Original Research Paper

Volume - 13 | Issue - 10 | October - 2023 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Machine Learning

BREAST CANCER DIAGNOSIS MODEL USING MULTI-CLASSIFIERS FUSION

Halima Bouden

Somya Arach

ABSTRACT In this paper, we present a comparison between different classi-fiers, decision tree (J48), Multi-Layer Perception (MLP), Bayes net (BN), Sequential Minimal Optimization (SMO), Random forest (RF) and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method. We present a fusion at classification level between these classifiers to get the most suitable multi-classifier approach for each data set. 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.

KEYWORDS : Breast Cancer; Performance; WEKA; Classification techniques; Fusion; UCI;

I. INTRODUCTION

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

Breast cancer is a general 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 on a psychological and medical level, and allows to improve the prognosis of cancer. The means of diagnosis are based on the clinical examination and Radiographic breast exploration by mammography sometimes associated with ultrasound.

Studies have shown that the implementation of an early detection program, during 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 effectiveness of these approaches to 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, reduce the cost of medicine, and help 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 literature. Many of them show 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 SVM-RBF kernel is more accurate than other classifiers.

A comparative study among three diverse datasets over different classifiers was introduced [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.

A comparison in [6] between diverse classifiers on WBC dataset was introduced using two data mining tools 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.

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]. 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 datasets and evaluation principles are discussed. A proposed model is shown in section 4. Section 5 Reports the experimental results. Section 6 introduces the conclusion of this paper.

II. 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 is a feed-forward back-propagation 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. 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 or updating weights between different layers [12].

K-Nearest Neighbor (KNN) classification [13] 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 [14] 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. J48 can handle both continuous and discrete attributes, training data with missing attribute values and attributes with differing costs. Further, it provides an option for pruning trees after creation.

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 mtry variability rather than all variables [22].

Support Vector Machine (SVM) is introduced by Vapnik et al. [15] it is a very powerful method that has been applied 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 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) [17]. 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.

Naive Bayes (NB) classifier is a probabilistic classifier based on the Bayes theorem. Rather than predictions, the Naïve Bayes classifier produces probability estimates. For each class value, they 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 [16].

III. Dataset Description

We used tree datasets: (WBC), (WDBC), (WPBC) from UCI Machine Learning Repository [21]. 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.

Table 1: 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

IV. Evaluation principle

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.

Table 2. 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 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.

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

V. Proposed Breast Cancer Diagnosis Model

We proposed 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.

A. 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.

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 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.

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, 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.

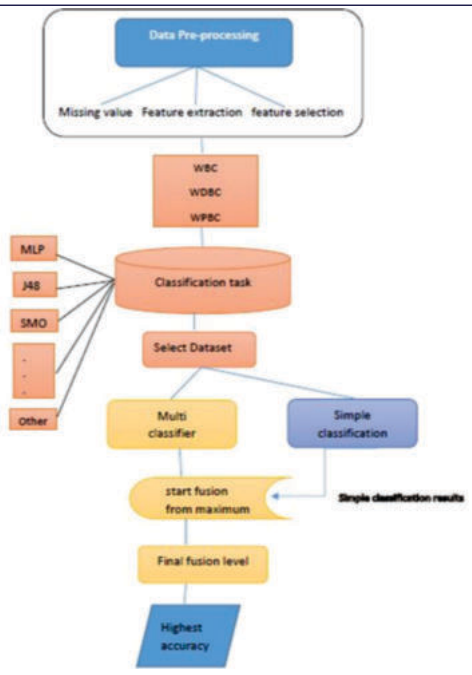

Figure1: Proposed Breast Cancer Diagnosis Model

VI. 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 three datasets:

A. *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 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.

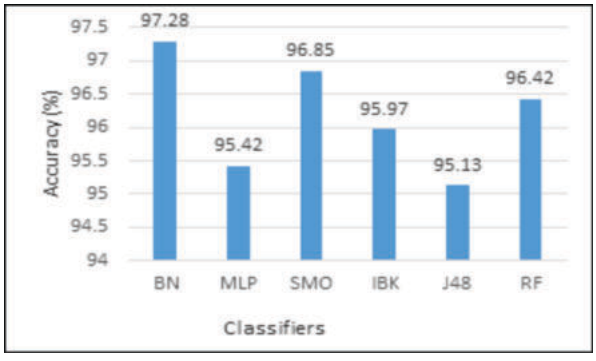

Figure 2: Single classifier in WBC

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%).**

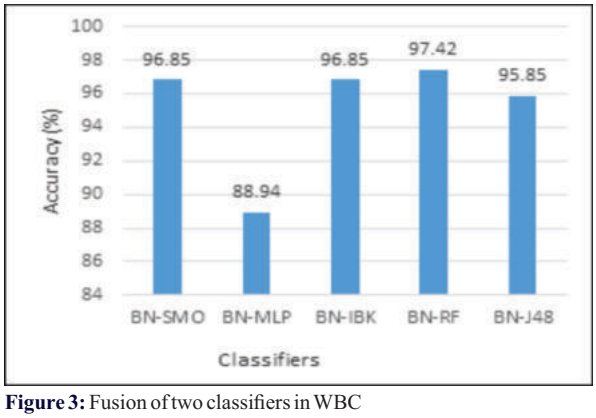

Figure 3: Fusion of two classifiers in WBC

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%.

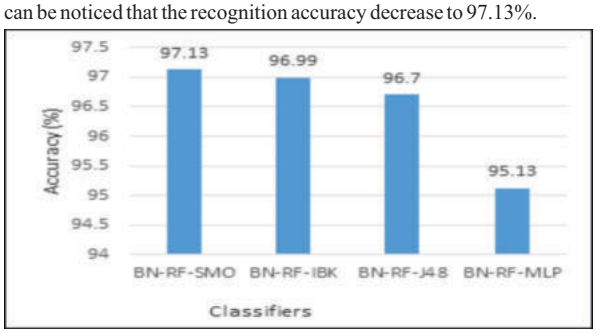

Figure 4. Fusion of three classifiers in WBC

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.

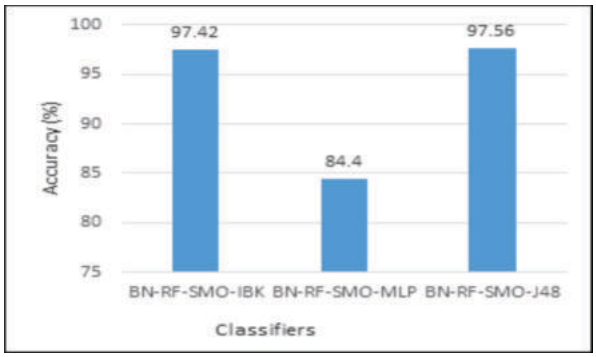

Figure 5: Fusion of four classifiers in WBC

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

Fig. 6 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%).

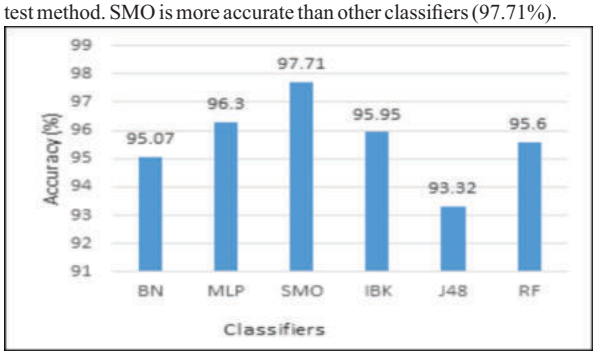

Figure 6: Single classifier in WDBC

Fig. 7 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.

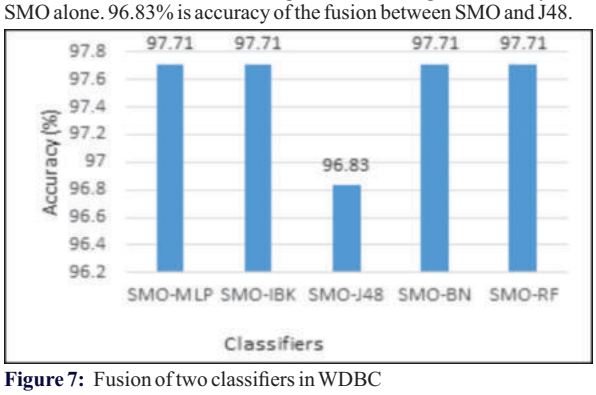

Figure 7: Fusion of two classifiers in WDBC

Fig. 8 shows that after we try to fuse SMO with each two of the other classifiers, the accuracy decreases.

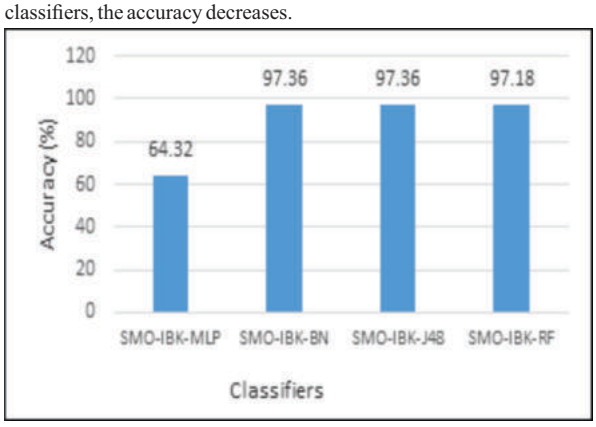

Figure 8: Fusion of three classifiers in WDBC

Fig. 9 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.

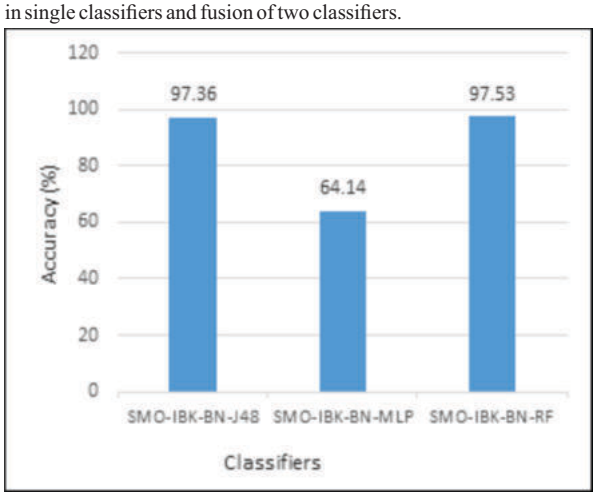

Figure 9: Fusion of four classifiers in WDBC

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

Fig. 10 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%).

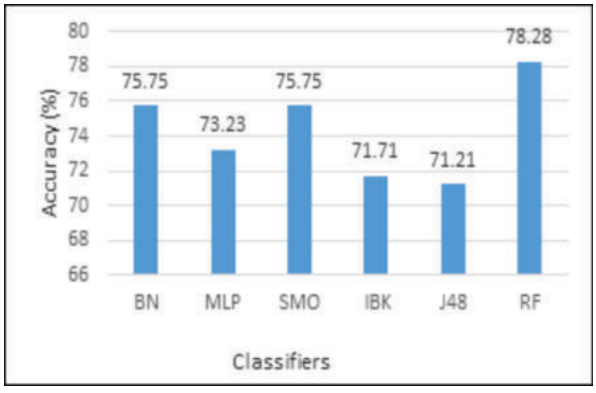

Figure 10: Single classifier in WPBC

Fig. 11 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.

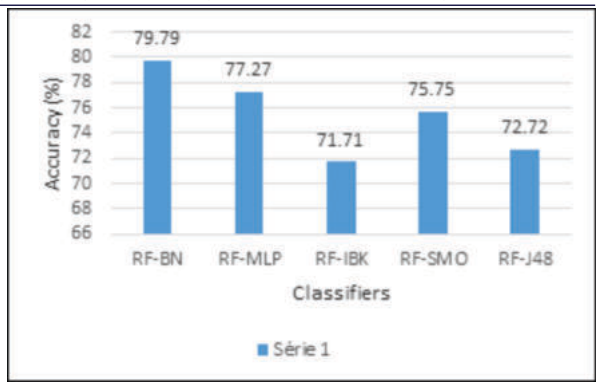

Figure 11. Fusion of two classifiers in WPBC

Fig. 12 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.

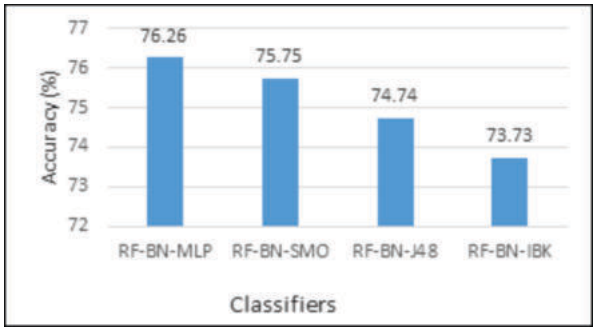

Figure 12: Fusion of three classifiers in WPBC

Fig. 13 shows that the fusion between RF, BN, MLP and SMO is superior to the other classifiers. It achieves accuracy of (79.26%).

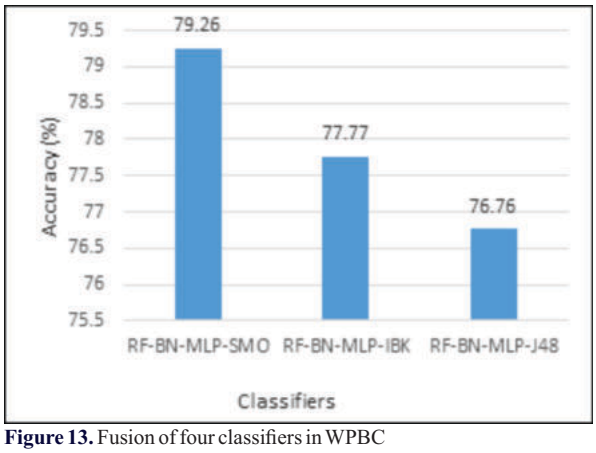

Figure 13. Fusion of four classifiers in WPBC

VII. CONCLUSION

The experimental results in WBC dataset show that the fusion between MLP and J48 classifiers with features selection (PCA) is superior to the other classifiers. On the other hand WDBC dataset shows that using single classifiers (SMO) or using fusion of SMO and MLP or SMO and IBK is better than other classifiers. Finally, the fusion of MLP, J48, SMO and IBK is superior to the other classifiers in WPBC dataset.

REFERENCES

[1] https://www.contrelecancer.ma/fr/detection_precoce_action (3-3-2019).
[2] https://www.contrelecancer.ma/site_media/uploaded_files/Guide_de_detection_pre%C3%BCcoce_des_cancers_du_sein_et_du_col_de_lute%C3%BCrus.pdf
[3] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." AI magazine 17.3 (1996).
[4] S. Aruna et al. (2011). Knowledge based analysis of various statistical tools in detecting breast cancer.
[5] Salama, Gouda I., M. B. Abdelhalim, and Magdy Abdelghany Zeid. "Breast cancer diagnosis on three different datasets using multi-classifiers."
[6] 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.
[7] 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.
[8] Vaibhav Narayan Chunekar, Hemant P. Ambulgekar (2009). Approach of Neural Network to Diagnose Breast Cancer on three different Data Set. 2009 International Conference on Advances in Recent Technologies in Communication and Computing.
[9] 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.
[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) pp:37-44. 2017
[11] Duda, R.O., Hart, P.E.: "Pattern Classification and Scene Analysis", In: Wiley-Interscience Publication, New York (1973)
[12] Bishop, C.M.: "Neural Networks for Pattern Recognition". Oxford University Press,New York (1999).
[13] 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.
[14] Ross Quinlan, (1993) C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.
[15] Vapnik, V.N., The Nature of Statistical Learning Theory, 1st ed., Springer-Verlag,New York, 1995.
[16] J. Han and M. Kamber,"Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2000.
[17] Platt, J.C.: Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSRTR-98- 14, Microsoft Research, 1998.
[18] L. Breiman, J. Friedman., R. Olshen, C. Stone, (1984), Classification and Regression Trees. Wadsworth, Belmont, CA.
[19] D.Steinberg., and P.L. Colla, (1995) "CART: Tree-Structured Nonparametric Data Analysis", Salford Systems: SanDiego, CA.
[20] D.Steinberg., and P.L. Colla, (1997) "CART-Classification and Regression Trees", Salford Systems: San Diego, CA.
[21] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[22] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.