# New Experiments on Ensembles of MF

Carlos Hernández-Espinosa, Joaquín Torres-Sospedra
and Mercedes Fernández-Redondo

Universidad Jaume I. ICC Department. Avda Vicente
Sos Baynat s/n. CP 12071 Castellon. Spain.

{espinosa, redondo}@icc.uji.es

## Abstract

As shown in the bibliography, training an ensemble of networks is an interesting way to improve the performance. However there are several methods to construct the ensemble. In this paper, some new results are presented in a comparison of twenty different methods. Ensembles of 3, 9, 20 and 40 networks have been trained to show results in a wide spectrum of values. The results show that the improvement in performance above 9 networks in the ensemble depends on the method but it is usually low. Also, the best method for a ensemble of 3 networks is called “CVC version 2” and uses a partition of the training and cross-validation sets according to the more usual method CVC. For the case of 9 and 20 networks the best method is “Conservative Boosting”, a modification of “Adaboost”. And finally for 40 networks the best method is “Cels”.

**Keyword**: Multilayer Perceptron, Backpropagation, Ensembles of Neural Networks.

## I. Introduction

Probably the most important property of a neural network is the generalization capability, i.e., the ability to correctly respond to new inputs not present in the training set.
One technique to increase the generalization capability with respect to a single neural network consist on training an ensemble of neural networks, i.e., to train a set of neural networks with different weight initialization or properties and combine the outputs in a suitable manner.
It is clear from the bibliography that this procedure in general increases the generalization capability [1, 2].
The two key factors to design an ensemble are how to train the individual networks to obtain not correlated outputs in the different networks and how to combine the different outputs to give a single output.
Among the methods of combining the outputs, the two most popular are *voting* and *output averaging* [3]. In this paper we will normally use *output averaging*. This procedure gives a reasonable performance [5] and it has not problems of ties.
In the other aspect, nowadays, there are several different methods in the bibliography to train the individual networks and construct the ensemble [1-3].

However, there is a lack of comparison among the different methods. One comparison can be found in [4], it is a previous work developed by our research group. In paper [4], eleven different methods are compared. The conclusions of this paper are that the method called “Decorrelated” is the best among the eleven methods analyzed.
Now, more complete results are presented by including nine new methods, so the number of methods is increased in the comparison to a total of twenty. These new empirical results are quite interesting because one of the new methods analyzed in this paper, called “conservative boosting” seems to have the best performance in several situations.

## II. Theory

In this section we briefly review the new nine ensemble methods introduced in this paper for comparison. The description of the rest of methods denoted by “Simple Ensemble”, “Ola”, “Evol”, “Decorrelated”, “Decorrelated2”, “CVC”, “Cels”, “Boosting”, “Bag_Noise” and “Adaboost”, can be found our previous reference [4], and in the references cited there.

### A. *CVC version 2*

In the usual CVC the available data is divided in training, cross-validation and testing subsets. After that, the data for training is divided by the number of networks giving several subsets. Then, one different subset is omitted for each network and the network is training with the rest of subsets.
The version 2 of CVC included in this paper is used in reference [5]. The data for training and cross-validation is jointed in one set and with this jointed set the usual division of CVC is performed. In this case, one subset is omitted for each network and the omitted subset is used for cross-validation.

### B. *Aveboost*

Aveboost is the abbreviation of Average Boosting. This method was proposed in reference [6] as a variation of Adaboost. In Adaboost, it is calculated a probability for each pattern of being included in the training set for the following network. In this case a weighted adaptation of the probabilities is performed.
The method of combination of outputs employed was not averaging in this case. Aveboost has the same method of combination of Adaboost. There are different weights in the combination according to the performance of the different networks.

### C. *TCA, Total Corrective Adaboost*

It was also proposed in [6] and it is another variation of Adaboost. In this case the calculation of the probability distribution for each network is treated as an optimization problem and an iterative process is performed.

### D. *Aggressive Boosting*

Aggressive Boosting is a variation of Adaboost. It is reviewed in [7]. In this case it is used a common step to modify the probabilities of a pattern for being included in the next training set.

### E. *Conservative Boosting*

It is another variation of Adaboost reviewed in [7]. It is a y technique similar to Aggressive Boosting. In this case the probability of the well classified patterns is decreased and the probability of wrong classified patterns is kept unchanged.

### F. *ArcX4*

It is another variation of Boosting. It was proposed and studied in reference [8]. The method selects training patterns according to a distribution, and the probability of the pattern depend on the number of times the pattern was not correctly classified by the previous networks.

The combination procedure proposed in the reference is the mean average. In our experiments we have used this procedure and also voting.

### G. *EENCL Evolutionary Ensemble with Negative Correlation*

This method is pro-posed in reference [9]. The ensemble is build as a population of a genetic algorithm. The fitness function is selected to consider the precision in the classification of the individual networks and also to penalize the correlation among the different networks in the ensemble.

Two variations of the method are used, EENCL UG and MG in UG we select as the networks of the ensemble the last population of networks, in the case of MG we select the best population according to the criterion of the mean squared error of cross-validation.

## III. Experimental Results

We have applied the twenty ensemble methods to ten different classification problems. They are from the UCI repository of machine learning databases. Their names are Cardiac Arrhythmia Database (Aritm), Dermatology Database (Derma), Protein Location Sites (Ecoli), Solar Flares Database (Flare), Image Segmentation Database (Image), Johns Hopkins University Ionosphere Database (Ionos), Pima Indians Diabetes (Pima), Haberman's survival data (Survi), Vowel Recognition (Vowel) and Wisconsin Breast Cancer Database (Wdbc).

We have constructed ensembles of a wide number of networks, in particular 3, 9, 20 and 40 networks in the ensemble. We repeated the process of training the ensembles of 3, 9, 20 and 40 networks ten times for ten different partitions of data in training, cross-validation and test.

With this procedure we can obtain a mean performance of the ensemble for each database (the mean of the ten ensembles) and an error in the performance calculated by standard error theory. The results are in table I for the case of ensembles of three networks, in table II for nine, in table III for twenty and in table IV for forty networks in the ensemble.

By comparing the results of table 1, and 2 with the results of a single network we can see that the improvement by the use of the ensemble methods depends clearly on the problem. For example, in databases Aritm (except for the case of CVC version 2), Flare, Pima and Wdbc there is not a clear improvement. In the rest of databases there is an improvement; perhaps the most important one is in database Vowel.

This result (the improvement depends strongly of the database) was already known in the bibliography.

There is, however, one exception in the performance of the method Evol. This method did not work well in our experiments. In the original reference the method was tested in the database Heart. The result for a single network was 60%, for a simple ensemble 61.42% and for Evol 67.14%. We have performed experiments with database Heart from the UCI repository and our result for a simple network is $82.0 \pm 0.9$, clearly different.

Now, we can compare the results of tables I, II, III and IV for ensembles of different number of networks. We can see that the results are in general similar and the improvement of training an increasing number of networks, for example 20 and 40, is in general low. Taking into account the computational cost, we can say that the best alternative for an application is an ensemble of three or nine networks.

Table I. Results for the ensemble of three networks.

| | ARITM | DERMA | ECOLI | FLARE | IMAGEN |
|---|---|---|---|---|---|
| **Single Net.** | 75.6 ± 0.7 | 96.7 ± 0.4 | 84.4 ± 0.7 | 82.1 ± 0.3 | 96.3 ± 0.2 |
| **Adaboost** | 71.8 ± 1.8 | 98.0 ± 0.5 | 85.9 ± 1.2 | 81.7 ± 0.6 | 96.8 ± 0.2 |
| **Bagging** | 74.7 ± 1.6 | 97.5 ± 0.6 | 86.3 ± 1.1 | 81.9 ± 0.6 | 96.6 ± 0.3 |
| **Bag_Noise** | 75.5 ± 1.1 | 97.6 ± 0.7 | 87.5 ± 1.0 | 82.2 ± 0.4 | 93.4 ± 0.4 |
| **Boosting** | 74.4 ± 1.2 | 97.3 ± 0.6 | 86.8 ± 0.6 | 81.7 ± 0.4 | 95.0 ± 0.4 |
| **Cels_m** | 73.4 ± 1.3 | 97.7 ± 0.6 | 86.2 ± 0.8 | 81.2 ± 0.5 | 96.82±0.15 |
| **CVC** | 74.0 ± 1.0 | 97.3 ± 0.7 | 86.8 ± 0.8 | 82.7 ± 0.5 | 96.4 ± 0.2 |
| **Decorrelated** | 74.9 ± 1.3 | 97.2 ± 0.7 | 86.6 ± 0.6 | 81.7 ± 0.4 | 96.7 ± 0.3 |
| **Decorrelated2** | 73.9 ± 1.0 | 97.6 ± 0.7 | 87.2 ± 0.9 | 81.6 ± 0.4 | 96.7 ± 0.3 |
| **Evol** | 65.4 ± 1.4 | 57 ± 5 | 57 ± 5 | 80.7 ± 0.7 | 77 ± 5 |
| **Ola** | 74.7 ± 1.4 | 91.4 ± 1.5 | 82.4 ± 1.4 | 81.1 ± 0.4 | 95.6 ± 0.3 |
| **CVC version 2** | 76.1 ± 1.6 | 98.0 ± 0.3 | 86.8 ± 0.9 | 82.5 ± 0.6 | 96.9 ± 0.3 |
| **AveBoost** | 73.4 ± 1.3 | 97.6 ± 0.7 | 85.3 ± 1.0 | 81.8 ± 0.8 | 96.8± 0.2 |
| **TCA** | 70.7 ± 1.9 | 96.1 ± 0.6 | 85.4 ± 1.3 | 81.9 ± 0.7 | 94.8 ± 0.5 |
| **ArcX4** | 75.4 ± 0.8 | 97.8 ± 0.5 | 85.3 ± 1.1 | 78.3± 0.9 | 96.6 ± 0.2 |
| **ArcX4 Voting** | 73.0 ± 0.8 | 97.0 ± 0.5 | 85.7 ± 1.1 | 80.6 ± 0.9 | 96.5 ± 0.2 |
| **Aggressive B** | 72.3 ± 1.9 | 97.0 ± 0.5 | 85.7 ± 1.4 | 81.9 ± 0.9 | 96.6 ± 0.3 |
| **Conservative B** | 74.8 ± 1.3 | 96.9 ± 0.8 | 85.4 ± 1.3 | 82.1 ± 1.0 | 96.5 ± 0.3 |
| **EENCL UG** | 71 ± 2 | 96.8 ± 0.9 | 86.6 ± 1.2 | 81.4 ± 0.8 | 96.3 ± 0.2 |
| **EENCL MG** | 74.5 ± 1.3 | 97.2 ± 0.8 | 86.6 ± 1.2 | 81.9 ± 0.5 | 96.0 ± 0.2 |
| **Simple Ens.** | 73.4 ± 1.0 | 97.2 ± 0.7 | 86.6 ± 0.8 | 81.8 ± 0.5 | 96.5 ± 0.2 |

Table I (continuation). Results for the ensemble of three networks.

| | IONOS | PIMA | SURVI | VOWEL | WDBC |
|---|---|---|---|---|---|
| **Single Net.** | 87.9 ± 0.7 | 76.7 ± 0.6 | 74.2 ± 0.8 | 83.4 ± 0.6 | 97.4 ± 0.3 |
| **Adaboost** | 88.3 ± 1.3 | 75.7 ± 1.0 | 75.4 ± 1.6 | 88.43 ± 0.9 | 95.7 ± 0.6 |
| **Bagging** | 90.7 ± 0.9 | 76.9 ± 0.8 | 74.2 ± 1.1 | 87.4 ± 0.7 | 96.9 ± 0.4 |
| **Bag_Noise** | 92.4 ± 0.9 | 76.2 ± 1.0 | 74.6 ± 0.7 | 84.4 ± 1.0 | 96.3 ± 0.6 |
| **Boosting** | 88.9 ± 1.4 | 75.7 ± 0.7 | 74.1 ± 1.0 | 85.7 ± 0.7 | 97.0 ± 0.4 |
| **Cels_m** | 91.9 ± 1.0 | 76.0 ± 1.4 | 73.4 ± 1.3 | 91.1 ± 0.7 | 97.0 ±0.4 |
| **CVC** | 87.7 ± 1.3 | 76.0 ± 1.1 | 74.1 ± 1.4 | 89.0 ± 1.0 | 97.4 ± 0.3 |
| **Decorrelated** | 90.9 ± 0.9 | 76.4 ± 1.2 | 74.6 ± 1.5 | 91.5 ± 0.6 | 97.0 ± 0.5 |
| **Decorrelated2** | 90.6 ± 1.0 | 75.7 ± 1.1 | 74.3 ± 1.4 | 90.3 ± 0.4 | 97.0 ± 0.5 |
| **Evol** | 83.4 ± 1.9 | 66.3 ± 1.2 | 74.3 ± 0.6 | 77.5 ± 1.7 | 94.4 ± 0.9 |
| **Ola** | 90.7 ± 1.4 | 69.2 ± 1.6 | 75.2 ± 0.9 | 83.2 ± 1.1 | 94.2 ± 0.7 |
| **CVC version 2** | 89.7 ± 1.4 | 76.8 ± 1.0 | 74.1 ± 1.2 | 89.8 ± 0.9 | 96.7 ± 0.3 |
| **AveBoost** | 89.4 ± 1.3 | 76.5 ± 1.1 | 75.1 ± 1.2 | 88.1 ± 1.0 | 95.6 ± 0.5 |
| **TCA** | 87.9 ± 1.2 | 75.4 ± 0.8 | 73.0 ± 1.5 | 87.5 ± 1.1 | 91 ± 4 |
| **ArcX4** | 89.4 ± 1.0 | 76.0 ± 0.8 | 68 ± 2 | 90.8 ± 0.9 | 96.3 ± 0.6 |
| **ArcX4 Voting** | 89.0 ± 1.0 | 76.3 ± 0.8 | 74 ± 2 | 86.2 ± 0.9 | 96.1 ± 0.6 |
| **Aggressive B** | 90.3 ± 0.9 | 74.3 ± 1.5 | 73.8 ± 1.5 | 86.9 ± 1.2 | 96.6 ± 0.6 |
| **Conservative B** | 89.4 ± 1.0 | 75.6 ± 1.2 | 75.6 ± 1.1 | 88.8 ± 1.1 | 97.0 ± 0.6 |
| **EENCL UG** | 93.0 ± 1.0 | 74.7 ± 1.0 | 73.9 ± 1.2 | 87.2 ± 0.8 | 96.2 ± 0.4 |
| **EENCL MG** | 93.7 ± 0.9 | 75.3 ± 1.0 | 73.9 ± 0.8 | 87.4 ± 0.7 | 96.4 ± 0.5 |
| **Simple Ens.** | 91.1 ± 1.1 | 75.9 ± 1.2 | 74.3 ± 1.3 | 88.0 ± 0.9 | 96.9 ± 0.5 |

Table II. Results for the Ensemble of nine networks.

| | ARITM | DERMA | ECOLI | FLARE | IMAGEN |
|---|---|---|---|---|---|
| **Adaboost** | 73.2 ± 1.6 | 97.3 ± 0.5 | 84.7 ± 1.4 | 81.1 ± 0.7 | 97.3 ± 0.3 |
| **Bagging** | 75.9 ± 1.7 | 97.7 ± 0.6 | 87.2 ± 1.0 | 82.4 ± 0.6 | 96.7 ± 0.3 |
| **Bag_Noise** | 75.4 ± 1.2 | 97.0 ± 0.7 | 87.2 ± 0.8 | 82.4 ± 0.5 | 93.4 ± 0.3 |
| **Cels_m** | 74.8 ± 1.3 | 97.3 ± 0.6 | 86.2 ± 0.8 | 81.7 ± 0.4 | 96.6 ± 0.2 |
| **CVC** | 74.8 ± 1.3 | 97.6 ± 0.6 | 87.1 ± 1.0 | 81.9 ± 0.6 | 96.6 ± 0.2 |
| **Decorrelated** | 76.1 ± 1.0 | 97.6 ± 0.7 | 87.2 ± 0.7 | 81.6 ± 0.6 | 96.9 ± 0.2 |
| **Decorrelated2** | 73.9 ± 1.1 | 97.6 ± 0.7 | 87.8 ± 0.7 | 81.7 ± 0.4 | 96.84 ± 0.18 |
| **Evol** | 65.9 ± 1.9 | 54 ± 6 | 57 ± 5 | 80.6 ± 0.8 | 67 ± 4 |
| **Ola** | 72.5 ± 1.0 | 86.7 ±1.7 | 83.5 ± 1.3 | 80.8 ± 0.4 | 96.1 ± 0.2 |
| **CVC version 2** | 76.1 ± 1.6 | 98.0 ± 0.3 | 86.8 ± 0.9 | 82.5 ± 0.6 | 96.9 ± 0.3 |
| **AveBoost** | 73.4 ± 1.3 | 97.6 ± 0.7 | 85.3 ± 1.0 | 81.8 ± 0.8 | 96.8 ± 0.2 |
| **TCA** | 70.7 ± 1.9 | 96.1 ± 0.5 | 85.4 ± 1.3 | 81.9 ± 0.7 | 94.8 ± 0.5 |
| **ArcX4** | 75.4 ± 0.8 | 97.8 ± 0.5 | 85.3 ± 1.1 | 78.3 ± 0.9 | 96.6 ± 0.2 |
| **ArcX4 Voting** | 73.3 ± 0.8 | 97.6 ± 0.5 | 84.9 ± 1.1 | 80.1 ± 0.9 | 97.2 ± 0.2 |
| **Aggressive B** | 72.3 ± 1.9 | 97.0 ± 0.5 | 85.7 ± 1.4 | 81.9 ± 0.9 | 96.6 ± 0.3 |
| **Conservative B** | 74.8 ± 1.3 | 96.9 ± 0.8 | 85.4 ± 1.3 | 82.1 ± 1.0 | 96.5 ± 0.3 |
| **EENCL UG** | 71 ± 2 | 96.8 ± 0.9 | 86.6 ± 1.2 | 81.4 ± 0.8 | 96.3 ± 0.2 |
| **EENCL MG** | 74.5 ± 1.3 | 97.2 ± 0.8 | 86.6 ± 1.2 | 81.9 ± 0.5 | 96.0 ± 0.2 |
| **Simple Ens** | 73.8 ± 1.1 | 97.5 ± 0.7 | 86.9 ± 0.8 | 81.6 ± 0.4 | 96.7 ± 0.3 |

Table II (continuation). Results for the ensemble of nine networks.

| | IONOS | PIMA | SURVI | VOWEL | WDBC |
|---|---|---|---|---|---|
| **Adaboost** | 89.4 ± 0.8 | 75.5 ± 0.9 | 74.3 ± 1.4 | 94.8 ± 0.7 | 95.7 ± 0.7 |
| **Bagging** | 90.1 ± 1.1 | 76.6 ± 0.9 | 74.4 ± 1.5 | 90.8 ± 0.7 | 97.3 ± 0.4 |
| **Bag_Noise** | 93.3 ± 0.6 | 75.9 ± 0.9 | 74.8 ± 0.7 | 85.7 ± 0.9 | 95.9 ± 0.5 |
| **Cels_m** | 91.9 ± 1.0 | 75.9 ± 1.4 | 73.4 ± 1.2 | 92.7 ± 0.7 | 96.8 ± 0.5 |
| **CVC** | 89.6 ± 1.2 | 76.9 ± 1.1 | 75.2 ± 1.5 | 90.9 ± 0.7 | 96.5 ± 0.5 |
| **Decorrelated** | 90.7 ± 1.0 | 76.0 ± 1.1 | 73.9 ± 1.3 | 92.8 ± 0.7 | 97.0 ± 0.5 |
| **Decorrelated2** | 90.4 ± 1.0 | 76.0 ± 1.0 | 73.8 ± 1.3 | 92.6 ± 0.5 | 97.0 ± 0.5 |
| **Evol** | 77 ± 3 | 66.1 ± 0.7 | 74.8 ± 0.7 | 61 ± 4 | 87.2 ± 1.6 |
| **Ola** | 90.9 ± 1.7 | 73.8 ± 0.8 | 74.8 ± 0.8 | 88.1 ± 0.8 | 95.5 ± 0.6 |
| **CVC version 2** | 89.7 ± 1.4 | 76.8 ± 1.0 | 74.1 ± 1.2 | 89.8 ± 0.9 | 96.7 ± 0.3 |
| **AveBoost** | 89.4 ± 1.3 | 76.5 ± 1.1 | 75.1 ± 1.2 | 88.1 ± 1.0 | 95.6 ± 0.5 |
| **TCA** | 87.9 ± 1.2 | 75.4 ± 0.8 | 73.0 ± 1.5 | 87.5 ± 1.1 | 91 ± 4 |
| **ArcX4** | 89.4 ± 1.0 | 76.0 ± 0.8 | 68 ± 2 | 90.8 ± 0.9 | 96.3 ± 0.6 |
| **ArcX4 Voting** | 91.3 ± 1.0 | 76.3 ± 0.8 | 73.9 ± 1.0 | 94.6 ± 0.9 | 96.6 ± 0.6 |
| **Aggressive B** | 90.3 ± 0.9 | 74.3 ± 1.5 | 73.8 ± 1.5 | 86.9 ± 1.2 | 96.6 ± 0.6 |
| **Conservative B** | 89.4 ± 1.0 | 75.6 1.2 | 75.6 ± 1.1 | 88.8 ± 1.1 | 97.0 ± 0.6 |
| **EENCL UG** | 93.0 ± 1.0 | 74.7 ± 1.0 | 73.9 ± 1.2 | 87.2 ± 0.8 | 96.2 ± 0.4 |
| **EENCL MG** | 93.7 ± 0.9 | 75.3 ± 1.0 | 73.9 ± 0.8 | 87.4 ± 0.7 | 96.4 ± 0.5 |
| **Simple Ens** | 90.3 ± 1.1 | 75.9 ± 1.2 | 74.2 ± 1.3 | 91.0 ± 0.5 | 96.9 ± 0.5 |

Table III. Results for the ensemble of twenty networks.

| | ARITM | DERMA | ECOLI | FLARE | IMAGEN |
|---|---|---|---|---|---|
| **Adaboost** | 71.4 ± 1.5 | 97.5 ± 0.6 | 86.0 ± 1.3 | 81.1 ± 0.8 | 97.3 ± 0.2 |
| **Bagging** | 75.9 ± 1.7 | 97.6 ± 0.6 | 87.1 ± 1.0 | 82.2 ± 0.5 | 97.0 ± 0.3 |
| **Bag_Noise** | 76.0 ± 1.1 | 97.3 ± 0.6 | 87.4 ± 0.8 | 82.1 ± 0.5 | 93.3 ± 0.3 |
| **Cels_m** | 75.4 ± 1.2 | 93.9 ± 1.4 | 86.3 ± 1.3 | 81.5 ± 0.4 | 95.7 ± 0.2 |
| **CVC** | 74.8 ± 1.3 | 97.3 ± 0.6 | 86.5 ± 1.0 | 81.7 ± 0.7 | 96.8 ± 0.2 |
| **Decorrelated** | 76.1 ± 1.1 | 97.6 ± 0.7 | 87.1 ± 0.7 | 81.3 ± 0.5 | 96.9 ± 0.2 |
| **Decorrelated2** | 73.9 ± 1.1 | 97.6 ± 0.7 | 88.1 ± 0.7 | 81.6 ± 0.5 | 96.8 ± 0.2 |
| **Evol** | 65.9 ± 1.9 | 47 ± 5 | 55 ± 4 | 81.2 ± 0.5 | 63 ± 5 |
| **Ola** | 72.5 ± 1.1 | 87.0 ± 1.4 | 84.3 ± 1.2 | 80.7 ± 0.4 | 96.4 ± 0.2 |
| **CVC version 2** | 74.3 ± 1.2 | 97.5 ± 0.6 | 86.6 ± 1.1 | 81.8 ± 0.4 | 97.0 ± 0.2 |
| **AveBoost** | 75.5 ± 1.1 | 97.9 ± 0.5 | 86.2 ± 1.2 | 82.4 ± 0.7 | 97.3 ± 0.3 |
| **TCA** | 71.6 ± 1.8 | 92 ± 2 | 85.4 ± 1.5 | 79.7 ± 0.9 | 95.7 ± 0.3 |
| **ArcX4** | 74.4 ± 1.4 | 97.8 ± 0.6 | 85.6 ± 0.8 | 78.4 ± 1.4 | 97.4 ± 0.2 |
| **ArcX4 Voting** | 75.1 ± 1.2 | 97.3 ± 0.7 | 86.0 ± 1.2 | 78.6 ± 1.0 | 97.3 ± 0.2 |
| **Aggressive B** | 74.8 ± 1.5 | 97.0 ± 0.6 | 87.1 ± 1.1 | 82.0 ± 0.5 | 97.2 ± 0.3 |
| **Conservative B** | 74.7 ± 0.9 | 97.9 ± 0.6 | 86.9 ± 1.2 | 82.8 ± 0.6 | 97.2 ± 0.3 |
| **EENCL UG** | 72.9 ± 0.9 | 95.1 ± 1.1 | 87.2 ± 0.7 | 82.0 ± 0.8 | 96.9 ± 0.3 |
| **EENCL MG** | 73.5 ± 1.6 | 96.2 ± 0.9 | 87.7 ± 1.0 | 81.4 ± 0.6 | 96.6 ± 0.3 |
| **Simple Ens** | 73.8 ± 1.1 | 97.3 ± 0.7 | 86.9 ± 0.8 | 81.5 ± 0.5 | 96.7 ± 0.2 |

Table III (continuation). Results for the ensemble of twenty networks.

| | IONOS | PIMA | SURVI | VOWEL | WDBC |
|---|---|---|---|---|---|
| **Adaboost** | 91.4 ± 0.8 | 74.8 ± 1.0 | 74.3 ± 1.5 | 96.1 ± 0.7 | 96.3 ± 0.5 |
| **Bagging** | 89.6 ± 1.1 | 77.0 ± 1.0 | 74.6 ± 1.7 | 91.3 ± 0.6 | 97.5 ± 0.4 |
| **Bag_Noise** | 92.7 ± 0.6 | 76.3 ± 0.8 | 74.6 ± 0.7 | 86.7 ± 0.7 | 96.1 ± 0.5 |
| **Cels_m** | 93.3 ± 0.7 | 75.4 ± 1.0 | 64 ± 3 | 87.5 ± 0.8 | 96.5 ± 0.5 |
| **CVC** | 89.6 ± 1.3 | 76.2 ± 1.3 | 73.8 ± 0.9 | 91.9 ± 0.5 | 97.4 ± 0.4 |
| **Decorrelated** | 91.1 ± 0.9 | 76.1 ± 1.0 | 74.1 ± 1.4 | 93.3 ± 0.6 | 97.0 ± 0.5 |
| **Decorrelated2** | 90.9 ± 0.9 | 76.1 ± 1.0 | 74.3 ± 1.3 | 93.3 ± 0.5 | 97.0 ± 0.5 |
| **Evol** | 66.1 ± 1.2 | 65.2 ± 0.9 | 74.8 ± 0.7 | 60 ± 3 | 78 ± 3 |
| **Ola** | 69.4 ± 1.2 | 74.2 ± 1.1 | 74.1 ± 0.7 | 88.7 ± 0.8 | 95.3 ± 0.6 |
| **CVC version 2** | 91.0 ± 0.9 | 76.7 ± 0.8 | 73.6 ± 1.0 | 93.3 ± 0.6 | 95.9 ± 0.6 |
| **AveBoost** | 91.4 ± 1.0 | 76.0 ± 1.1 | 74.8 ± 1.2 | 95.8 ± 0.6 | 95.8 ± 0.6 |
| **TCA** | 86.1 ± 1.0 | 73.5 ± 0.9 | 71.3 ± 1.8 | 84 ± 3 | 94.4 ± 0.7 |
| **ArcX4** | 92.0 ± 0.9 | 72.7 ± 1.1 | 69 ± 2 | 96.6 ± 0.5 | 96.4 ± 0.6 |
| **ArcX4 Voting** | 92.6 ± 0.9 | 75.0 ± 0.9 | 73.8 ± 1.5 | 96.1 ± 0.7 | 96.6 ± 0.6 |
| **Aggressive B** | 91.6 ± 0.9 | 75.5 ± 1.3 | 73.9 ± 1.7 | 96.9 ± 0.6 | 96.8 ± 0.6 |
| **Conservative B** | 92.4 ± 1.0 | 76.7 ± 1.2 | 72.8 ± 1.3 | 96.6 ± 0.6 | 96.4 ± 0.6 |
| **EENCL UG** | 92.3 ± 1.1 | 75.2 ± 0.8 | 72.5 ± 1.5 | 88.2 ± 0.9 | 95.8 ± 0.4 |
| **EENCL MG** | 92.3 ± 1.0 | 76.2 ± 1.3 | 74.1 ± 1.0 | 88.3 ± 0.9 | 96.5 ± 0.4 |
| **Simple Ens** | 90.4 ± 1.0 | 75.9 ± 1.2 | 74.3 ± 1.3 | 91.4 ± 0.8 | 96.9 ± 0.5 |

Table IV Results for the ensemble of forty networks.

| | ARITM | DERMA | ECOLI | FLARE | IMAGEN |
|---|---|---|---|---|---|
| **Adaboost** | 73.8 ± 1.1 | 97.8 ± 0.5 | 85.7 ± 1.4 | 81.1 ± 0.7 | 97.3 ± 0.2 |
| **Bagging** | 74.7 ± 1.5 | 97.6 ± 0.6 | 86.9 ± 1.1 | 82.0 ± 0.6 | 97.1 ± 0.3 |
| **Bag_Noise** | 75.7 ± 1.3 | 97.5 ± 0.6 | 87.5 ± 0.8 | 82.1 ± 0.6 | 93.4 ± 0.3 |
| **Cels_m** | 74.5 ± 1.7 | 95.3 ± 1.2 | 81.9 ± 1.8 | 81.5 ± 0.4 | 95.7 ± 0.2 |
| **CVC** | 73.4 ± 1.9 | 97.3 ± 0.6 | 86.8 ± 0.9 | 81.7 ± 0.7 | 96.6 ± 0.2 |
| **Decorrelated** | 75.6 ± 1.3 | 97.6 ± 0.7 | 87.5 ± 0.7 | 81.4 ± 0.5 | 96.9 ± 0.2 |
| **Decorrelated2** | 74.4 ± 1.2 | 97.6 ± 0.7 | 88.2 ± 0.7 | 81.7 ± 0.4 | 96.8 ± 0.2 |
| **Evol** | 59 ± 2 | 41 ± 7 | 52 ± 6 | 81.2 ± 0.5 | 63 ± 4 |
| **Ola** | 75.1 ± 1.1 | 87.7 ± 1.6 | 84.9 ± 1.3 | 80.7 ± 0.4 | 96.3 ± 0.2 |
| **CVC version 2** | 77.0 ± 0.8 | 97.2 ± 0.6 | 86.3 ± 0.9 | 82.2 ± 0.5 | 96.7 ± 0.3 |
| **AveBoost** | 76.3 ± 1.0 | 97.2 ± 0.7 | 86.0 ± 1.1 | 80.7 ± 1.1 | 97.5 ± 0.2 |
| **TCA** | 70.6 ± 1.7 | 82 ± 5 | 84.0 ± 1.4 | 80.1 ± 1.0 | 95.8 ± 0.3 |
| **ArcX4** | 74.0 ± 1.4 | 97.5 ± 0.6 | 86.2 ± 1.0 | 80.0 ± 1.1 | 97.4 ± 0.2 |
| **ArcX4 Voting** | 74.6 ± 1.0 | 97.2 0.7 | 85.9 ± 1.2 | 80.1 ± 0.8 | 97.4 ± 0.2 |
| **Aggressive B** | 75.5 ± 1.2 | 96.6 ± 0.5 | 87.8 ± 0.9 | 82.0 ± 0.6 | 97.3 ± 0.3 |
| **Conservative B** | 75.1 ± 1.0 | 97.6 ± 0.7 | 87.8 ± 1.1 | 82.7 ± 0.6 | 97.2 ± 0.2 |
| **EENCL UG** | 70.7 ± 1.9 | 95.8 ± 1.1 | 86.5 ± 0.8 | 82.1 ± 0.7 | 97.0 ± 0.2 |
| **EENCL MG** | 74.1 ± 1.2 | 96.1 ± 1.1 | 88.1 ± 0.7 | 82.1 ± 0.5 | 96.8 ± 0.3 |
| **Simple Ens** | 73.8 ± 1.1 | 97.6 ± 0.7 | 86.9 ± 0.7 | 81.6 ± 0.5 | 96.8 ± 0.2 |

Table IV (continuation). Results for the ensemble of forty networks.

| | IONOS | PIMA | SURVI | VOWEL | WDBC |
|---|---|---|---|---|---|
| **Adaboost** | 91.6 ± 0.7 | 73.3 ± 1.0 | 73 ± 2 | 97.0 ± 0.6 | 96.7 ± 0.9 |
| **Bagging** | 90.0 ± 1.1 | 77.0 ± 1.1 | 74.2 ± 1.3 | 91.2 ± 0.8 | 97.4 ± 0.3 |
| **Bag_Noise** | 93.0 ± 0.6 | 76.4 ± 0.9 | 74.6 ± 0.7 | 86.5 ± 0.8 | 95.9 ± 0.5 |
| **Cels_m** | 92.9 ± 0.9 | 75.7 ± 0.7 | 71.3 ± 1.9 | 79.1 ± 1.3 | 96.3 ± 0.6 |
| **CVC** | 88.3 ± 1.0 | 76.6 ± 1.0 | 74.6 ± 1.0 | 92.2 ± 0.8 | 96.8 ± 0.5 |
| **Decorrelated** | 91.0 ± 1.0 | 75.9 ± 1.0 | 73.9 ± 1.4 | 93.1 ± 0. | 97.0 ± 0.5 |
| **Decorrelated2** | 90.7 ± 0.9 | 76.2 ± 1.0 | 74.3 ± 1.3 | 93.5 ± 0.6 | 96.9 ± 0.4 |
| **Evol** | 64.1 ± 1.3 | 65.9 ± 1.0 | 74.8 ± 0.7 | 54 ± 3 | 77.6 ± 1.9 |
| **Ola** | 69.4 ± 1.4 | 74.4 ± 0.7 | 74.8 ± 0.8 | 88.6 ± 1.0 | 95.7 ± 0.6 |
| **CVC version 2** | 92.0 ± 1.0 | 76.1 ± 0.9 | 73.4 ± 1.2 | 92.9 ± 0.7 | 96.0 ± 0.5 |
| **AveBoost** | 91.6 ± 0.9 | 76.6 ± 1.0 | 74.6 ± 1.1 | 96.4 ± 0.6 | 96.0 ± 0.5 |
| **TCA** | 79 ± 5 | 73.7 ± 0.8 | 70.2 ± 1.7 | 79 ± 4 | 92.8 ± 1.7 |
| **ArcX4** | 91.6 ± 1.0 | 73.4 ± 0.6 | 72.3 ± 1.4 | 96.9 ± 0.5 | 96.5 ± 0.6 |
| **ArcX4 Voting** | 91.7 ± 1.0 | 74.5 ± 0.9 | 73.0 ± 1.4 | 97.0 ± 0.5 | 96.7 ± 0.6 |
| **Aggressive B** | 92.3 ± 0.9 | 75.7 ± 1.2 | 73.9 ± 1.4 | 97.5 ± 0.5 | 96.7 ± 0.6 |
| **Conservative B** | 91.9 ± 0.8 | 76.2 ± 1.2 | 73.3 ± 1.5 | 97.3 ± 0.6 | 96.3 ± 0.5 |
| **EENCL UG** | 92.9 ± 0.8 | 74.3 ± 0.8 | 72.3 ± 1.2 | 90.0 ± 1.0 | 96.5 ± 0.8 |
| **EENCL MG** | 93.7 ± 0.7 | 76.5 ± 1.2 | 75.3 ± 1.1 | 89.6 ± 0.8 | 96.8 ± 0.6 |
| **Simple Ens** | 90.3 ± 1.0 | 75.9 ± 1.2 | 74.3 ± 1.3 | 92.2 ± 0.7 | 96.9 ± 0.5 |

We have also calculated the percentage of error reduction of the ensemble with respect to a single network. We have used equation I for this calculation.

$$\text{PerError}_{\text{reduction}} = 100 \frac{\text{PerError}_{\text{single network}} - \text{PerError}_{\text{ensemble}}}{\text{PerError}_{\text{single network}}} \qquad \text{(I)}$$

The value of the percentage of error reduction ranges from 0%, where there is no improvement by the use of a particular ensemble method to 100%. There can also be negative values when the performance of the ensemble is worse than the single network.
This new measurement is relative and can be used to compare more clearly the different methods. Furthermore we can calculate the mean performance of error reduction across all databases this value is in table V for ensembles of 3, 9, 20 and 40 nets.

Table V. Mean percentage of error reduction for the different ensembles.

| | **Ensemble 3 Nets** | **Ensemble 9 Nets** | **Ensemble 20 Nets** | **Ensemble 40 Nets** |
|---|---|---|---|---|
| **Adaboost** | 1.33 | 4.26 | 9.38 | 12.21 |
| **Bagging** | 6.86 | 12.12 | 13.36 | 12.63 |
| **Bag_Noise** | -3.08 | -5.08 | -3.26 | -3.05 |
| **Boosting** | -0.67 | --- | --- | --- |
| **Cels_m** | 9.98 | 9.18 | 10.86 | 14.43 |
| **CVC** | 6.18 | 7.76 | 10.12 | 6.48 |
| **Decorrelated** | 9.34 | 12.09 | 12.61 | 12.35 |
| **Decorrelated2** | 9.09 | 11.06 | 12.16 | 12.10 |
| **Evol** | -218.23 | -297.01 | -375.36 | -404.81 |
| **Ola** | -33.11 | -36.43 | -52.53 | -47.39 |
| **CVC version 2** | 10.25 | 10.02 | 7.57 | 7.49 |
| **AveBoost** | 1.13 | 10.46 | 9.38 | 10.79 |
| **TCA** | -9.71 | -25.22 | -43.98 | -53.65 |
| **ArcX4** | 1.21 | 2.85 | 7.85 | 10.05 |
| **ArcX4 Voting** | -2.08 | 9.73 | 10.76 | 11.14 |
| **Aggressive B** | 1.22 | 7.34 | 13.03 | 13.54 |
| **Conservative B** | 4.45 | 13.07 | 14.8 | 14.11 |
| **EENCL UG** | 0.21 | -3.23 | -3.59 | 1.10 |
| **EENCL MG** | 3.96 | 1.52 | 2.84 | 7.89 |
| **Simple Ens** | 5.89 | 8.39 | 8.09 | 9.72 |

According to this global measurement *Ola*, *Evol* and *BagNoise* performs worse than the Simple Network, i.e., it is not worthy to use an ensembles with these three methods. The best methods are *Bagging*, *Cels*, *Decorrelated*, *Decorrelated2* and *Conservative Boosting*.
The best method for 3 nets in the ensemble is *CVC version 2*, the best method for the case of 9 and 20 nets is *Conservative Boosting* and the best method for the case of 40 networks is *Cels* but the performance of *Conservative Boosting* is also good.
So, we can conclude that if the number of networks is low it seems that the best method is *CVC version 2* and if the number of network is high the best method is in general *Conservative Boosting*.
Also in table V, we can see the effect of increasing the number of networks in the ensemble. There are several methods (*Adaboost, Cels, ArcX4, ArcX4 Voting, Aggressive Boosting and Conservative Boosting*) where the performance seems to increase slightly with the number of networks in the ensemble. But other methods like *Bagging, CVC, Decorrelated, Decorrelated2* and *Simple Ensemble* does not increase the performance beyond 9 or 20 networks in the ensemble. The reason can be that the new networks are correlated to the first ones or that the combination method (the average) does not exploit well the increase in the number of networks.

## IV. Conclusion

In this paper we have presented experimental results of twenty different methods to construct an ensemble of networks, using ten different databases. We trained ensembles of 3, 9, 20 and 40 networks in the ensemble, so we have cover a wide spectrum in the number of networks in the ensemble. The results showed that in general the improvement by the use of the ensemble methods depends clearly on the database as it was already known in the bibliography. Also the improvement in performance from three or nine networks in the ensemble to a higher number of networks depends on the method, but it is usually low. Taking into account the computational cost, an ensemble of nine networks may be the best alternative for most of the methods. Finally, we have obtained the mean percentage of error reduction over all databases. It is a relative measurement that it is not statistically significant but we think it can be useful to have an insight in the performance. According to the results of this measurement the best methods are *Bagging*, *Cels*, *Decorrelated*, *Decorrelated2* and *Conservative Boosting*. The best method for 3 networks in the ensemble is CVC version 2, the best method for the case of 9 and 20 nets is Conservative Boosting and the best method for 40 is *Cels* but the performance of *Conservative Boosting* is also good. So we can conclude that if the number of networks is low it seems that the best method is *CVC Version 2* and if the number of network is high the best method is in general *Conservative Boosting*.

## IV. Acknowledgments

This research was supported by the project MAPACI TIC2002-02273 of CICYT in Spain.

## References

[1] Tumer, K., Ghosh, J., “Error correlation and error reduction in ensemble classifiers”, Connection Science, vol. 8, nos. 3 & 4, pp. 385-404, 1996.

[2] Raviv, Y., Intrator, N., “Bootstrapping with Noise: An Effective Regularization Technique”, Connection Science, vol. 8, no. 3 & 4, pp. 355-372, 1996.

[3] Drucker, H., Cortes, C., Jackel, D., et alt., “Boosting and Other Ensemble Methods”, Neural Computation, vol. 6, pp. 1289-1301, 1994.

[4] Fernandez-Redondo, Mercedes, Hernández-Espinosa, Carlos, Torres-Sospedra, Joaquín, “Classification by Multilayer Feedforward ensambles”, Internacional Symposium on Neural Networks, Lecture Notes in Computer Science, Vol. 3173, pp. 852-857, 2004.

[5] Verikas, A., Lipnickas, A., Malmqvist, K., Bacauskiene, M., Gelzinis, A., “Soft Combination of neural classifiers: A comparative study“, Pattern Recognition Letters, Vol. 20, pp 429-444, 1999.

[6] Oza, N.C., “Boosting with Averaged Weight Vectors”, Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 2709, pp. 15-24, 2003.

[7] Kuncheva, L.I., “Error Bounds for Aggressive and Conservative Adaboost”, Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 2709, pp. 25-34, 2003.

[8] Breiman, L., “Arcing Classifiers”, Annals of Statistic, vol. 26, no. 3, pp. 801-849, 1998.

[9] Liu, Y., Yao, X., Higuchi, T., “Evolutionary Ensembles with Negative Correlation Learning”, IEEE Trans. On Evolutionary Computation, vol. 4, no. 4, pp. 380-387, 2000.

**Carlos Hernández-Espinosa** was born in Murcia, Spain. He received the Physics degree from *Universidad de Valencia* in 1989 and the Ph.D degree in 1994. Nowadays, he is a lecturer and the leader of the Neural Networks and Soft Computing research group at *Universitat Jaume I.*

**Joaquín Torres-Sospedra** was born in Castellón, Spain. He received the Computer Science degree from *Universitat Jaume I* in 2003. Nowadays, he is working for *Universidad Politécnica de Madrid* at MAPACI project and he is a PhD Student at *Universitat Jaume I*.

**Mercedes Fernandez-Redondo** was born in Murcia, Spain. She received the Physics degree from *Universidad de Valencia* and the Ph.D degree from *Universitat Jaume I* in 2001. Nowadays, she is a lecturer at *Universitat Jaume I*, Castellón, Spain.