Automatic diagnosis of defects of rolling element bearings based on computational intelligence techniques

Marco Cococcioni, Beatrice Lazzerini, Sara Lioba Volpi

Department of Information Engineering, University of Pisa, Via Diotisalvi, 2, Pisa, Italy

{m.cococcioni, b.lazzerini, sara.volpi}@iet.unipi.it

Abstract—This paper presents a method, based on classification techniques, for automatic detection and diagnosis of defects of rolling element bearings. We used vibration signals recorded by four accelerometers on a mechanical device including rolling element bearings: the signals were collected both with all faultless bearings and after substituting one faultless bearing with an artificially damaged one. We considered four defects and, for one of them, three severity levels. In all the experiments performed on the vibration signals represented in the frequency domain we achieved a classification accuracy higher than 99%, thus proving the high sensitivity of our method to different types of defects and to different degrees of fault severity. We also assessed the degree of robustness of our method to noise by analyzing how the classification performance varies on variation of the signal-to-noise ratio and using statistical classifiers and neural networks. We achieved very good levels of robustness.

Keywords-automatic fault diagnosis; robust classification; statistical classifiers; neural networks; classifier fusion

I. INTRODUCTION

Machine condition monitoring and fault diagnostics are key factors to guarantee a continuous and reliable production process in all manufacturing and production industries [1]-[4]. In particular, condition-based maintenance, which consists of real-time monitoring of the state of rotating machines and performing appropriate maintenance actions when necessary, is the most affective type of maintenance.

Rotating machines usually operate by means of bearings which may be affected by several types of faults, e.g., indentation on the roll, indentation on the inner raceway, unbalanced cage. These faults may be responsible for machine breakdown and performance level reduction.

Different methods for detection and diagnosis of faults in bearings have been developed. Traditional techniques for bearing performance analysis include time-domain analysis [5,6] and frequency-domain analysis [4] used separately or together [7,8]. Time-domain analysis is usually based on performance indexes such as RMS (Root Mean Square), Crest Factor and Kurtosis, while frequency-domain analysis is mainly based on the Fourier Transform technique. Most methods found in the literature, however, consider the bearing fault diagnosis as a two-class problem because they just distinguish between integral and damaged bearings independently of the type and/or the severity of the defect.

This paper aims to achieve the following objectives: given a mechanical object containing rolling bearings, i) to detect the presence of a defect, ii) to recognize the specific kind of defect, iii) to recognize the severity of the defect. To this aim, we have dealt with the problem as a classification problem, adopting two statistical classifiers, namely the Linear Discriminant Classifier (LDC) [9] and the Quadratic Discriminant Classifier (QDC) [9], Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) neural networks [10]. Actually, we use LDC and QDC to perform both feature selection and signal classification, whereas MLPs and RBFs only perform classification of signals represented by means of the features selected by LDC or QDC.

Finally, in order to improve classification performance, we have resorted to the concept of classifier ensemble [11,12]. Four different approaches exist to building a classifier ensemble [11,12], depending on the level at which the fusion occurs, namely: data level (different datasets are used), feature level (different feature sets are used), classifier level (different base classifiers are used), combinational level (different combiners are used). In this paper we have adopted the first three levels.

II. EXPERIMENTAL DATASET

In the experiments we used vibration signals coming from a mechanical device including more than ten rolling element bearings monitored, in the time domain, by means of four accelerometers: the signals were collected both with all faultless bearings and after substituting one faultless bearing with a damaged one.

The bearings were artificially damaged and experimental data were collected before and after each damage. Four types of damages were considered so that the data can be classified into five classes:

- C1: faultless bearing,
- C2: indentation on the inner raceway,
- C3: indentation on the roll,
- C4: sandblasting of the inner raceway,
- C5: unbalanced cage.

The fault of class C2 consists of a 450 μm indentation on the inner raceway, whereas bearings of class C3 can be divided into three subclasses depending on the severity of the damage, namely:

- C3.1: 450 μm indentation on the roll (light),
- C3.2: 1.1 mm indentation on the roll (medium),
- C3.3: 1.29 mm indentation on the roll (high).

We used the PRTools software in the Matlab environment[13].

IEEE computer society

The data were recorded by the four accelerometers for time intervals of ten minutes. We considered a data set consisting of one-second signals distributed as shown in Table I.

TABLE I. Signals per class

Class	C1	C2	C3			C4	C5
			C3.1	C3.2	C3.3		
Number of signals (sec)	2890	1770	1770	1250	1770	1520	1770

We worked in the frequency domain by transforming the signals by the Fast Fourier Transform (FFT). Unlike the classical approach, which identifies specific characteristic frequencies associated with given defects, we tried to automatically find out the frequencies able to discriminate among the different defects taken into consideration.

Based on heuristic considerations, for each accelerometer, we considered the frequency interval [1-300] Hz (the interval does not contain the continuous component), sampled every 1 Hz. Within this interval, we took into account six frequency ranges: [1-50] Hz, [51-100] Hz, [101-150] Hz, [151-200] Hz, [201-250] Hz, [251-300] Hz. As there are four accelerometers, up to $300 \times 4 = 1200$ frequency samples (obtained by concatenating the four groups of 300 frequency samples relative to the four accelerometers) could be used to represent each signal (Fig. 1). In other words, each signal can be represented in $\Re^n$ with $n \leq 1200$. The frequency samples will be referred to as features in the following.
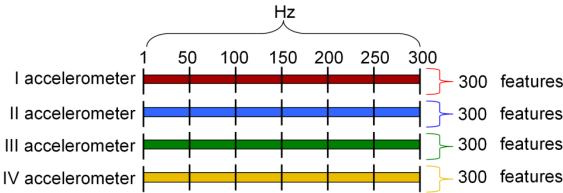

Figure 1. Organization of the features.

In the experiments the data were balanced using a random technique so that each class contains the same number of samples as the least numerous one. Then the training set was built using the hold-out method [11], i.e., by randomly choosing 70 % of the total data, while the remaining data were used as test set.

A. First experiment

The aim of this experiment is to classify the signals into five classes C1, C2, C3.1, C4, C5, i.e., to recognize the different types of faults, but considering only the lowest level of fault severity for class C3.

The choice as how to represent the signals takes into account the memory necessary for signal representation and the time needed for classifier training. Both of them should be kept reasonably low. Besides we desire to rank the accelerometers according to their contribution to classification accuracy so as to identify the most significant accelerometer(s).

We therefore consider the four accelerometers separately from each other, and, for each of them, we look for the frequencies that are able to provide the best accuracy when used to represent the signals to be classified. Indeed, features are not all equally relevant [11]. In this way, we can decrease the space dimension and the training time.

This step is performed using the forward feature selection (FFS), based on the *featself* function of PRTools. We chose to use FFS because it is a reasonable compromise between exhaustive search and random search. Further, we adopted LDC and QDC to perform both feature selection and classification of the signals represented through the selected features. This choice stems from the fact that LDC and QDC are fast trainable classifiers with only one parameter r (called regularization parameter). We fix r to 0 for both LDC and QDC classifiers in all the experiments.

We use 4 LDCs and 4 QDCs: each LDC/QDC works on the frequency range [1-300 Hz] of a particular accelerometer.

We experimentally verified that the classification accuracy of a typical classifier (see Fig. 2) increases with the number of features up to a point in which the accuracy remains almost constant and eventually decreases reaching a value that is equal to 1/n, with n being the number of classes (we recall that we work with balanced classes). Please note that in Fig. 2 the numbers on the x-axis refer to the number of features used for classification and not to the ordered features in the interval [1-300] Hz).
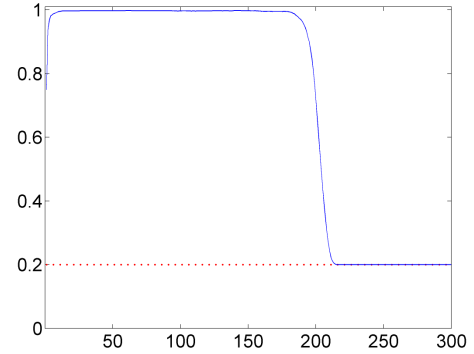

Figure 2. Typical curve of classification accuracy (y-axis) versus number of features (x-axis) for a five-class problem.

To solve this classification problem we perform the following steps.

1) First, fixed a maximum number of features equal to 10 so as to keep the computational complexity at an acceptable level, we repeat the FFS for a reasonable number t of times (t=30 in our case based on heuristic considerations) using both the LDC and QDC classifiers for each accelerometer. In each trial the H-method is applied to generate the training set.

2) We identify the stable features (SFs), i.e., the features that are the same and selected in the same order, in all the trials, by the FFS. Actually, the features selected by the FFS may vary from one trial to another. In order to guarantee a higher level of generalization, we are, therefore, interested in

identifying the features that are significant for all the training sets.

3) Once identified the SFs, we use them to compute the classification accuracy expressed in the form (mean ± standard deviation) for each accelerometer over 30 more trials. We then compare the four accelerometers (Table II) and identify the best one(s).

From Table II, considering the accuracies, we identify, as the most promising accelerometers, the second and the third accelerometers using LDC.

TABLE II. ACCURACY OF FIRST EXPERIMENT

Accelero meter	LDC		QDC	
	Number of SFs	Accuracy Mean±Std.Dev.	Number of SFs	Accuracy Mean±Std.Dev.
1	1	62.85±0.51%	1	62.26±0.49%
2	3	**98.13±0.22%**	2	96.54±0.12%
3	3	**96.88±0.22%**	2	94.70±0.61%
4	3	95.70±0.41%	3	96.55±0.16%

4) Based on design specifications, we consider good, and thus acceptable, an accuracy higher than a threshold θ (θ = 99.00 % in the experiments). If at least one accelerometer meets this requirement, we consider the best of them. Otherwise, like in the case under consideration, we try to improve the performance making use of the feature-level approach to building classifier ensembles [11,12]. To this aim, we consider the SFs for the two best accelerometers (the second and third, in our case, using the LDC classifier (Table III)).

TABLE III. SFS FOR ACCELEROMENTERS 2 AND 3

LDC, Accelerometer 2	296, 295, 277
LDC, Accelerometer 3	277, 296, 96

Then we identify the stable ranges, i.e., the frequency ranges containing the stable features above. In this case, we obtain the range [251-300] Hz for the second accelerometer, and the ranges [251-300] Hz and [51-100] Hz for the third accelerometer.

Finally, we consider the union of these three ranges and perform the FFS on them with LDC and QDC. We repeat the feature selection 30 times, and again we collect the SFs (Table IV). Please note that in Table IV the ranges [1-50] Hz, [51-100] Hz, [101-150] Hz correspond, respectively, to the range [251-300] Hz of the second accelerometer, and the ranges [51-100] Hz and [251-300] Hz of the third accelerometer. Thus the SFs 127 and 146 correspond, respectively, to the features 277 and 296 of the third accelerometer while the SFs 46 and 45 correspond, respectively, to the features 296 and 295 of the second accelerometer. Using only these SFs, over 30 more trails, we compute the classification accuracy (Table V). From Table V we can see that the accuracy (99.82 ± 0.06 %) obtained by the QDC classifier meets our design specifications. Table VI shows a typical confusion matrix using the QDC classifier and the selected SFs.

In this experiment, we managed to reduce the space dimension and thus the complexity to an acceptable level (signals are represented only in $\Re^4$), drastically decreasing the memory and time required for training and classification.

TABLE IV. LIST OF THE SFS

QDC, Accelerometers 2,3	127, 146, 46, 45

TABLE V. ACCURACY

Accelerometer Frequency range (Hz)	LDC		QDC	
	Num. of SFs	Accuracy Mean±Std.Dev.	Num. of SFs	Accuracy Mean±Std.Dev.
2 (251-300) 3 (251,300)	4	99.46±0.05%	4	**99.82±0.06%**

TABLE VI. CONFUSION MATRIX

		Estimated labels					Total
		C1	C2	C3.1	C4	C5	
True labels	C1	456	0	0	0	0	456
	C2	0	455	0	1	0	456
	C3.1	0	0	456	0	0	456
	C4	0	1	0	455	0	456
	C5	0	0	0	0	456	456
Total		456	456	456	456	456	2280

However, considering class C3, we have worked using only the lowest level of fault severity, i.e., C3.1. Of course, we are interested not only in identifying the defect as soon as possible, but also in recognizing higher levels of severity. In other words, we want to classify signals of classes C3.2 and C3.3 as belonging to the same class C3.1, as they represent the same category of defect. On the other hand the choice not to include C3.2 and C3.3 in the training test is motivated by the fact that we cannot expect to have always all the different types of severity of a particular defect at disposal to train a classifier. Thus we want to check whether the classifier, trained using only the basic defects (i.e., the defects at their lowest level of severity), can correctly recognize also the derived defects (i.e., the defects at higher levels of severity). This is of the utmost importance to make our method a practical tool.

Using the SFs previously found and the QDC classifier we tested our classifier on a test set composed not only by C1, C2, C3.1, C4 and C5 but also by C3.2 and C3.3. Over 30 trials, we obtained an accuracy of 99.80±0.02 %. A typical confusion matrix for this problem is shown in Table VII. From Table VII we can notice that all the elements of class C3 were classified correctly.

B. *Second experiment*

The second experiment aims to assess the robustness of the proposed method to noise. To this aim, we repeated the previous experiment (training on C1, C2, C3.1, C4 and C5, and test on C1, C2, C3.1, C3.2, C3.3, C4 and C5) using the optimal configuration previously found (see Table IV). We trained the QDC classifier with a training set consisting of

signals not affected by noise, then we tested it on signals affected by different levels of noise. In particular, we generated a signal of white *gaussian* noise (stochastic process with zero mean and unity variance), which was multiplied by an increasing positive coefficient, *noise level* (*NL*), to raise its power.

TABLE VII. CONFUSION MATRIX

		Estimated labels					Total
		C1	C2	C3	C4	C5	
True labels	C1	455	0	0	0	0	456
	C2	0	455	0	1	0	456
	C3	0	0	456	0	0	456
	C4	0	3	0	456	0	456
	C5	0	0	0	0	456	456
Total		455	458	456	455	456	2280

The noise was added to all the signals of the test set in the time domain; we then computed the signal-to-noise ratio, *SNR*, as the ratio between the average power of the signals of all classes and the noise power. At this point we studied the trend of the percentage of correct classification of the signals on the variation of *SNR*. More formally, the power of the *i*-th signal si is computed as follows:

$$P_{s_i} = \frac{1}{N_{sam}} \sum_{j=1}^{N_{sam}} s_{i_j}^2$$

where N_{sam} represents the total number of samples for each signal and s_{i_j} is the *j*-th temporal sample of the signal s*i*. The average power of the signals of all classes is:

$$\overline{P_s} = \frac{1}{N_{sig}} \sum_{i=1}^{N_{sig}} P_{s_i}$$

where N_{sig} is the total number of signals in all the classes.

We performed 100 trials. The process of noise generation has been repeated for each trial and for each sample of the signals, each time generating a new process of random noise n_k, $k = 1,..., N_{sam}$. For each trial we affected the test set with an increasing level of noise, multiplying each stochastic noise process by an increasing positive integer *NL*. In the experiments, we used 10 noise levels NL_h with $h = 1, ..., 10$:

$$NL_h \in \{5, 10, 15, 20, 25, 30, 40, 60, 80, 100\} .$$

In particular we affected the test data with noise balanced so as to obtain the same *SNR* for each accelerometer. Indeed the power of the signals of the accelerometers can be different and, for this reason, we increased the noise for the features of the accelerometer with higher power. The reason for that is that we are interested in analyzing the robustness of each accelerometer.

The *k*-th extracted Gaussian noise gives origin to 10 different noise signals:

$$n_{kh} = NL_h n_k , \quad h = 1,...,10.$$

Consequently, the power of the generic noise signal n_k becomes:

$$P_h = \frac{1}{N_{sam}} \sum_{k=1}^{N_{sam}} n_{kh}^2 .$$

Therefore, for each noise level NL_h, we compute the *SNR* as follows:

$$SNR_h = 10 \log_{10} \frac{\overline{P_s}}{P_h} .$$

Table VIII shows the appreciable level of robustness to noise of our classification system over 100 trials. We obtain very good results up to $SNR \approx 16.52$ db (accuracy higher than 90.00 %) and acceptable results (accuracy between 80.00 % and 90.00 %) up to $SNR \approx 9.59$ db.

TABLE VIII. QDC CLASSIFIER. TEST SET WITH NOISE

NL	SNR (db)	Accuracy Mean±Std.Dev.	NL	SNR (db)	Accuracy Mean±Std.Dev.
5	40.55 db	99.35 ± 0.22 %	30	9.59 db	86.20 ± 1.01 %
10	28.62 db	97.93 ± 0.15 %	40	4.59 db	77.41 ± 1.25 %
15	21.47 db	96.24 ± 0.47 %	60	-2.75 db	62.44± 1.78 %
20	16.52 db	92.50 ± 0.77 %	80	-7.61 db	51.55 ± 1.49 %
25	12.45 db	89.75 ± 0.74 %	100	-11.35 db	44.28 ± 0.90 %

In order to increase the robustness to noise, we adopted MLP and RBF neural networks, and compared them with the QDC classifier.

We used MLPs with one hidden layer and all neurons characterized by a logarithmic sigmoid transfer function. We tried different numbers of hidden neurons (10, 15, 20, 25, 30, 35, 40, 45, 50). We adopted RBFs with one hidden layer and all neurons characterized by a gaussian transfer function. We tried different numbers of hidden neurons (10, 15, 20, 25, 30, 35, 40, 45, 50) and different spread values (0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1, 1.1, 1.2, 1.3).

The inputs to the MLPs and RBFs are the SFs previously selected by QDC, i.e., the features 296 and 295 of the second accelerometer and the features 277 and 296 of the third accelerometer.

The MLP with 50 hidden neurons provides the best performance among all the MLPs, while, among the RBFs, the best performance is obtained by the RBF with 45 hidden neurons and a spread of 1.2 (Tables IX and X).

TABLE IX. MLP WITH 50 HIDDEN NEURONS. TEST SET WITH NOISE

NL	SNR (db)	Accuracy Mean±Std.Dev.	NL	SNR (db)	Accuracy Mean±Std.Dev.
5	40.55 db	99.62 ± 0.01 %	30	9.59 db	91.63 ± 0.84 %
10	28.62 db	99.32 ± 0.21 %	40	4.59 db	84.73 ± 0.65 %
15	21.47 db	98.48 ± 0.26 %	60	-2.75 db	72.67 ± 0.81 %
20	16.52 db	96.51 ± 0.47 %	80	-7.61 db	63.75 ± 1.75 %
25	12.45 db	93.87 ± 0.76 %	100	-11.35 db	56.00 ± 1.24 %

Comparing the results in Tables VIII, IX, and X, we can see that the MLP and RBF do increase the robustness to noise. In particular, both of them improve both the good results (accuracy higher than 90.00 %) up to 9.59 db, and the

acceptable results (accuracy higher than 80.00 %) up to 4.59 db.

TABLE X. RBF WITH 45 HIDDEN NEURONS AND SPREAD=1.2. TEST SET WITH NOISE

NL	SNR (db)	Accuracy Mean±Std.Dev.	NL	SNR (db)	Accuracy Mean±Std.Dev.
5	40.55 db	99.63 ± 0.07 %	30	9.59 db	92.11 ± 0.59 %
10	28.62 db	99.43 ± 0.18 %	40	4.59 db	84.16 ± 1.05 %
15	21.47 db	99.05 ± 0.19 %	60	-2.75 db	69.41 ± 1.10 %
20	16.52 db	97.13 ± 0.33 %	80	-7.61 db	55.92 ± 1.63 %
25	12.45 db	94.21 ± 0.76 %	100	-11.35 db	47.00 ± 1.19 %

Analyzing the results in greater detail, we can affirm that the RBF obtains higher accuracy and lower standard deviation for the first levels of noise (5, 10, 15, 20, 25, 30) compared to the QDC and the MLP. However, for the subsequent levels of noise (40, 60, 80, 100), the performance of the RBF starts to decrease and becomes quite similar to the one achieved by the QDC classifier. The MLP provides better accuracy for all the different levels of noise compared to the QDC, not only concerning the average accuracy but also the standard deviation, proving to be more stable to the noise compared to the QDC classifier. Besides, even though the RBF is better than the MLP for the first levels of noise, the MLP offers a more graceful performance degradation for high levels of noise.

Thus, for acceptable levels of noise, the best results and, consequently, the best robustness are obtained by the RBF, while, for higher levels of noise, the MLP results to be the best. Figure 3 clarifies the comparison among these three classifiers (QDC, MLP and RBF).

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

IV. CONCLUSIONS

In this paper we have presented an automatic method, based on classification techniques, for diagnosing defects of rolling element bearings.

The proposed method has been applied to experimental data, registered by four accelerometers, and related to four different defects on rolling bearings and different levels of severity for one of them. The method has proved to be highly sensitive both to different defects and to different degrees of severity for the considered defects. We achieved an accuracy on the test set higher than 99.00 %.

We have also performed a noise analysis to assess the robustness of our method to noise. In particular, we have classified the noisy signals by means of a classifier trained on signals without noise. The appreciable levels of robustness to noise achieved could be even increased if, in

real world situations, we could filter noise out of the acquired signals before their classification. Alternatively, or in addition, we could train the classifier both with signals without noise and with signals with added noise, choosing the noise level to be added depending on the specific situation of interest.
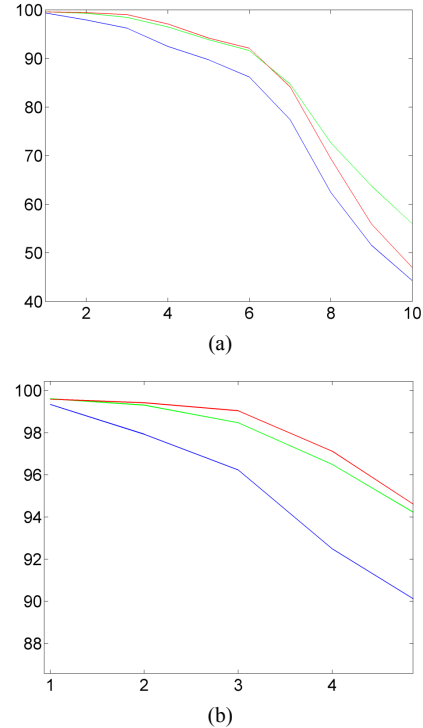

Figure 3. (a) Robustness of QDC (blue), MLP (green), RBF (red) to the noise. x-axis: level of noise, y-axis: accuracy. (b) Zoom of (a) for the first levels of noise.

The main novelties of the method are, therefore, automatic feature selection and fault classification, multi-class classification, higher sensitivity with respect to traditional vibration analysis techniques, and good robustness to noise.

REFERENCES

[1] A. H. C. Tsang, "Condition-based maintenance: tools and decision making", *Journal of Quality in Maintenance Engineering*, vol. 1, no. 3, 1995, pp. 3-17.

[2] J. H. Williams, *Condition-based maintenance machine diagnostics*, Kluwer Academic Publisher, London UK, 1994.

[3] B. Li, M.-Y Chow, Y. Tipsuwan, and J. C. Hung, "Neural-Network-Based Motor Rolling Bearing Fault Diagnosis", *IEEE Transaction Industrial Electronics*, vol. 47, no. 5, 2000, pp. 1060-1069.

[4] M. Cococcioni, P. Forte, S. Manconi, and C. Sacchi, "Rolling bearing monitoring using classification techniques", *8th International Conference on Vibrations in Rotating Machines (SIRM'09)*, Vienna,

Austria, February 2009.

[5] B. Samanta, K. R. Al-Balushi, and S.A. Al-Araimi, "Artificial neural networks and genetic algorithm for bearing fault detection", *Soft Computing*, vol. 10, 2006, pp. 264-271.

[6] B. Samanta, K. R. Al-Balushi, and S. A. Al-Araimi, "Bearing fault detection using artificial neural networks and genetic algorithm", *Journal on Applied Signal Processing*, vol. 3, 2004, pp. 366-377.

[7] N. Nguyen, and H. Lee, "Bearing fault diagnosis using adaptive network based fuzzy inference system", *International Symposium on Electrical & Electronics Engineering*, HCM City, Vietnam, 2007.

[8] N. Tandon, and A. Choudhury, "A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings", *Tribology International*, vol. 32, no. 8, 1999, pp. 469-480.

[9] A. Webb, *Statistical Pattern Recognition, 2nd Edition*, John Wiley & Sons, New York, 2002.

[10] S. Haykin, Neural Networks and Learning Machines, 3rd Edition, Prentice Hall, 2008.

[11] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley Interscience, New Jersey, USA, 2004.

[12] T. G. Dietterich, "Ensemble methods in machine learning", *in J. Kittler and F. Roli, editors, Multiple Classifier Systems, Lecture Notes in Computer Science*, vol. 1857, Springer, Cagliari Italy, 2000, pp. 1-15.

[13] R. P. W. Duin, P. Juszczak, and P. Paclik, *PRTools4: A Matlab Toolbox for Pattern Recognition, version 4.1*, Delft Pattern Recognition Research, Delft University of Technology, Delft, The Netherlands, August 2007.