International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 03, March-2015

Robust Breast Cancer Diagnosis on Four Different Datasets Using Multi-Classifiers Fusion

Ahmed Abd El-Hafeez Ibrahim[1], Atallah I. Hashad[2], Nigm El-Deen M. Shawky[3] and Aly Maher[4]
Arab Academy for Science,
Technology & Maritime Transport,
Cairo, Egypt

***Abstract*- The goal of this paper is to compare between different classifiers or multi-classifiers fusion with respect to accuracy in discovering breast cancer for four different data sets. We present an implementation among various classification techniques which represent the most known algorithms in this field on four different datasets of breast cancer two for diagnosis and two for prognosis. We present a fusion between classifiers to get the best multi-classifier fusion approach to each data set individually. By using confusion matrix to get classification accuracy which built in 10-fold cross validation technique. Also, using fusion majority voting (the mode of the classifier output). The experimental results show that no classification technique is better than the other if used for all datasets, since the classification task is affected by the type of dataset. By using multi-classifiers fusion the results show that accuracy improved in three datasets out of four.**

Keywords- ***Breast Cancer; Classification techniques; Fusion; UCI; WEKA.***

I. INTRODUCTION

Nowadays, a main class of difficulties in medical learning includes the diagnosis of malady based upon many tests executed upon the patient. For this reason the use of classifier systems in medical diagnosis is growing in regularly. However, diverse artificial intelligence methods for classification also help reduce probable mistakes that can be done because of unskilled specialists and also offer medical data to be examined in shorter time and more detailed.

Breast cancer is a common disease among women. Accurate preoperative diagnosis of a breast lesion is significant for ideal treatment planning. To evade unnecessary patient suffering, it is essential to achieve the certain diagnosis without delay and with as few surgeries as possible [1]. In Egypt 18.9% of total cancer cases among the Egypt National Cancer Institute [2]. The success of treatment depends on an early recognition of breast cancer, which achieve more exact and less violent treatment options and mortality from breast cancer falls. Classification methods can achieve very high sensitivities up to 98% in classifying malignant lesions.

In later years, data-mining has become one of the most valued tools for and operating data in order to produce valuable information for decision-making [3]. Supervised learning, including classification is one of the most significant brands in data mining, with a recognized output variable in the dataset.

The implementations had been executed with a "WEKA" tool which stands for the Waikato Environment for Knowledge Analysis. A lot of papers about applying machine learning procedures for survivability analysis in the field of breast cancer diagnosis. Here are some examples:

A comparative study among three diverse datasets over different classifiers was introduced [4]. In Wisconsin Diagnosis Breast Cancer [WDBC] data set using SMO classifier only achieved the best results. In Wisconsin Prognosis Breast Cancer [WPBC] data set using a fusion between MLP, J48, SMO and IBK achieved the best results and In Wisconsin Breast Cancer [WBC] data set using a fusion between MLP and J48 with the principle component analysis [PCA] is achieved the best results.

A comparison between some of the open source data mining tools [5]. The type of dataset and the method the classification techniques were applied inside the toolkits affected the performance of the tools. The WEKA has achieved the best results, but we note that the comparison between tools in breast cancer results only show that Tanagra is the best tool not the WEKA as stated.

A comparative study between three classification techniques in WEKA [6]. The (SMO) has the best expectation accuracy, by removing the 16 instances with missing values from the dataset to build a different dataset with 683 cases. Class distribution Benign: 458 (65.5%) and Malignant: 241 (34.5%) as stated. But we note that the right percent after removing 14 benign and 2 malignant instances is benign: 444 (65%) and malignant: 239 (35%) hence percentage had been stated wrong.

A comparison between diverse classifiers on WBC dataset was introduced using two data mining tools [7] the classification technique, random tree outperforms has the highest accuracy rate, but we note that they don't state which accuracy data mining metrics was used.

The rest of this paper is prearranged like this: In sector II, Classification algorithms are discussed.in sector III datasets and evaluation principles are discussed. In sector IV a proposed model is shown.in sector V reports the experimental results. Finally, Sector VI introduces the conclusion.

II. DIFFERENT CLASSIFIERS TECHNIQUES

Bayesian networks are very smart for medical analytic systems, they can be executed to make extrapolations in cases where the input data is incomplete [8].

K-Nearest Neighbor (KNN) [9] classifies examples based on their similarity. It is one of the most popular techniques for pattern recognition. It is a kind of Lazy learning where the function is only approached locally and all computation is delayed until classification. An article is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of items for which the correct classification is known. In WEKA this classifier is called IBK

Decision tree J48 implements Quinlan's C4.5 algorithm [10] for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. It used for classification. J48 forms decision trees from a set of categorized training data using the theory of information entropy. Splitting the data into smaller subsets of each attribute can be used to make a decision.

Examining the normalized information gain, which outcomes from selecting an attribute on behalf of splitting the data. To make the decision, we use the attribute which has the maximum normalized information gain. Then the procedure repeats on the lesser subsets. The splitting process ends if all cases in a subset fit into the similar class. Then a leaf node is created in the decision tree telling to choose that class.

Support Vector Machine (SVM) [11] constructs one or more than one hyper plane in the high-dimensional feature space for classification. A hyper plane is applied to discriminate between data classes. If a hyper plane has the longest distance to the contiguous training data point of any class, then a valued separation can be reached; since a longer border signifies the classifier has a lesser generalization error. When classes overlap with other, a hyper plane is chosen to decrease the errors of data points along or across the borderline between classes; these points are referred to as the support points or support vectors

Sequential Minimal Optimization (SMO) is a new technique for training (SVMs) [12]. It is a simple and fast method for training an SVM. Solving double quadratic optimization problem by improving the least subset including two features at each repetition. It can be implemented simply and analytically. Training a support vector machine needs the solution of a very much quadratic programming optimization problems. SMO breakdowns this great quadratic programming problem into a sequence of minimum possible quadratic programming problems. These small quadratic programming problems are solved analytically, which saves a time in numerical quadratic programming optimization as an internal round. One of the advantages allows SMO to deal very big training sets that The total of memory required is linear in the training set size. One of the main differences between SMO and SVM that matrix computation in SVM balances in the middle of linear and cubic in the training set size of diverse test problems. While in SMO scales between linear and quadratic in the training set size. The computation time for SMo is subjected to SVM evaluation; hence SMO is fastest for linear SVMs and light datasets.

Multilayer Perceptron (MLP): it consists of 3 layers the input, hidden and output layers. The weighted sum of the inputs and bias term are conceded to the motivation level over a transmission function to produce the output. And the units are arranged in a layered feed-forward Neural Network (FFNN). The input layer consists of as several neurons as the number of features in a feature vector. Second layer, named hidden layer, has h number of Perceptions, where the value of h is determined by trial. The output layer has only one neuron representing either benign or malignant value (in case of diagnosis datasets). We used sigmoid activation function for hidden and output layers. The batch learning method is used for updating weights between different layers.

Random Forest: is a combined classifier that contains of several decision trees and productions the class that is the mode of the class's production of separate trees. Random forest introduces two bases of randomness: "Bagging" and "Random input vectors". Respectively a tree is grown by a bootstrap model of training data. At each node, greatest divided is selected from a random model of m_{try} variability rather than all variables [13].

III. Datasets and Evaluation principles

A. Dataset description

We copied the first three breast cancer databases from the UCI machine-learning repository [15], and the fourth dataset from the Lubiana University.

Table 1. Datasets Description

dataset	No of instances	No of attributes	Missing values
WBC	699	11	16
WDBC	569	32	-
WPBC	198	34	4
BCD	286	10	9

1) Confusion matrix

Evaluation method is based on the confusion matrix. The confusion matrix is an imagining implement usually used to show presentations of classifiers. It is used to display the relationships between real class attributes and predicted classes. The grade of efficiency of the classification task is calculated with the number of exact and unseemly classifications in each conceivable value of the variables being classified in the confusion matrix [14]

Table 2. Confusion matrix

		Predicted Class	
		Negative	Positive
Outcomes	Negative	TP	FN
	Positive	FP	TN

For instance, in a 2-class classification problem with two predefined classes (e.g., Positive diagnosis, negative diagnosis) the classified test cases are divided into four categories:

- True positives (TP) correctly classified as positive instances.
- True negatives (TN) correctly classified negative instances.
- False positives (FP) incorrectly classified negative instances
- False negatives (FN) incorrectly classified positive instances.

To evaluate classifier performance. We use accuracy term which is defined as the entire number of misclassified instances divided by the entire number of available instances for an assumed operational point of a classifier.

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

IV. RROPOSED METHODOLOGY

We proposed a robust method for discovering breast cancer using four different data sets based on data mining using WEKA as follows:

A. Data preprocessing

Pre-processing steps are applied to the data before classification:

1) Data Cleaning: eliminating or decreasing noise and the treatment of missing values. There are 16 instances in WBC and 4 instances in WPBC that contain a single missing attribute value, denoted by "?"And there are 9 instances in BCD that contain two missing values which typically substituted by the mean value for that attribute based on statistics.

2) Relevance Analysis: Statistical correlation analysis is used to discard the redundant features from further analysis.

The WBC, WPBC and WDBC have one irrelevant attribute named 'Sample code number' which has no effect in the classification process; therefore the attribute is not considered.

3) Data Transformation: The dataset is transformed by normalization, which is one of the greatest public tools used by inventors of automatic recognition classifications to get superior results. Data normalization hurries up training time by initialing the training procedure to reach feature within the same scale. The aim of normalization is to transform the attribute values to a small-scale range.

B. Single classification task

Classification is the procedure of determining a classifier that designates and distinguishes data classes so that it could expect the class of units or entities with unknown class label value. The assumed model depends on the training dataset analysis. The derivative model characterized in several procedures, such as simple classification rules, decision trees and another. Basically data classification is a two-stage process, in the initial stage; a classifier is built signifying a predefined set of notions or data classes. This is the training stage, where a classification technique builds the classifier by learning from a training dataset and their related class label columns or attributes. In next stage the model is used for prediction. In order to guess the predictive accuracy of the classifier an independent set of the training instances is used.

We evaluate the state of the art classification techniques which stated in recent published researches in this field to figure out the highest accuracy classifier's result with each dataset.

C. MULTI-CLASSIFIERS FUSION CLASSIFICATION TASK

A fusion of classifiers is combining multiple classifiers to get the highest accuracy. It is a set of classifiers whose separate predictions are united in some method to classify new instances. Combination ought to advance predictive accuracy. In WEKA the class for uniting classifiers is called Vote Different mixtures of probability guesses for classification are available.

1) According to results of single classification task, multi-classifiers fusion process starts using the classifier achieved best accuracy with other single classifiers predicting to improve accuracy.
2) Repeating the same process till the latest level of fusion, according to the number of single classifiers to pick the highest accuracy through all processes.

We propose our algorithm as follows.

- Import the Dataset.
- Replace missing values with the mean value.
- Normalize each variable of the data set, so that the values range from 0 to 1.
- Create a separate training set and testing set by haphazardly drawing out the data for training and for testing.
- Select and parameterize the learning procedure
- Perform the learning procedure
- Calculate the performance of the model on the test set

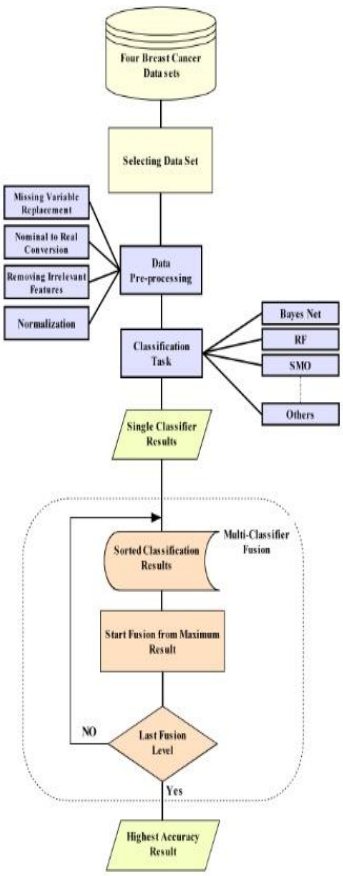

Figure 1.Proposed Breast Cancer diagnosis model.

V. EXPERIMENTAL RESULTS

To calculate the proposed model, two experiments were implemented. First one in the single classification task and second for multi-classifiers fusion task each of them using four datasets:

A. Single classification task

Table (3) shows the comparison of accuracies for the four single classifiers (Bays Net, SMO, RF and J48).

Using (WBC) dataset the highest accuracy at Bays Net (97.28%).

Using (WDBC) dataset:

The highest accuracy at SMO (97.72%).

Using (WPBC) dataset:

The highest accuracy at RF (78.29%).

Using Breast Cancer dataset (BCD) dataset Lubiana University the highest accuracy of J48 (76.63%).

TABLE 3. Single classifiers highest accuracies

	Bayes Net	SMO	RF	J48
WBC	97.28	96.99	95.85	95.14
WDBC	95.08	97.72	95.08	93.15
WPBC	74.79	75.79	78.29	74.74
BCD	75.5	74.83	75.2	76.63

B. Multi-classifiers fusion task

Table (4) shows the comparison of highest accuracies for different levels of multi-classifiers fusion task.

Using (WBC) dataset the accuracy of the fusion between the four classifiers Bays Net, MLP, SMO and J48 similar to fusion between the four classifiers Bays Net, MLP, SMO and IBK which achieves accuracy (97.57%).

Using (WDBC) dataset:

The accuracy of the fusion between the two classifiers SMO and Bays Net similar to fusion between the two classifiers SMO and MLP similar to fusion between the two classifiers SMO and IBK which achieves accuracy (97.72%).

Using (WPBC) dataset:

The accuracy of the fusion between the four classifiers Bays Net, MLP, RF and Zero R achieves accuracy (80.84%).

Using (BCD) dataset the accuracy of the fusion between the three classifiers Bays Net, RF and J48 achieves accuracy (78.67%).

TABLE 3. multi-classifers fusion highest accuracies

Fusion level	2nd	3rd	4th	5th	6th
WBC	97.28	97.28	97.57	97.28	97.28
WDBC	97.72	97.71	97.72	97.72	97.72
WPBC	77.84	77.29	80.84	77.82	79.82
BCD	77.64	78.67	78.33	78.66	78.29

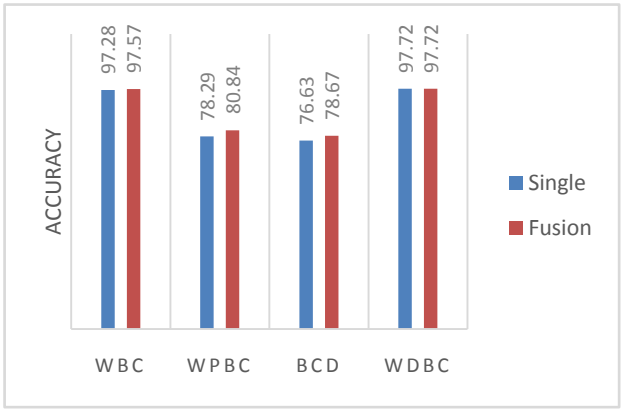

Figure 2. Comparison of the highest accuracies for single and fusion classifier

Also, we note that eliminating the instances which have missed values don't achieve better accuracies in both experiments.

VI. CONCLUSION

The experimental results illustrated that multi-classifiers fusion achieved better accuracy than single classifier. Each dataset has its own best single classifier and no pioneer single classifier for all datasets. In WPBC data set using a 4th level fusion better than the other fusion levels and single classifier too. In WBC data set using a 4th level fusion better than the other fusion levels and single classifier too. In BCD data set using a 3rd level fusion better than the other fusion levels and single classifier too. In WDBC data set Using a 2nd level fusion better than or equal the other fusion levels and has the same accuracy of the best single classifier. We concluded that, no multi-classifier fusion level or combination is pioneer for all datasets, and using more levels of fusion classifiers does not mean better accuracy. Bayes net classifier is unique in enhancement accuracy in multi-classifiers fusion for all datasets. The best enhancement in accuracy level using multi-classifiers fusion achieved in the prognostic datasets WPBC dataset by 2.55% and by 2.04% of BCD dataset more over any single classifier. Substituting the instances which have missed values with mean or mode values is better than eliminating it for any breast cancer dataset and achieve better accuracies

REFERENCES

[1] US Cancer Statistics Working Group. "United States cancer statistics: 1999–2010 incidence and mortality web-based report." Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute (2014).

[2] Elattar, Inas. "Breast cancer, magnitude of the problem: Egyptian society of surgical Oncology Conference, Sinai, Egypt;(30 March-1 April 2005)."

[3] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996).

[4] Salama, Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid. "Breast cancer diagnosis on three different datasets using multi-classifiers." .

[5] Wahbeh, Abdullah H., et al. "A comparison study between data mining tools over some classification methods." IJACSA International Journal of Advanced Computer Science and Applications, , pp. 18-26. 2011.

[6] Chaurasia, Vikas, and Saurabh Pal. "A Novel Approach for Breast Cancer Detection using Data Mining Techniques." International Journal ofr Innovative in Computer and Communication Engineering (2014).

[7] S. Syed Shajahaan, S. Shanthi and V. ManoChitra 'Application of Data Mining Techniques to Model Breast Cancer Data"International Journal of Emerging Technology and Advanced Engineering, 2008 Certified Journal, Volume 3, Issue 11, November 2013.

[8] Jyotirmay Gadewadikar, Ognjen Kuljaca1, Kwabena Agyepong, Erol Sarigul3, Yufeng Zheng and Ping Zhang, Exploring Bayesian networks for medical decision support in breast cancer detection, African Journal of Mathematics and Computer Science Research Vol. 3(10), pp. 225-231, October 2010.

[9] Angeline Christobel. Y, Dr. Sivaprakasam (2011). An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer Information Systems, Vol. 3, No. 2, 2011.

[10] Ross Quinlan, (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

[11] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.

[12] Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSR-TR-98- 14, Microsoft Research, 1998.

[13] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[14] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees and A. Zanasi, Discovering data mining from concept to implementation. Upper
Saddle River, N.J.: Prentice Hall, 1998.

[15] UCI Machine Learning Repository, Available at: http:// http://archive.ics.uci.edu/ml/, (Accessed 22 january 2015).

[16] Wolberg, William H., and Olvi L. Mangasarian. "Multisurface method of pattern separation for medical diagnosis applied to breast cytology." Proceedings of the national academy of sciences, 1990: pp. 9193-9196.