International Journal of Mathematics and
Computer Science, **14**(2019), no. 4, 935–952

Performance Analysis on Three Breast Cancer Datasets using Ensemble Classifiers Techniques

Soumya Arach, Halima Bouden

Equipe de recherche modélisation et théorie de l'information
MTI
Tétouan, Morocco

email: soumya.arach@gmail.com, halima@gmail.com

(Received June 29, 2019, Accepted August 5, 2019)

Abstract

In this paper, we present a comparison between different classifiers or multi-classifiers fusion with respect to accuracy in discovering breast cancer for three different datasets using a classification accuracy and confusion matrix based on a 10-fold cross validation method. We present an implementation among various classification techniques which represent the most known algorithms in this field on three different datasets of breast cancer. To get the most suitable results, we had referred to attribute selection using Correlation and Recursive Feature Elimination (RFE) in three datasets to get a set of correlated features.

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 and feature selection methods did not have a strong influence on WDBC and WPBC datasets, but in WBC the selected attributes (Uniformity of Cell Size, Mitoses, Clump thickness, Bare Nuclei, Single Epithelial cell size, Marginal adhesion, Bland Chromatin and Class) improved the accuracy.

Key words and phrases: Breast cancer, classification techniques, fusion, performance Analysis.
AMS (MOS) Subject Classification: 68T10.
ISSN 1814-0432, 2019, http://ijmcs.future-in-tech.net

1 Introduction

In Morocco, breast and cervical cancer are real public health problems; Not only they represent the most common cancers in women (36.1% for breast cancer and 12.8% for cervix cancer) but they also cause a significant number of deaths because of the delay in their diagnosis [1]. The age of breast cancer affection in Morocco and, more generally, Arab countries is ten years compared to foreign countries as the disease targets women, on average, at the age of 30 compared to 45 years in European countries, for instance.

Breast cancer is a disease for which there is currently no means of primary prevention since the etiology of this cancer is not completely elucidated. Nevertheless, breast cancer can be curable or at least have a better prognosis when detected early. Its early detection allows to establish a therapeutic conservative surgery and allows to improve the prognosis of cancer. The means of diagnosis are based on clinical examination and radiographic breast exploration by mammography sometimes associated with ultrasound.

Studies have shown that the implementation of an early detection program over a span of several years can reduce the mortality rate of this disease by 25% [2].

Data mining approaches in medical domains is increasing rapidly due to the improvement and effectiveness of these approaches in classification and prediction systems especially in helping medical practitioners in their decision-making [3].

In addition to its importance in finding ways to improve patient outcomes, this reduces the cost of medicine and helps in enhancing clinical studies. Supervised learning, including classification, is one of the most significant brands in data mining with a recognized output variable in the dataset.

Many experiments are performed on medical and non-medical datasets using multiple classifiers and feature selection techniques. A good amount of research on breast cancer datasets is found in the literature with good classification accuracy.

In [4], the performance criterion of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART are compared, to find the best classifier in breast cancer datasets (WBC and Breast tissue). The experimental result shows that the SVM-RBF kernel is more accurate than other classifiers.

A comparative study among three diverse datasets over different classifiers was introduced in [5]. In Wisconsin Diagnosis Breast Cancer [WDBC] dataset using SMO classifier only achieved the best results. In Wisconsin

Prognosis Breast Cancer [WPBC] dataset 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.

In [6], a comparison between diverse classifiers on WBC dataset was introduced using two data mining tools, the classification technique, random tree outperforms has the highest accuracy rate. We note, however, that the authors do not state which accuracy data mining metrics was used.

In [7], SVM proves to be the most accurate classifier, when the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) are compared.

In [8], the neural network classifier is used on WPBC dataset.

A comparison between some of the open source data mining tools [9] show that 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.

In [10], three classifications algorithms neural networks, SVM, and decision trees (J48) are performed. By examining confusion matrix and error rates, decision tree (J48) has the highest accuracy rate.

The rest of this paper is organized as follows: In section 2, Classification algorithms are discussed. Section 3 is about dataset description. In section 4, evaluation principles are discussed. Section 5, is about Feature Extraction and Selection. A proposed model is shown in section 6. In section 7, we report our experimental results. We conclude the paper with section 8.

2 Classifiers Techniques

The Multilayer Perceptrons (MLPs), are supervised learning classifiers that consist of an input layer, an output layer, and one or more hidden layers that extract useful information during learning and assign modifiable weighting coefficients to components of the input layer. MLP, a feed-forward backpropagation network, is the most frequently used neural network technique in pattern recognition [11]. The weighted sum of the inputs and bias term are conceded to the motivation level over a transmission function to produce the output. 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. A 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 a sigmoid activation function for hidden and output layers. The batch learning method is used for updating weights between different layers [12].

K-Nearest Neighbor (KNN) classification [7] classifies instances based on their similarity. It is one of the most popular algorithms for pattern recognition. It is a type of Lazy learning where the function is only approximated locally and all computation is deferred until classification. An object is classified by a majority of its neighbors. K is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known. In WEKA this classifier is called IBK.

Decision tree J48 implements Quinlan's C4.5 algorithm [13] for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. It is 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. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Moreover, it provides an option for pruning trees after creation.

Random Forest is a combined classifier that contains several decision trees and produces 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". A tree is grown by a bootstrap model of training data. At each node, greatest divided is selected from a random model of mtry variability rather than all variables [14].

Support Vector Machine (SVM) is introduced by Vapnik et al. [15] and is a very powerful method that has been used in a wide variety of applications. The basic concept in SVM is the hyper plane classifier or linear separability. Two basic ideas are applied to achieve linear separability, SVM: margin maximization and kernels; that is, mapping input space to a higher-dimension space (or feature space).

SVM projects the input data into a kernel space. Then it builds a linear model in this kernel space. A classification SVM model attempts to separate the target classes with the widest possible margin. A regression SVM model tries to find a continuous function such that the maximum number of data points lie within an epsilon-wide tube around it. Different types of kernels and different kernel parameter choices can produce a variety of decision boundaries (classification) or function approximators (regression). In WEKA, this classifier is called SMO.

Sequential Minimal Optimization (SMO) is a new technique for training (SVMs) [16]. It is a simple and fast method for training an SVM solving a 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 quadratic programming optimization problem.

Naive Bayes (NB) classifier is a probabilistic classifier based on Bayes theorem. Rather than predictions, the Naïve Bayes classifier produces probability estimates. For each class value, we estimate the probability that a given instance belongs to that class. Requiring a small amount of training data to estimate the parameters necessary for classification is the advantage of the Naive Bayes classifier. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence [17].

3 Dataset Description

We use tree datasets (WBC), (WDBC), (WPBC) from UCI Machine Learning Repository [18]. A brief description of these datasets is presented in table 1. Each dataset consists of some classification patterns or instances with a set of numerical features or attributes.

1. TABLE I. TABLE . DESCRIPTION OF THE BREAST CANCER DATASETS

Dataset	***No of instances***	***No of attributes***	***Missing values***
WBC	699	11	16
WDBC	569	32	-
WPBC	198	34	4

4 Evaluation Principles

4.1 Confusion Matrix

The Evaluation method is based on the confusion matrix which 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 at each conceivable value of the variables.

1. TABLE II. TABLE . Confusion matrix

		Predicted	
		Negative	***Positive***
Actual	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 an 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.

$$AC = \frac{TP + TN}{FP + FN + TP + TN}...............(1)$$

5 Feature Extraction and Selection

Too many features pose the problem of overfitting the model. We generally want to restrict the features in our models to those that are most relevant to the response variable we want to predict. Using as few features as possible will also reduce the complexity of our models which results in less time and computer power not o mention that it is easier to understand.
There are several ways to identify how much each feature contributes to the model and to restrict the number of selected features. Here, we are going to examine the effect of feature selection via

Correlation,

Recursive Feature Elimination (RFE)

Additionally, we want to know how different data properties affect the influence of these feature selection methods on the outcome. For that, we use three breast cancer datasets, one of which has few features (WBC); the other two are larger.

Based on our comparisons of the correlation method and RFE, we conclude that:
Removing highly correlated features is not a generally suitable method.
The correlation method operates regardless of feature importance. In WDB dataset, the features with the highest importance were also flagged as highly correlated. Correlation models performed worst in all three datasets. RFE tends to include features with high importance, but that alone is not a good indicator for whether several features will work well in combination when predicting an outcome.
Our conclusions are of course not to be generalized to any data since there are many more feature selection methods and we are only looking at a limited number of datasets and only at their influence on six models (RF, MLP, SMO, BN, IBK, and J48).
WBC dataset was small with only 9 features; here, removing highly correlated features was the least successful selection method. RFE improved the predictions compared to no feature selection.
In WDBC and WPBC Datasets, different feature selection methods did not have a strong influence.

6 Proposed Breast Cancer Diagnosis Model

We propose a method for discovering breast cancer using three different data sets based on data mining using WEKA. Fig. 1 shows the diagram of the Proposed Breast Cancer Diagnosis Model. It consists of three phases namely: data preprocessing, single classification and multi-classifiers fusion classification task.

6.1 Data Preprocessing

Preprocessing 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 ”?”.
2) Feature extraction and Relevance Analysis: Statistical correlation analysis is used to discard the redundant features from further analysis. Feature extraction considers the whole information content and maps the useful information content into a lower dimensional feature space. Feature selection is based on omitting those features from the available measurements which

do not contribute to class separability; that is, redundant and irrelevant features are ignored. In the classification step, different classifiers are applied to get the best result of diagnosing and prognosing the tumor.

6.2 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 second stage the model is used for prediction. In order to guess the fusion level 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.

6.3 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 a single classification task, multiclassifiers 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.

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.

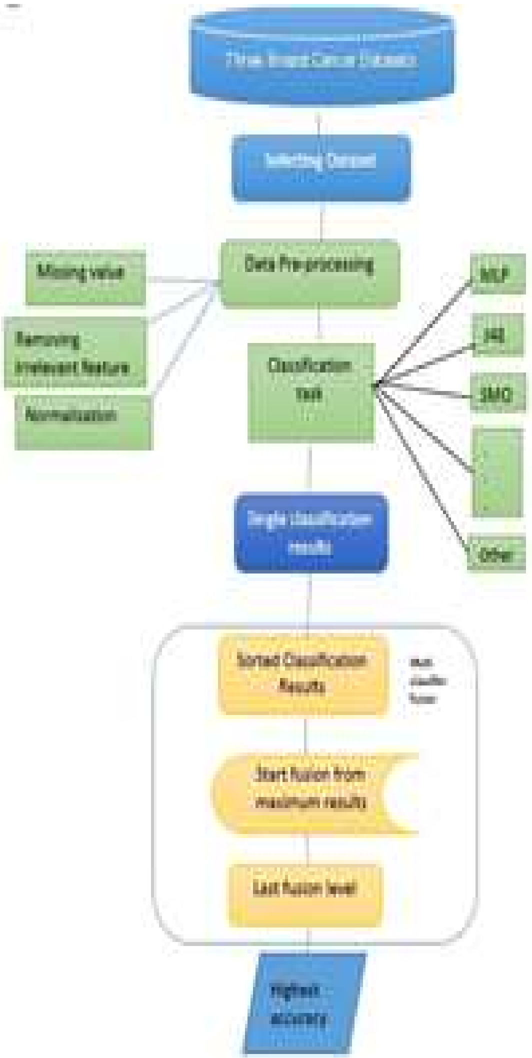

Fig.1. Proposed breast cancer diagnosis model

7 Experimental Results

To calculate the proposed model, two experiments were implemented: one in the single classification task and the other for multi-classifiers fusion task. Each experiment uses three datasets:

7.1 Experiment (1) using Wisconsin Breast Cancer (WBC) dataset:

Fig. 2 shows the comparison of accuracies for the six classifiers (BN, MLP, J48, SMO ,IBK and RF) based on a 10-fold cross validation as a test method. The accuracy of BN (97.28%) is the best classifier and the accuracy obtained by SMO is better than that produced by RF, IBK, MLP and J48.

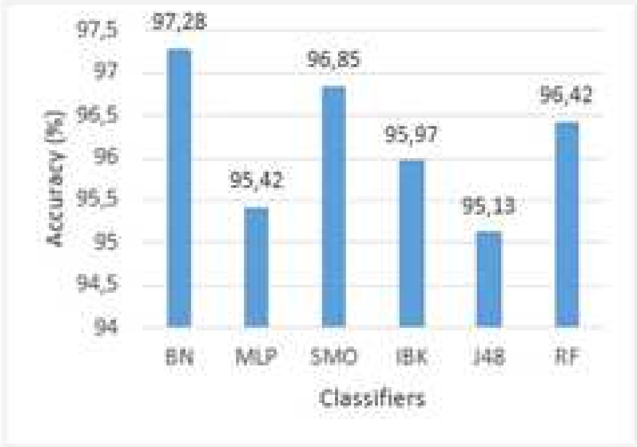

Fig.2. Single classifier in WBC

Fig. 3 shows the result of combining BN and each of the other classifiers. The fusion between BN and RF achieves the best accuracy (97.42%).

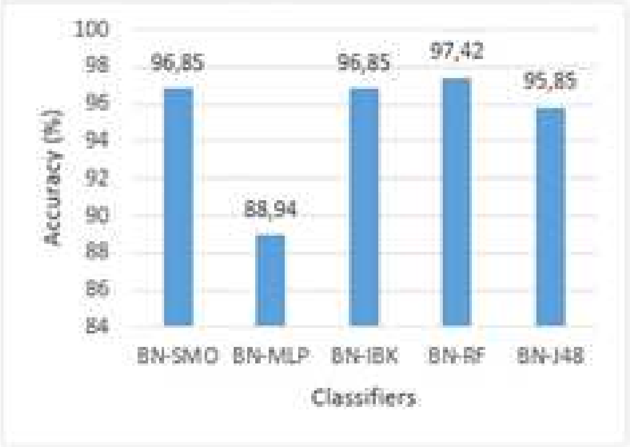

Fig.3. Fusion of two classifiers in WBC

Fig. 4 shows the result of fusion between the three classifiers BN+RF+SMO, BN+RF+MLP and BN+RF+J48 and BN+RF+IBK. It can be noticed that the recognition accuracy decrease to 97.13%.

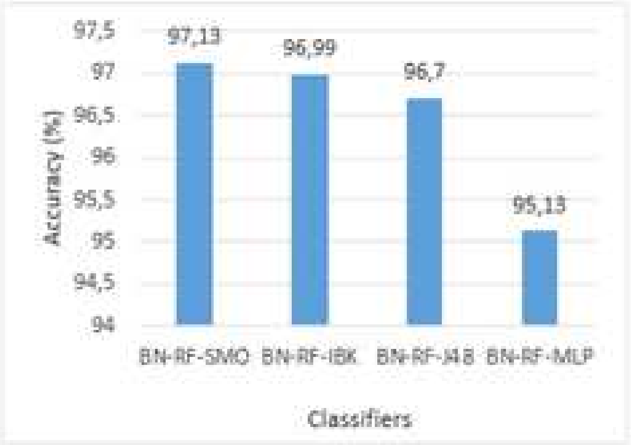

Fig. 4. Fusion of three classifiers in WBC

Fig. 5 shows that the fusion between the four classifiers BN,RF,SMO and J48 achieves accuracy (97.56%).This fusion is better than single classifiers, fusion of 2 classifiers and fusion of 3 classifiers.

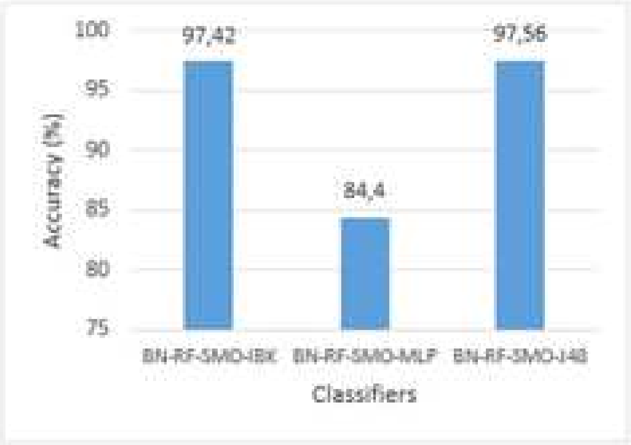

Fig. 5. Fusion of four classifiers in WBC

When using features selection on WBC dataset, with PCA as a select attribute in the third level of fusion, the accuracy increases (97.99%) with the fusion between BN, RF, SMO and IBK as shown in Fig. 6

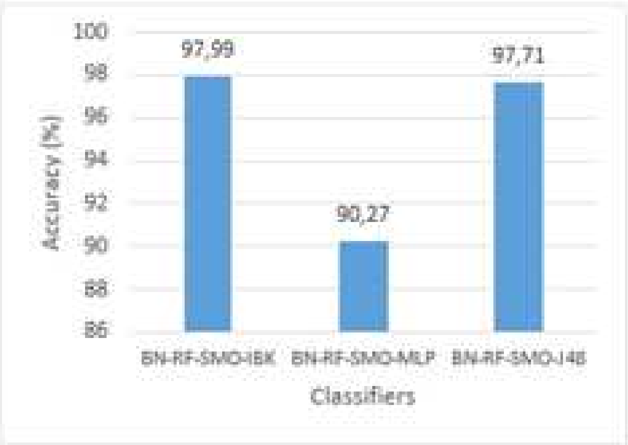

Fig. 6. Fusion of four classifiers in WBC with PCA

7.2 Experiment (2) using Wisconsin Diagnosis Breast Cancer (WDBC) dataset without feature selection:

Fig. 7 shows the comparison of accuracies for the six classifiers (BN, MLP, J48, SMO, RF and IBK) based on cross validation of 10-fold as a test method. SMO is more accurate than other classifiers (97.71%).

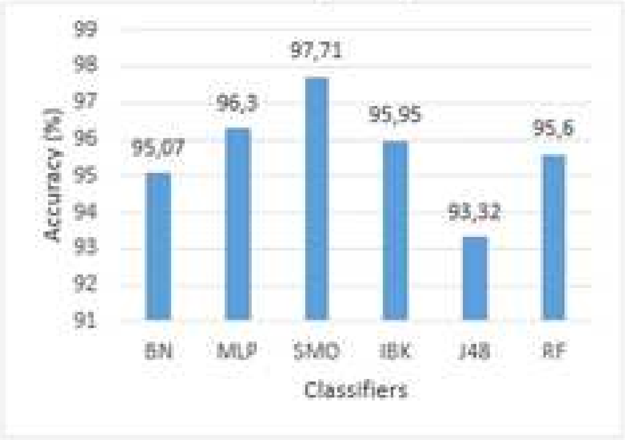

Fig. 7. Single classifier in WDBC

Fig. 8 shows that fusion between SMO and each of other classifiers led to the following results: the fusion between SMO and MLP, SMO and IBK, SMO and BN, SMO and RF gives the same highest accuracy as of SMO alone. 96.83% is accuracy of the fusion between SMO and J48.

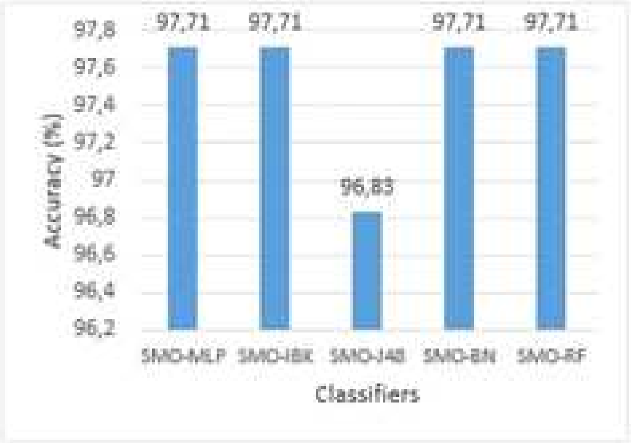

Fig. 8. Fusion of two classifiers in WDBC

Fig. 9 shows that after we try to fuse SMO with each two of the other classifiers, the accuracy decreases.

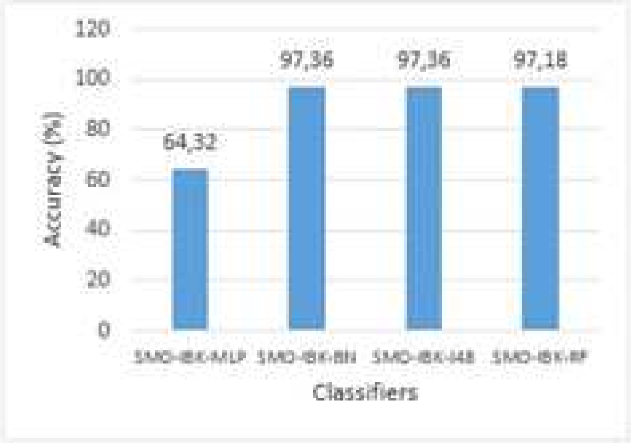

Fig. 9. Fusion of three classifiers in WDBC

Fig. 10 shows that the fusion between SMO, IBK and NB with MLP increases the accuracy slightly but still lower than the highest accuracy in single classifiers and fusion of two classifiers.

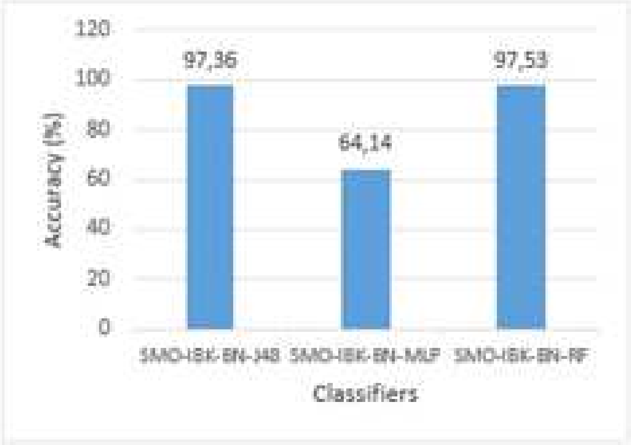

Fig.10. Fusion of four classifiers in WDBC

C. Experiment (3) using Wisconsin Prognosis Breast Cancer (WPBC) dataset without feature selection:

Fig. 11 shows the comparison of accuracies for the six classifiers (NB, MLP, J48, SMO, RF and IBK) based on 10-fold cross validation as a test method. The accuracy of RF (78.28%) is the highest.

Accuracy of BN and SMO is better than other classifiers and they are the same (75.75%).

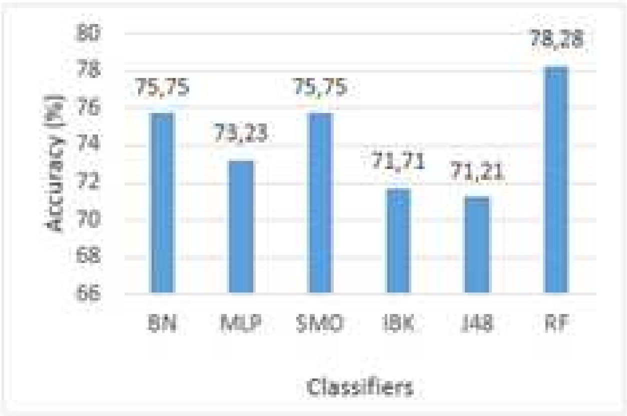

Fig.11. Single classifier in WPBC

Fig. 12 shows that the fusion between RF and each of other classifiers led to the following results:

Fusion between RF and BN gives the highest accuracy (79.79%) followed by fusion between RF and MLP (77.27%). The lower accuracy is given by fusion between RF and SMO.

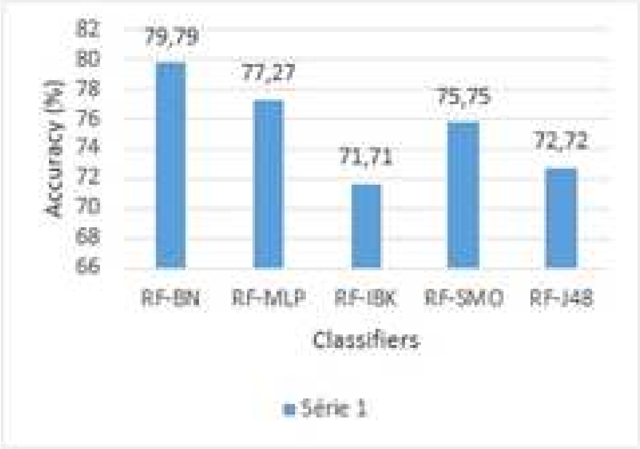

Fig. 12. Fusion of two classifiers in WPBC

Fig. 13 shows that the fusion between RF, BN and MLP achieves the best accuracy of (76.26%), but it is lower than accuracy of single classification and fusion between two classifiers.

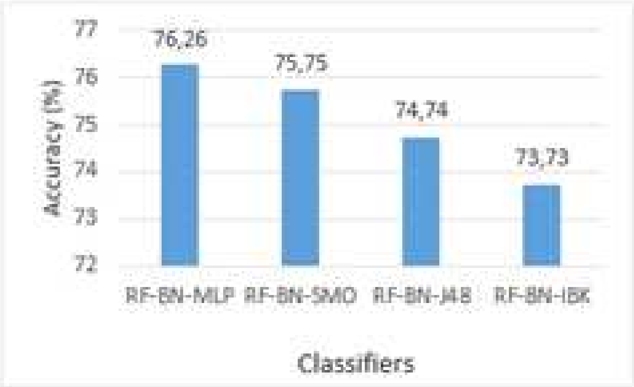

Fig. 13. Fusion of three classifiers in WPBC

Fig. 14 shows that the fusion between RF, BN, MLP and

SMO is superior to the other classifiers. It achieves an accuracy of (79.26%).

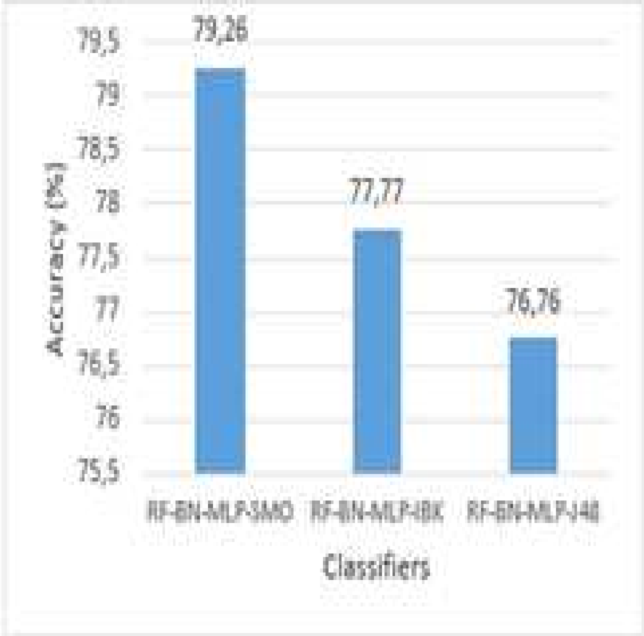

Fig. 14. Fusion of four classifiers in WPBC

8 Conclusion

The experimental results in WBC dataset show that the fusion between BN, RF, SMO and J48 is superior to the other classifiers, and when using PCA feature selection on the fusion between BN, RF, SMO and IBK the accuracy increases to 97.99%. On the other hand, WDBC dataset shows that using single classifiers (SMO) or using fusion of SMO and MLP or SMO and IBK or SMO and BN is better than other classifiers. Finally, the fusion of RF and BN is superior to the other classifiers in WPBC dataset.

Acknowledgement. S. Arach acknowledges financial support for this research from the "Centre National pour la Recherche Scientifique et Technique" CNRST, Morocco.

References

[1] http://www.contrelecancer.ma/fr/

[2] Guide de Détection Précoce des Cancers du Sein et du Col de L'utérus, http://www.contrelecancer.ma

[3] Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, From data mining to knowledge discovery in databases, AI magazine, **17,** no. 3, (1996).

[4] S. Aruna et al., Knowledge based analysis of various statistical tools in detecting breast cancer, 2011.

[5] Gouda I. Salama, M. B. Abdelhalim, Magdy Abdelghany Zeid, Breast cancer diagnosis on three different datasets using multi-classifiers.

[6] S. Syed Shajahaan, S. Shanthi, V. ManoChitra, Application of Data Mining Techniques to Model Breast Cancer Data, International Journal of Emerging Technology and Advanced Engineering, **3,** Issue 11, November 2013.

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

[8] Vaibhav Narayan Chunekar, Hemant P. Ambulgekar, Approach of Neural Network to Diagnose Breast Cancer on three different Data Set, International Conference on Advances in Recent Technologies in Communication and Computing, (2009).

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

[10] S. Arach, H. Bouden, Learning Experiences Using Neural Networks and Support Vector Machine (SVM), International Journal of New Computer Architectures and their Applications (IJNCAA), (2017), 37–44.

[11] R. O. Duda, P. E. Hart, Pattern Classification and Scene Analysis Wiley-Interscience Publication, New York, 1973.

[12] C. M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press,1999.

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

[14] Leo Breiman, Random forests, Machine learning **45,** No. 1, (2001), 5–32.

[15] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

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

[17] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kauffman Publishers, 2000.

[18] A. Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, (2010).