3. Prognosis of Breast Cancer Recurrence

Cancer is a family of diseases that involve uncontrolled cell growth wherein cells split and grow in exponential manner, generating malignant tumors and spread to other parts of the body. Breast cancer begins in the breast tissue that develops from cells of the breast. There are several risk factors that increase a woman's chance to obtain breast cancer, such as: age, family history, genetics, menstrual periods, not having children, and obesity. However, up to date, scientists do not yet know which factors that causes most breast cancer and increase its chance. As a result, researchers make a great jump in understanding how normal breast cancer cells change to become cancerous [18].

In this paper, the prognosis of breast cancer recurrence dataset (UCI Machine Learning Repository) was analyzed. The dataset consists of 286 patients with known diagnosis status five years after the operation is available. Each record of the dataset has nine attributes and associated with its class label, which is either recurrence or no recurrence. The nine attributes are detailed in Table 1. The measurements are assigned an integer value between 1 and 11. There are 201 patients who did not have recurrence after five years and 85 who did. This data was verified after collecting, and thus is likely to be relatively error-free. In 30% of the patients that undergo a breast cancer operation, the illness reappears after five years; thus prognosis of recurrence is important. This dataset has nine instances with missing attributes values. The remaining 277 instances are taken for use in this paper. Therefore, we used 208 (75%) instances for training and 70 (25%) instances for testing the proposed algorithm.

Table1: Prognosis of breast cancer recurrence attributes.

Attribute Number	Attribute Description	Values of Attribute	Mean	Standard Deviation
1	age	1- 6	4.64	1.01
2	menopause	1-3	1.90	0.98
3	tumor size	1-11	5.88	2.13
4	inv-nodes	1-6	1.50	1.13
5	node-caps	1-2	1.20	0.40
6	deg-malig	1-3	2.05	0.72
7	breast	1-2	1.47	0.50
8	breast-quad	1-5	2.17	1.20
9	irradiat	1-2	1.22	0.41

4. The Algorithm

In this section, two algorithms are hybridized to gain the advantages of each one. The first one is GA, which is used as a feature selection to reduce the number of inputs. The second is ANFIS algorithm which is used to build a rule based system for breast cancer recurrence. Figure 1 illustrates a pictorial diagram for the proposed system.

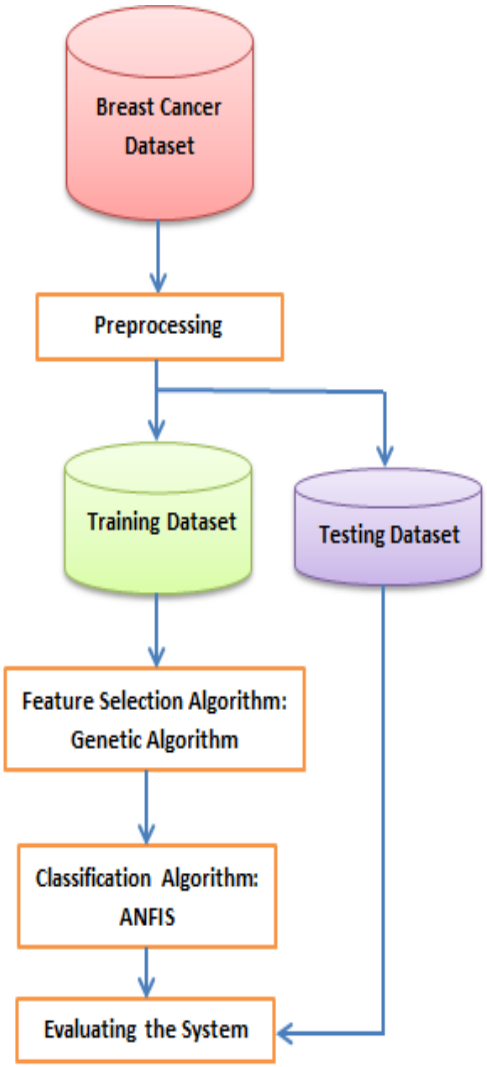

Fig. 1: Proposed GA-ANFIS classification system.

The first step of the proposed system is to divide the dataset into two partitions for training and testing processes. Training process is used to build an internal model; the fundamental procedures of this process include selection of the best features to reduce the size of training dataset and apply ANFIS as a classifier algorithm. While the testing process is used to evaluate the system. The next subsections present the genetic algorithm concepts as a feature selection algorithm, followed by illustration of ANFIS classification algorithm.

4.1 Genetic Algorithm as a Feature Selection

Feature selection is a pre-process approach that is usually used on high-dimensional data. its objectives include reducing the dimensionality, removing irrelevant data, redundant data, facilitating data understanding, reduce the amount of needed data for learning process, and enhancing the obtained accuracy of the algorithms. The main target of feature selection is to choose a small subset of features that ideally is necessary and sufficient for the classification system [19].

A GA is considered as a heuristics search and optimization algorithm that is inspired natural evaluation [20], which simulates the notation of natural evolution to the computing era. The father of GA is John Holland who introduces it by explaining adaptive processes of natural systems. The evolution process of the genetic algorithm starts by constructing an initial population (solutions) randomly, moves toward a global optimal solution, and stops the searching process when the stop criteria is met. In each generation, each individual is evaluated based on fitness function. Multiple individuals are stochastically selected from the current population based on their (fitness), modified (mutated or recombined) to generate a new solutions, which becomes current in the next iteration of the algorithm [21]. Figure 2 presents the basic GA algorithm.

In this paper, a GA is used as a feature selection algorithm. A chromosome is encoded as a string of bits whose size corresponds to the number of features. A *0* or *1*, at position *i*, indicates whether the feature *i* is selected (*1*) or not (*0*). Figure 3 presents a pictorial diagram for chromosome, where features 2, 3, 5, and 8 are selected. Moreover, a GA consists of three main operators, crossover, and mutation and selection mechanisms. These operators allow the GA to explore the search space.

Subset Size-Oriented Common Feature Crossover Operator (SSOCF) is used as a crossover mechanism, which keeps useful informative blocks and produce offsprings (new solutions) that have the same distribution as parents [22]. Features shared by the parents are kept by offsprings and non-shared features are inherited by offsprings corresponding to the i^{th} parent with the probability $(n_i - n_c / n_u)$, where n_i is the number of selected features of the i^{th} parent, n_c is the number of commonly selected features across both mating partners, and n_u is the number of non-shared selected features. Figure 4 illustrates the SSOCF operation where there are four common bits between two parents (as shown in the mask), while there are six common bits between the two offsprings.

The diversity of a GA is done based on mutation mechanism, a chromosome has a probability *Pmut* to mutate. We choose randomly number of bits *n* to be flipped. In order to create a large diversity, we select $n \in [1,5]$.

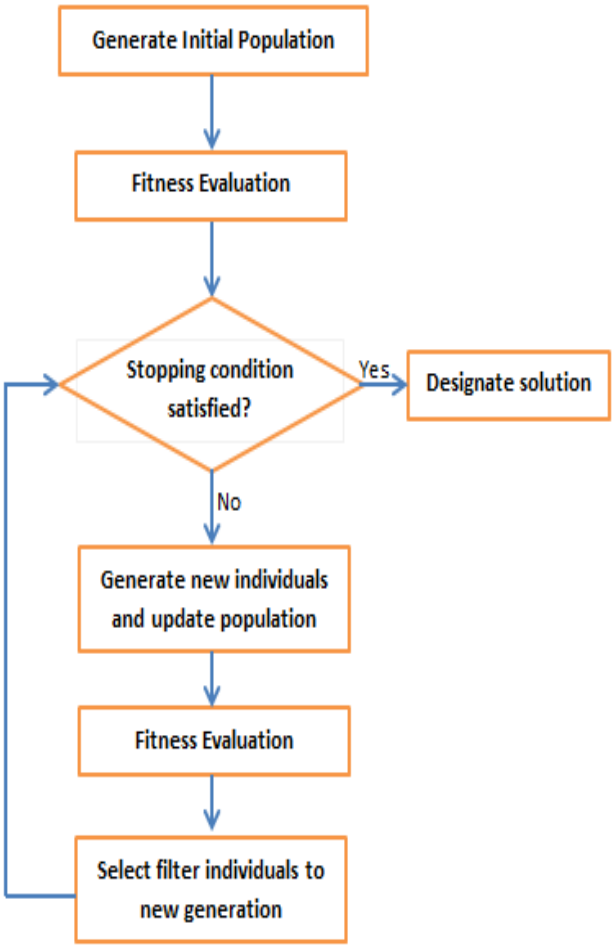

Fig. 2: Basic GA.

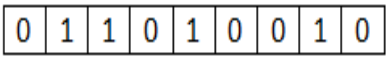

Fig. 3: Chromosome presentation.

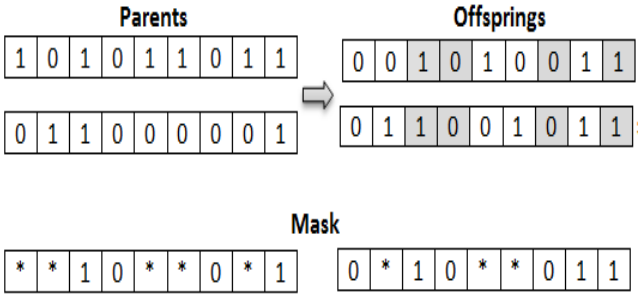

Fig. 4: SSOCF crossover mechanism.

A probabilistic binary tournament selection has been used as selection mechanism, where the tournament selection holds *n* tournaments to choose *n* individuals. Each tournament consists of sampling two elements of the population and choosing the best one with a probability $p \in [0.5, 1]$. The fitness function is evaluated as following:

$$F = ((1-S)\times\frac{\frac{T}{10}-1\times10\times SF}{T})+2\times(S\times\frac{\frac{T}{10}+10\times SF}{T})$$

Where:

$S=\frac{|A\cap B\cap C...|}{|A\cup B\cup C...|}$, where *A, B, C,* are the selected features

T= Total number of features

SF= Number of selected significant features (selected features that are not too close in term of chromosomal distance).

4.2 ANFIS Algorithm

Jang [23] introduced ANFIS, which is a fuzzy Sugeno model. ANFIS combines the Fuzzy Inference System (FIS) into the framework of adaptive networks. The ANFIS is a fuzzy system that is incorporates with neural network to determine the fuzzy sets and fuzzy rules. An adaptive network is a network structure that consists of number of nodes connected to each other. The output of these nodes depends on adaptable parameters relating to these nodes. To present the ANFIS architecture, two fuzzy if–then rules based on a first order Sugeno model are considered:

Rule 1:

if x is A_1 and y is B_1, then $f_1 = p_1x + q_1y + r_1$

Rule 2:

if x is A_2 and y is B_2, then $f_2 = p_2x + q_2y + r_2$

Where x and y are the inputs for the FIS, A_i and B_i are the fuzzy sets, f_i are the outputs within the fuzzy region specified by the fuzzy rule, p_i, q_i and r_i are the design parameters that are determined during the training process . The ANFIS architecture representation for this system is as shown in Figure 5, for more details about ANFIS see [23].

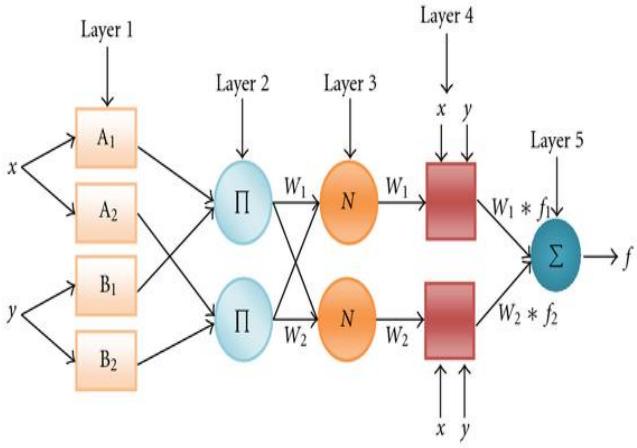

Fig. 5: ANFIS architecture.

5. Performance Metrics

To evaluate medical systems, there are several measurements usually used. In this paper, three metrics were used; accuracy, specificity, and sensitivity. A system of a single prediction has four different possible outputs, as shown in Table 2.

Table 2: Different outputs of two class prediction.

Actual Class	Prediction Cass	
	Correct	Incorrect
Yes	TP	FN
No	FP	TN

The True Positive (TP) and True Negative (TN) are correct classification. A False Positive (FP) occurs when the output is incorrectly predicted as yes (or positive) when it is actually no (negative). A False Negative (FN) occurs when the output is incorrectly predicted as no (or negative) when it is actually is yes (positive). The following equations are used to measure the accuracy, specificity and sensitivity [25].

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

$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

A sensitivity of 100% means that the test recognizes all sick people as such. Thus, in a high sensitivity test, a negative result is used to rule out the disease. A specificity of 100% means that the test recognizes all healthy people as healthy. Thus, a positive result in a high specificity test is used to confirm the disease.

6. Simulation Results

The proposed algorithm was programmed using Matlab and simulations were performed on the Intel Core i3 2.4 GHz computer using the prognosis of breast cancer recurrence dataset "UCI Machine Learning Repository", as presented in Section 3. Table 3 shows the parameters for the proposed system.

Table 3: Parameters setting.

	Parameter	Value
Genetic Algorithm	Generation number	1000
	Population size	100
	Crossover rate	0.75
	Mutation rate	[0.5, 1]
	Selection mechanism	Tournament
	Crossover type	SSOCF
ANFIS Algorithm	Number of neurons in input layer	9
	Number of neurons in hidden layer	20
	Number of neuron in output layer	1
	Epoch	100
	Learning method	learngdm

The GA selects four best features (Age, Tumor Size, breast, and breast-quad), as shown in Figure 6. These features are used as inputs to the ANFIS classifier. Table 4 presents the results of the classifier algorithm that have been applied in this paper. The accuracy obtained for training and testing are 88% and 71%, respectively. It is clear that the proposed model is able to obtain high accuracy.

1	0	1	0	0	0	1	1	0

Fig. 6: GA solution.

Table 4: Proposed algorithm results.

	Accuracy	Specificity	Sensitivity
Training dataset	88 %	75%	83%
Testing dataset	71%	64%	73%

Our algorithm is capable to find a high accuracy value for both training and testing dataset. Table 5 shows the results obtained and the comparison with other approaches in the literature such as Dumitru (2009) that employed Naïve Bayesian, Howell (2009) who applied Inductive Logic Programming, and Stefanowski & Wilk (2007) that employed Inductive Rule-based Classification.

Table 5: Comparison between various methods and our method.

Results	Accuracy
Our Algorithm	88%
Dumitru [26]	76%
Howell [27]	75%
Stefanowski & Wilk [28]	80%

7. Conclusion and Future works

In this paper we present a new approach to solve breast cancer recurrences problem. GA is used as a feature selection to find the most sufficient features for ANFIS classifier approach. The robustness of the proposed hybrid methodology is examined using classification accuracy, sensitivity and specificity. The obtained accuracy is 88% for training and 71% for testing. The classification results were consistent with highest result obtained from other classifiers published in the literature. Future work will be aimed to enhance the performance of hybrid algorithm and tested on different medical datasets.

References

[1] The American Cancer Society (year). *Breast Cancer Facts & Figures 2013-2014.* Atlanta, Georgia, USA.

[2] American Cancer Society, http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-040951.pdf, last vistited 10-10-2015.

[3] Floyd, C., Lo, J., Yun, A., Sullivan, D., and Kornguth, P. (1994). Prediction of breast cancer malignancy using an artificial neural network. *Cancer*, 74(11):2944–2998.

[4] Polat, K. and Güneş, S. (2006). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, 17(4), pp. 694-701.

[5] Akay, M. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2): 3240–3247.

[6] Karabatak, M. and Cevdet-Ince, M. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2):3465–3469.

[7] Übeyli, E. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4):1054-1062.

[8] Rodrigues, P., Chang, R., and Suri, J. (2006). Non-extensive entropy for cad systems of breast cancer images. *Computer Graphics and Image Processing*, Brazilian Symposium, 121–128.

[9] Maglogiannis, I., Zafiropoulos, E. and I.Anagnostopoulos. (2009). An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Applied Intelligence, 30(1):24–36.

[10] Michie D., Spiegelhalter D., and Taylor C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood.

[11] Wu, Y., Giger. M., Doi, K., Vyborny ,C. , Schmidt, R., and Metz, C. (1993). Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology*, 187 (1):81–87.

[12] Wilding, P., Morgan, M., Grygotis, A., Shoffner, M., and Rosato E.(1994). Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. *Cancer Letter*, 77(2-3):145–153.

[13] Fogel, D., Wasson, E., and Boughton, E. (1995). Evolving neural networks for detecting breast cancer. *Cancer letters*, 96(1):49–53.

[14] Blake, C., and Merz, C. (1998). UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn/mlrepository.html. *University of California*, Irvine, Dept. of Information and Computer Sciences,.

[15] Peng, L., Yang, B., and Jiang, J. (2009). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*, 179(1), 809–819.

[16] Marcano-Cedeño, A., Quintanilla-Domínguez, J., and Andina, D. (2011). WBCD breast cancer database classification applying artificial metaplasticityv neural network. *Expert Systems with Applications*, 38(8):9573–9579.

[17] Abonyi, J., and Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters*, 14(24): 2195–2207.

[18] Polat, K., and Günes, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, 17(4):694–701.

[19] Jerez-Aragonés, J., Gómez-Ruiz, J., Ramos-Jiménez, G., Muñoz-Pérez J., and Alba-Conejo E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial Intelligence Methods*, 27(1):45-63.

[20] Oreski, S., Oreski, D., and Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert Systems with Applications*, 39(16):12605–12617.

[21] Goldberg, D. and Holland, J. (1988). Genetic algorithms and machine learning. *Machine Learning*, 3(2):95–99.

[22] Abdullah, S., Turabieh, H., (2008). Generating university course timetable using genetic algorithm and local search. *Proceeding of the 3rd International Conference on Hybrid Information Technology*, 254-260.

[23] Emmanouilidis, C., Hunter, A., and MacIntyre, J. (2000). A multiobjective evolutionary setting for feature selection and a commonality-based crossover operator. *The 2000 Congress on Evolutionary Compuatation*, San Diego, California, USA.

[24] Jang, J. (1993). ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(3);665–685.

[25] Loo, C. (2005). Accurate and reliable diagnosisand classification using probabilistic ensemble simplified fuzzy ARTMAP. *IEEE Transaction on Knowledge and Data Engineering*, 17(1): November 2005.

[26] Dumitru, D. (2009). Prediction of recurrent events in breast cancer using the Naive Bayesian classification. *Annals of University of Craiova, Mathematical Computer Science Series,* 36(2):92–96.

[27] Howell, C. (2009). *Application of ILP to a Breast Cancer Recurrence Data Set with Aleph*. Fall 2009.

[28] Stefanowski, J. and Wilk S. (2007). Improving rule based classifiers induced by MODLEM by selective pre-processing of imbalanced data. *Proceedings of the RSKD Workshop at ECML/PKDD*, Warsaw, 54–65.

Hamza Turabieh is an Assistant professor at Computer Science department- Faculty of Science and Information Technology- Taif University. Hamza Turabieh received his BA, M.Sc. degrees in Computer Science from Balqa Applied University in 2004 and 2006 respectively in Jordan. Turabieh obtained his PhD from National University of Malaysia (UKM) in 2010, his research interests and activities lie at the interface of Computer Science and Operational Research. Intelligent decision support systems, search and optimization (combinatorial optimization, constraint optimization, multi-modal optimization and multi-objective optimization) using heuristics, local search, hyper-heuristics, met heuristics (in particular memetic algorithms, particle swarm optimization), hybrid approaches and their theoretical foundations.