IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 26, 2016, accepted October 13, 2016, date of publication October 27, 2016, date of current version November 18, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2622679

Advances in Classifier Evaluation: Novel Insights for an Electric Data-Driven Motor Diagnosis

IGNACIO MARTIN-DIAZ[1,2], (Student Member, IEEE), DANIEL MORINIGO-SOTELO[1], (Member, IEEE), OSCAR DUQUE-PEREZ[1], AND RENE DE J. ROMERO-TRONCOSO[2], (Senior Member, IEEE)

[1]Escuela de Ingenierías Industriales, University of Valladolid, 47002 Valladolid, Spain
[2]CA Telematica at DICIS, University of Guanajuato, 37320 León, Mexico

Corresponding author: R. de J. Romero-Troncoso (troncoso@hspdigital.org)

This work was supported in part by the Universidad de Guanajuato, DAIP, under Grant 733/2016, in part by the Mexican Council of Science and Technology (CONACyT) under Grant 598078, in part by a Mobility Grant from the University of Valladolid, Spain, and in part by the Spanish Ministerio de Economía y Competitividad and FEDER Program in the framework of the Proyectos I+D del Subprograma de Generación de Conocimiento, Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia under Grant DPI2014-52842-P.

ABSTRACT Fault diagnosis of inductions motors has received much attention recently. Most of the works use data obtained either from the time domain or by applying advanced techniques in the frequency domain. Some researchers have employed a considerable effort in designing sophisticated algorithms to achieve the best performance of the diagnosis system. However, some contributions in the field have not taken advantage of the benefits that a good evaluation stage can bring to the developing of classifiers for fault diagnosis. In this paper, novel insights for the classifier evaluation are presented to promote better assessment practices in the field of electric machine diagnosis based on supervised classification. A case of study consisting of a motor with a broken rotor bar is described to analyze the performance of two classifiers by using scores focused on the fault detection. Also, different error estimation methods are considered to obtain unbiased predictive performances. Two statistical tests are also discussed to confirm the significance of the results under a single data set.

INDEX TERMS Broken rotor bar, classification algorithm, condition monitoring, electric machines, fault diagnosis, performance evaluation.

I. INTRODUCTION

Induction motors (IMs) are a fundamental part of any current industrial facility, due mainly to their robustness, reliability, low price and lower maintenance requirements. Nonetheless, maintenance policies are required to avoid faults of critical motors in an industrial process. The success of any diagnosis system is crucial for future predictive maintenance programs [1]. The prime objective of diagnosis is to detect faults as early as possible [2] and to discriminate between different types of faults [3] to prevent harmful consequences. Fig.1 shows a typical data-based fault diagnosis scenario.

The range of approaches to data-driven fault diagnosis is broad, from intelligent algorithmic tools [3]–[7] to statistical techniques for time-dependent signals [8], [9]. A significant number of research papers describe the use of Machine Learning (ML) procedures to diagnose IM. These methods include supervised learning techniques [10]–[13], where a known data set (containing input data and target values) is employed to make predictions on unknown observations. Exploratory unsupervised learning, where the data labels are not available [14], [15] as semi-supervised learning [16], with only some data labels available, also belong to the ML family. Regression of continuous-response values [17] and reinforcement learning based on an agent policy optimization in a rewarding setting have also been considered [18].

Nevertheless, in this broad context, there are still some unsolved concerns. One of them is the problem of choosing the best algorithm or combination of algorithms for the diagnosis phase, according to its prediction performance. A feasible implementation of the classifier avoiding unnecessary spending of resources is another open matter. These questions certainly point out that there is a need to find the best procedure to evaluate the classifier performance but always according to a specific practical context. Therefore, it is advisable to understand the evaluation process to make a critical analysis of the performance classifier,

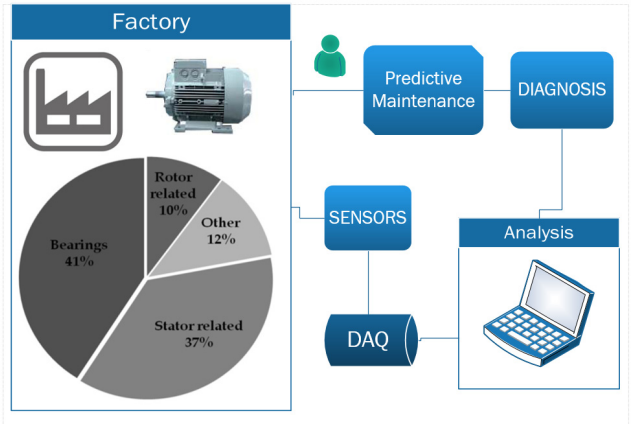

FIGURE 1. Typical data-based fault diagnosis scenario.

applied to a particular situation [19]. Depending on the diagnosis system requirements, the advantages and disadvantages of each step of the evaluation scheme should be considered. The application context (or the role that the diagnosed motor plays in the industrial process) may result important for estimating the performance of the classifier and do a proper evaluation.

Furthermore, it has been observed, from the analysis of the state-of-the-art of IM fault diagnosis systems, that the performance measurement of classifiers is very heterogeneous. It can be concluded that the evaluation criteria used do not permit performance comparisons of the results [20]–[23]. The assessment of the classifier in some papers is very limited because it is merely based on accuracy and error rate. Others use the confusion matrix (CM) without exploring the scores that can be calculated from it. Besides, regardless of the score, it is advisable to consider the error estimation in order to optimize the tuning parameters of the classifier properly. Moreover, other aspects may be taken into account in the diagnosis design stage, such as the size of the training data, the qualitative or quantitative nature of the data observations or the imbalance between classes [24], [25].

Therefore, there is a need and an opportunity to present the most recent ML literature contributions, concerning performance evaluation, to help the researchers to design more homogeneous and comparable evaluation processes. This performance evaluation has three steps: 1) to measure the classification quality (performance measures) for a particular goal; 2) to estimate the classifier quality with error estimation methods; and 3) to observe if the classifier behavior presents significant differences using statistical tests..

The contribution of this paper is to present a novel insight, with the aim to homogenize the performance evaluation of IM fault classifiers keeping in mind the diagnosis tool to be developed. This paper is an extended contribution presented in [26]. It includes a wider set of experimental results and describes in detail the performance evaluation of the classifiers. The case of study for experimentation consists in an IM having a broken rotor bar.

II. CLASSIFIER PERFORMANCE EVALUATION

As aforementioned, some questions related to the classifier evaluation need to be addressed. These questions can be answered using performance metrics focused on the diagnosis goals. Various methods can provide good results depending on the dataset characteristics [19]. This step needs to be carefully studied to obtain estimations close to the true measured value [24]. Finally, the use of the available statistical significance tests, with their intrinsic limitations [27]–[29], can be carried out to observe whether the results are attributed to a real classifier behavior or if they are obtained by chance. These tests may also provide information about the representative character of the dataset to solve the real problem. All these aspects are explained in the following subsections.

A. PERFORMANCE METRICS

When designing a classifier, it is required to choose proper metrics to assess the classifier quality. The use of a particular set of scores depends, among other things, on the application domain, case characteristics, data set features and the variable or condition to be diagnosed. Performance measures can be classified according to the return values. The following performance scores belong to the single value assessment category: Accuracy, True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), F-measure and G-mean. Among dual value scores can be mentioned: Precision-Recall Curves and Sensitivity-Specificity Curves.

Alternatively, results can be presented in a table form as a CM, or graphically with Receiver Operating Characteristic (ROC) Curves, Cost Curves, Lift Chart, etc. [19], [24]. Additionally, some supervised classification problems using probabilistic models can be evaluated with calibration scores [19].

Frequently, the choice of inappropriate performance measures causes misleading classifier evaluations. The convenience of one measure over other performance metrics will depend on the research objectives. Sometimes the optimization criteria may vary with the application: robustness of the diagnosis systems for detecting simultaneous faults on a multi-fault scenario; priority on some particular false positive rate versus a given false negative rate; state of degradation of a particular type of fault, etc. Therefore, the classifier can be considered as a multi-objective problem. Thus, for these purposes not every metric contributes with the same information to solve the problem and it is required the use of additional measures. Furthermore, for those models with the same value of their metrics, it is preferable to choose the simplest one according to the Occam's razor definition [30]. Next, the most useful and common performance metrics in a discrete scenario are described.

1) CM AND SCORES

Unlike regression problems, in classification, the empirical risk [19] is typically obtained with the following

loss function:

$$L(y, y_e) = \begin{cases} 1 & if\ y \neq y_e \\ 0 & otherwise \end{cases} \quad (1)$$

where y_e is the predicted class of the observations (or in the case of probabilistic classifiers, the estimated probabilities of class memberships) and y is the true response of the observations. A CM provides general information about the supervised classifier performance according to the assignments to every class of interest [19]. For the sake of simplicity, a binary (two-class) problem is considered and its CM is shown in Table 1. Multiclass problems can be treated using One Against One (OAO) or One Against All (OAA) approaches [19], [24] in order to convert them into a set of binary problems.

TABLE 1. Confusion matrix for a two-class problem.

Actual	Prediction		
	$Class_1$	$Class_2$	TOTAL
$Class_1$	True Positive (TP)	False Negative (FN)	P_{actual}
$Class_2$	False Positive (FP)	True Negative (TN)	N_{actual}
TOTAL	P_{Pred}	N_{Pred}	M

In Table 1, M is the number of total instances; subscripts Pred and actual indicate whether they are predicted or actual instances, respectively. Finally, $Class_1$ (positive) and $Class_2$ (negative) are the considered classes.

a: ACCURACY AND CLASSIFICATION ERROR (1-ACCURACY)

One of the most used scores to evaluate discrete classification in electrical machines diagnosis is accuracy, also known as predictive accuracy. This score is indicative of the classification error committed evaluating two or more classes.

$$Accuracy = \frac{TP + TN}{M} \quad (2)$$

When classifying several classes, this score is optimistic since all classification errors are considered equally, and each class is not evaluated individually. Two classifiers could have the same accuracy but provide a different classification for each class. In such a case, the characteristics of the data are crucial when various types of faults have different implications [31]. Besides, in practical cases, classes are usually imbalanced [25]. For example, in most industrial environments, there will be more data of the healthy class since this is the normal condition of the machines. The following scores deal with this problem and discriminate how each class is classified.

b: TNR (OR SPECIFICITY)

TNR (also known as Specificity in two-class problems) is a performance metric focused in the negative class.

$$TNR = \frac{TN}{TN + FP} \quad (3)$$

c: PRECISION

Conversely, this performance metric evaluates the correct classification of the positive class.

$$\text{Prec} = \frac{TP}{TP + FP} \quad (4)$$

d: TPR (RECALL OR SENSITIVITY)

This measure, along with specificity, provides a proportion of one-class samples correctly classified. However, similarly to Precision, it only evaluates the positive class.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

Recall is interesting for those cases where the positive class observations were not classified as positive during the training stage [24].

e: F-MEASURE

This score can help to solve any contradiction that may appear between Precision and Recall scores. F-measure leaves out the TN performance. Several versions exist, depending on the value assigned to α, allowing to choose how Precision and Recall are weighted. For $\alpha \in \mathbb{R}$, $\alpha > 0$, the general expression is shown below:

$$F_\alpha = \frac{(1 + \alpha)(\text{Prec} \times \text{Recall})}{(\alpha \times \text{Prec}) + \text{Recall}} \quad (6)$$

When $\alpha = 1$, the resulting expression is the harmonic mean of Precision and Recall. One of the limitations of this measure is that the correct value of the weight to make a good comparison is unknown *a priori*.

f: GEOMETRIC MEAN

This metric is proposed in [32] and gives information about the classifier performance on an imbalanced problem. There are two versions of this score. $G_{mean,1}$ is used to evaluate the relative balance of the classifier performance on all classes:

$$G_{mean,1} = \sqrt{TPR \times TNR} \quad (7)$$

The other version, $G_{mean,2}$, puts the focus on the positive class, taking Precision into account.

$$G_{mean,2} = \sqrt{TPR \times \text{Prec}} \quad (8)$$

From this section, it can be stated that metrics with a single class focus, as sensitivity and specificity, provide relevant information that accuracy does not. In the case of a problem with imbalanced classes, CM derived metrics have some disadvantages. The performance of continuous learning algorithms (soft classifiers as fuzzy-based ones) is not evaluated adequately using CM based measures because the decision

threshold cannot be modified. Another disadvantage may derive from the fact that different error costs are not considered [33]. A solution to evaluate an imbalanced problem is the use of some ratios, described in the literature, which consider the number of instances of each class [19], [24]. On the other hand, the impact of each type of classification errors in the maintenance field can be very different. For this reason, a cost matrix should be created to weight the classification errors committed for each class [31], regarding any previous experience.

In real situations, it may result difficult to estimate the magnitude of these costs that would make undoubtedly a difference in the training stage of the classifier. Therefore, in this context, the problem is to minimize the misclassification associated costs and not the number of errors made by the classifier. It is evident that there are many more performance metrics, especially for continuous and probabilistic classifiers. The goal of the paper is not to make an exhaustive review of all metrics but introduce those most significant for developing a more appropriate diagnosis evaluation.

2) GRAPHICAL PERFORMANCE METRICS

In certain situations, the use of classifiers that obtain the class memberships by using a threshold (scoring classifiers) is required. For these cases, it would be necessary to incorporate some extra information in addition to that deduced from the CM. Accordingly, graphical analysis methods serve as effective tools to describe the performance of such algorithms. There are many graphical methods addressed in the literature; in this section, some of the most standard ones are succinctly introduced: ROC Curves, Lift charts, Precision-Recall Curves and Cost Curves.

a: ROC CURVE

The ROC analysis was introduced in signal detection theory and allows visualizing the performance of a classifier using a graphical plot. This graph illustrates the performance of a binary classifier whereas its discrimination threshold is varied. The curve is created plotting the TP rate against the FP rate at various threshold settings. The range of all possible variations of each rate represents the operating range of a classifier. Unlike the F-measure and accuracy, this approach is insensitive to class imbalance [34]. Additionally, it is indicative of the whole operating range of the classifier and allows identifying optimal performance regions as well. The optimal point of operation can be chosen according to various formulations [19], [34], and a conversion to a score is possible [34]. There are more specialized metrics to measure the performance of continuous classifiers dealing with different learning strategies [38], [24]. In the case of classifiers formed by a combination of learning algorithms, as in hybrid classifiers, the Convex Hull may be considered [19], [34]. In such case, the ROC Convex Hull [34] enables visualizing optimal performance points for a particular classifier in the ROC space. This is particularly useful in classification scenarios where the cost of an FP is different from an FN one. An example is the diagnosis implications of critical and non-critical motors.

Auc (Area Under ROC Curve): This measure allows obtaining a scalar to compare the performance of classifiers, but information from the whole classifier operating range is lost [34]. Given two instances, randomly chosen from the positive and negative classes, the AUC represents the probability that the classifier classifies better the positive instance over the negative one [24]. Considering the Convex Hull from the ROC curve, it identifies the best classifier only if one dominates the other [34]. In the case of two classifiers whose ROC curves intersect each other, AUC would not provide adequate information about the comparison [24].

b: LIFT CHARTS

As with the ROC curves, Lift charts permit to visualize the true positives but in this case against the dataset size used to achieve such number of true positives [19]. For this chart, the true positives are plotted in the vertical axis whereas the horizontal one indicates the number of observations in the dataset taken into account for the true positives on the vertical axis. Particularly, in the electrical motor diagnosis domain, this is very useful because it shows which classifiers can identify faulty cases by using the smallest sample size.

c: Precision-Recall Curves

This kind of chart is also known as PR Curves and is discussed in [35]. It serves to analyze the balance between the positive examples correctly classified and the negative examples misclassified. Basically, it is a plot where the classifier Precision is represented as a function of the Recall values. In other words, in a fault diagnosis scenario, these curves represent the ratio of cases from true fault detected referred to those which are identified as healthy (vertical axis) and in the x-axis referred regarding the occurrence rate of false positive indications. These curves have proved to be successful when highly imbalanced data are present [35].

d: COST CURVES

This type of curves makes use of the relative known misclassification costs by plotting them directly instead of employing those based on ROC metrics [19]. The main advantage is their simple usage when deciding the most suitable classifier in those cases where the error cost, class distribution or the imbalanced proportion of the classes are known. The difference with the ROC curves is that these give more practical information for those circumstances where the required information is available (i.e. when the operator has enough reported knowledge from its functioning equipment and its consequences).

B. ERROR ESTIMATION METHODS

Once a particular set of performance metrics is chosen for the algorithm evaluation within the diagnosis problem, the next step is to determine the best error estimation method to prove

that the designed algorithm obtains an adequate bias-variance tradeoff with the best possible use of all available data. Many error estimation methods are discussed in [19] and [36] and they can be classified in the following categories: Resubstitution, Holdout, and Resampling [19].

Particularly, in the case of fault diagnosis in IMs, the input data frequently comes from a motor current signature analysis (MCSA) or a vibration signal analysis. The available data is limited, and it is generally used in its entirety to train the classifier. It is required to build a large data set of samples, enough for the testing stage, to make the best use of the acquired data. This is easily achieved by resampling methods. Basically, there are two kinds of methods: simple resampling and multiple resampling. In the latter, the following methods are included: Random Subsampling, Bootstrapping, Randomization and Repeated k-fold cross-validation.

The advantage of multiple resampling methods over simple resampling is the stability on the estimations. This fact results in a large sampling number, but this may lead to violating the independence assumption between the test and the training sets [37].

In large data sets, Resubstitution is a satisfactory method. The same data is used in the training and testing stages. Thus, the estimations become optimistically biased. Due to this bias, the use of this method with small data samples provides weak results. The Holdout method is an alternative to Resubstitution, where the data set is split into two exclusive groups, one for training and the other for testing. The former is usually larger. This method provides a pessimistic and biased estimation of the classification error. As the number of samples in the dataset increases, this bias is decreased. But, on the other hand, the variance of the error estimation increases. In the Repeated k-fold cross validation (CV) method, the data set is divided into k equal-size and mutually exclusive folds. All except one fold are used for training and the remaining for testing. As the name of this method suggests, this procedure is repeated as many times as the evaluator considers. The classification error is obtained by averaging each k-error in every repetition. In the k-fold CV method, when the number of folds is the same as the size of observations, the method is named Leave-one-out. This method has a higher computational cost, which makes it impractical for some applications.

Bootstrap [37] may be a better choice than the k-fold cross-validation in those cases where the data set is relatively small. There are multiple variants of bootstrap used in statistics. Zero bootstrap is an improved variant of the simple bootstrap, which suffers from overfitting [19]. Basically, it consists of sampling with replacement m instances uniformly from the data set, denoted by S. The likelihood of each instance being chosen is 1/m and the likelihood of not being chosen is:

$$\left(1 - \frac{1}{m}\right) \tag{9}$$

For any given observation, the likelihood of not being chosen after m samples, when m is large, is:

$$\left(1 - \frac{1}{m}\right)^m \cong \frac{1}{e} \cong 0.368 \tag{10}$$

Thus, the expected number of different instances in the resulting sample of m observations is:

$$(1 - 0.368) \cdot m = 0.632 \cdot m \tag{11}$$

Therefore, the test set, T_{boot} is formed by all observations of S not present in S_{boot}. A classifier f_{boot} is obtained with S_{boot} and is tested with T_{boot}. The empirical risk estimated of f_{boot} is obtained over T_{boot}. This procedure is repeated k times, and the respective empirical risk is averaged to obtain the estimator $boot.0_e$ that may result pessimistic since the classifier is typically trained using only 63.2% of the whole data set in each step.

As a consequence, *0.632 Bootstrap* [37] tries to correct the pessimistic bias taking into account the optimistic bias from the resubstitution error over the remaining fraction 0.368.

$$boot.0.632_e = \frac{1}{k}\sum_{i=1}^{k} 0.632 \cdot boot.0_{ei} + 0.368 \cdot err_e(C) \tag{12}$$

where $err_e(C)$ represents the resubstitution error of the classifier C obtained with the training set, S. The latter variant may lead to estimations with lower variance as a result of increasing the dataset size.

An improvement to this cross validation method is the *0.632 plus bootstrap* (or 0.632+ bootstrap) [38], which corrects bias when there is a great amount of overfitting. Indeed, the weights are assigned individually for each model to reveal the goodness in creating the training sample. This version considers the error for cases in which dependent and independent variables were not associated. The following expressions are used to calculate the estimation error with this method:

$$boot.0.632+_e = (1 - \omega_e) \cdot err_e(C) + \omega_e \cdot boot.0_{ei} \tag{13}$$

$$\omega_e = \frac{0.632}{(1 - 0.632R_e)} \tag{14}$$

$$R_e = \frac{boot.0_{ei} - err_e(C)}{\gamma_e - err_e(C)} \tag{15}$$

$$\gamma_e = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\delta(C_i, (\varphi_x^j))}{n^2} \tag{16}$$

where γ_e is the non-information rate estimated by evaluating the prediction model on all possible combinations of targets and predictors. Generally, the number of repetitions k (Bootstraps) to achieve an appropriate estimation is much larger than for the cross-validation case, being usual values where $k \geq 200$ [19].

C. STATISTICAL TESTS

Using the aforementioned performance metrics, a better understanding about the desired classifier behavior according to the most critical aspects is achieved. Secondly, the

error estimation methods focused on the stability of the outcomes regarding slight changes in the training and test sets composition have been considered. Finally, the statistical tests (STs) help the researcher to be more precise to verify the significance of the results. Indeed, the importance of an appropriate choice of the significance test must be considered due to their implications to confirm the differences in the classifier performance [19]. This step can help to control the error probability of declaring a model better than others. For these purposes, several statistical tests have been discussed according to a wide variety of ML scenarios [24]. It is not the aim of the paper to analyze each test within the field of ML. However, some of those related to the use of a single dataset are introduced, which is a usual situation for the algorithm comparison employed in electrical machines data-based diagnosis.

In fact, an appropriate estimation of the chosen scores is not enough to evaluate differences among classifiers since the estimation does not take into account uncertainties related to the estimation process. It is true that an ST is not definitive to conclude the comparison due to its limitations and misuses, but it is still valid to evaluate the classifier outcomes when possible [19], [24]. However, leaving the criticisms aside, two of the most useful STs are described next:

1) TWO-MATCHED SAMPLES t-TEST

This is a parametric type test, i.e. it requires fulfilling each of the assumptions under consideration to be applied correctly. Its validity depends on the three following assumptions:

Normality: one requirement for this test is that the populations have to come from a normal distribution. The behavior under this assumption is quite robust, and it is required that the test set has at least 30 samples. Some of the most practical normality tests are the Kolmogorov-Smirnov test, the Shapiro-Wilk test or the Anderson-Darling test.

Randomness of the samples: This point supposes that the scores are representative of the underlying distribution, i.e. the scores have been calculated in such a way (method of estimating the error) that allows characterizing the distribution without biasing the sampling process.

Equal Variances: the paired t-test assumes that the two sample populations have the same variance. The similarity of variances can be checked with the F test or the Bartlett's test.

Given two samples, the goal of the test is to verify if there is a significant difference between the means. This appreciation is realized by looking at the first and second moments of the samples (mean and standard deviation). A statistical hypothesis test serves to make inference about the two datasets under study through the confirmation of the null hypothesis rejection, which was assumed first. Before starting with the test, normality is adopted, and the null hypothesis considers a zero difference between these means. To see whether the hypothesis can be rejected, it is needed to find out what differences can be expected just by chance (those related to the normal distributions). This difference is checked with the *t-statistic*:

$$t = \frac{\overline{sc}(C_1) - \overline{sc}(C_1)}{\frac{\sigma_{\mathrm{d}}}{\sqrt{n}}} \tag{17}$$

where:

$$\bar{d} = \overline{sc}(C_1) - \overline{sc}(C_2) \tag{18}$$

is the difference of the means of the scores under consideration by applying classifiers C_1 and C_2; σ_{d} denotes the sample standard deviation defined by:

$$\sigma_{\mathrm{d}} = \sqrt{\frac{\sum_{i=1}^{n} (d_{\mathrm{i}} - \overline{d})^2}{n-1}} \tag{19}$$

where d_{i} expresses the difference between the scores for each classifier at the trial i:

$$d_{\mathrm{i}} = sc(C_1) - sc(C_2) \tag{20}$$

n is the number of trials. The average value of the score is calculated as follows:

$$\overline{sc}(C) = \frac{1}{n} \sum_{i=1}^{n} sc_i(C) \tag{21}$$

Finally, the obtained t value must be compared with the probability values found for the assumed distribution. If this p-value (output probability) is small (for the significance level established) then the null hypothesis should be rejected. Otherwise, there is no evidence to conclude that there is a difference between the results.

2) McNemar's TEST

McNemars's test is a non-parametric test and it is only advisable for those cases where the assumptions on the distribution of the performance measures are not met [19]. In general, this test is applied to compare the classification errors of the two classifiers. Once the training and test sets are obtained separately, the classifiers are tested on the test set, and afterward the McNemar's contingency table (Table 2) is obtained. The elements of this table are computed as follows:

$$C_{Mc,00} = \sum_{i=1}^{|S_{test}|} [I\,(C_1(x_i) \neq y_i) \wedge I\,(C_2(x_i) \neq y_i)] \tag{22}$$

$$C_{Mc,01} = \sum_{i=1}^{|S_{test}|} [I\,(C_1(x_i) \neq y_i) \wedge I\,(C_2(x_i) = y_i)] \tag{23}$$

$$C_{Mc,10} = \sum_{i=1}^{|S_{test}|} [I\,(C_1(x_i) = y_i) \wedge I\,(C_2(x_i) \neq y_i)] \tag{24}$$

$$C_{Mc,11} = \sum_{i=1}^{|S_{test}|} [I\,(C_1(x_i) = y_i) \wedge I\,(C_2(x_i) = y_i)] \tag{25}$$

where the number of examples misclassified by the two classifiers is expressed as: $C_{\mathrm{Mc},00}$. $C_{\mathrm{Mc},01}$ represents the number of misclassified observations by C_1 but correctly assigned by C_2; $C_{\mathrm{Mc},10}$ denotes the reverse case and $C_{\mathrm{Mc},11}$ is the

TABLE 2. McNemar's test contingency table.

		Classifier C_2	
		0	1
Classifier C_1	0	$C_{Mc,00}$	$C_{Mc,01}$
	1	$C_{Mc,10}$	$C_{Mc,11}$

number of instances correctly classified by C_1 and C_2 at the same time. For this test, the null hypothesis considers that both classifiers have the same performance. Then, the statistic, which is an approximation of the χ^2, is computed as:

$$\chi^2_{Mc} = \frac{\left(\left|C_{Mc,01} - C_{Mc,10}\right| - 1\right)^2}{C_{Mc,01} + C_{Mc,10}} \tag{26}$$

Finally, the statistic obtained in (26) is compared against the χ^2 distribution values found in the typical tables. The null hypothesis is rejected if the obtained value surpasses those related to the considered level of significance. There are some cases in which the McNemar's test statistic cannot be approximated by χ^2 distribution and should be replaced by others (e.g. binomial distribution). This is particularly true for those cases in where only a limited amount of examples is available [19].

III. CASE OF STUDY

A case of study is presented in this section to illustrate the application of the performance evaluation procedures. The laboratory experiment consists of a destructive test of a low voltage commercial IM.

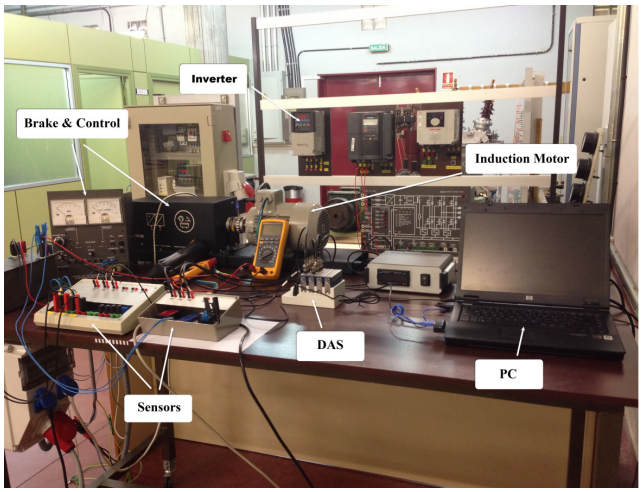

FIGURE 2. General view of the laboratory setup.

A. LABORATORY SETUP

A layout of the laboratory setup can be seen in Fig.2. An IM, star connected and fed directly from the line and from an inverter, was tested to collect data for this study. This squirrel cage motor with two pole pairs has a rated power of 0.75 kW, at 400 V_{AC}. The rated current is 1.9 A at a rated speed of 1395 RPM. The inverter is an Allen Bradley, Power Flex 40.

The motor is loaded with a magnetic powder brake and tested under two load conditions (medium and high load level). The operating frequency is 50 Hz. The stator current is acquired by a Hall-effect current transducer by LEM. A National Instruments NI cDAQ-9174 base platform with an NI9215 acquisition module is used for data acquisition, with a sampling frequency of 80 kHz and a steady-state sampling time of 10 s.

The motor is tested first under healthy conditions. Faulty conditions are produced by drilling a hole in one of the rotor bars. A partially broken bar is produced with a depth hole of 12 mm, and finally, a full-broken bar is obtained drilling an 18 mm hole. The characteristic fault harmonics, represented by the Lower Sideband Harmonic (LSH) and Upper Sideband Harmonic (USH) are extracted from the steady-state stator current as fault patterns. For this purpose, the power spectral density (PSD) of the stator current is computed with the Fast Fourier Transform. The RMS stator current values and the motor slip are considered as variables as well. Half of the trials are obtained with the motor fed from the line and the remaining ones with the inverter supply.

The number of classes chosen by using the OAA approach, the number of tests performed and the division of data for balanced classes are as follows: Class 1, or positive class, is the faulty motor having 120 instances (60 of a partially broken rotor bar and 60 of a fully broken rotor bar); and Class 2, or negative class, is the healthy motor with 120 instances.

B. EVALUATION

Two different classifiers (implemented under R [39]) are considered: Decision Tree (DT), which is implemented by the CART (Classification and Regression Trees) algorithm [40] and Support Vector Machine (SVM) [41] with a linear kernel. As the main goal is just to illustrate the differences between some performance measures, the classifiers are not optimized.

Firstly, a Holdout method is used to calculate the CM of each classifier using the balanced dataset divided into two groups: 50% for training and 50% for testing.

The CM and some other related scores are shown in Table 3. Some conclusions can be drawn from it. If only the accuracy score is considered, as it is usual in other works, the outcome will be that both algorithms show the same performance. Yet, other scores, derived from the CM, provide more information for the particular faulty class of interest. Table 3 shows that Recall and Gmean2 have different values for each classifier, and they give worthy information about the classifier performance on the faulty (positive) class. DT classifies slightly better than SVM the negative class (Healthy motor), which can be more interesting in a different scenario where this class may be more important. Nonetheless, SVM presents a better behavior for the faulty

TABLE 3. Confusion matrix and associated scores for the comparison between CART and SVM.

Confusion Matrix							
Decision Tree (CART)				Support Vector Machine (SVM)			
Actual	Prediction		Error	Actual	Prediction		Error
	F	H			F	H	
F	58	2	0.033	F	59	1	0.017
H	0	60	0.000	H	1	59	0.017

Scores	Classifier	
	DT (CART)	SVM
Accuracy/Classification Error	0.9833/0.0167	0.9833/0.0167
TNR	1.0000	0.9833
Recall (TPR)	0.9667	0.9833
Precision	1.0000	0.9833
F-measure (α=1)	0.9831	0.9833
G-mean,1	0.9832	0.9833
G-mean,2	0.9832	0.9833

observations, which is confirmed by the values of the Precision, Recall, and the F-measure. Thus, in a practical case where a false positive has not serious consequences, SVM could be more advisable (e.g., for critical motors where a non-detected failure could have a significant impact on the industrial process). However, for those cases where a false positive has costly implications (for example, when motor inspection costs are high) DT could be chosen under this balanced scenario. Once more, if accuracy had been the only score computed, the comparison between these two classifiers would have been inconclusive, as both have the same value in this case of study.

Secondly, to analyze the provided information by each score under imbalanced situations, different cases are studied and shown in Table 4. For that, the faulty class instances are reduced according to the imbalanced ratios (IRs) considered (the definition of the IR can be found in [25]), which are IR=10, IR=5 and IR=2. Now, the CM is obtained from a stratified 10-fold CV method. Under imbalanced situations, the class-oriented scores are highly required to observe the performance of the class of interest. For the IR=5 and IR=10 imbalanced situations, SVM presents clearly better behavior than DT classifier as shown by these scores. Nevertheless, for the IR=2 case the results are quite close, showing differences in the Recall and Precision scores as expected. The Recall and Precision values for the DT are similar, but on the other hand, SVM presents more distant values among them. Indeed, it is observed that as the imbalanced ratio is reduced being close to the balanced case, similar results to Table 3 are achieved. As mentioned earlier, this fact may somewhat determine the future behavior of the diagnosis tool under certain scenarios.

Additionally, returning to the balanced dataset, some graphical curves are computed (with the ROCR R package [42]) for the case of study, which are shown in Figs. 3-4. These curves provide information about the performance of these two classifiers in the whole operating range, obtained by varying the decision threshold. In Fig. 3, it is clearly observed that SVM presents a better performance in the ROC space, which is confirmed by its AUC value ($AUC_{SVM} = 0.97 > AUC_{DT} = 0.95$). On the Precision-Recall curves (Fig. 4) it is seen that SVM has a higher percentage of faulty classified instances identified as faulty against the percentage

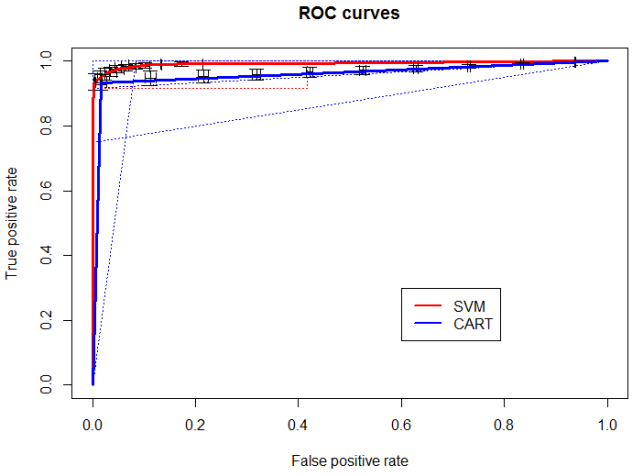

FIGURE 3. ROC curves of the two classifiers by a stratified 10-fold CV procedure.

TABLE 4. Scores for the comparison of CART and SVM under imbalanced data (by stratified 10-fold CV).

Imbalanced Ratio	Classifiers	Accuracy	F-measure	$G\text{-mean}_2$	Precision	Recall
10	DT	0,9924	0,9600	0,9608	1,000	0,9231
	SVM	1,0000	1,0000	1,0000	1,0000	1,0000
5	DT	0,9861	0,9600	0.9608	1,000	0,9231
	SVM	1,0000	1,0000	1,0000	1,0000	1,0000
2	DT	0,9689	0,9567	0,9535	0,9567	0,9503
	SVM	0,9306	0,9565	0,9584	0,9167	1,0000

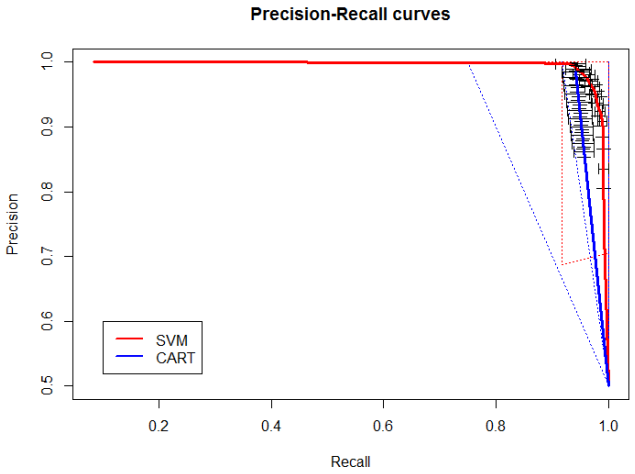

FIGURE 4. Precision-Recall curves of the two classifiers by a stratified 10-fold CV procedure.

of faults considered as such with respect all test observations. Therefore, focusing on the faulty cases, the SVM classifier presents a better performance behavior.

TABLE 5. Error estimation results using 10-fold cv and 0.632+ bootstrap for the balanced case.

Classifiers	Error Estimation Method	
	10-fold CV	0.632+ bootstrap
Decision Tree (CART)	0.0166	0.0167
Support Vector Machine	0.0164	0.0166

Regarding the error estimation, Table 5 shows results for the averaged accuracy metric by using 10-fold CV and 0.632+ bootstrap, for each classification algorithm in question. In this case, both classifiers show that the 0.632+ bootstrap presents a little bit more pessimistic behavior regarding the 10-fold cross validation method. Nevertheless, its use is suggested due to the improvement in the predictive behavior (bias-variance trade-off) under a small data set size [38].

Finally, to verify the significance of the previous results, the possible application of a significance test is checked. The procedure used is a stratified 10-fold cross-validation. Taking into account the Recall, Precision, and F1 (F-measure with $\alpha = 1$) metrics, and according to the Shapiro-Wilk test, the normality conditions are not fulfilled. Every p-value from each set of metrics is less than 0.05 for a significance level of 95%. Thus, the null hypothesis cannot be rejected, i.e. the observations do not come from the Gaussian distribution. For this reason, the parametric two-matched samples t-Test cannot be applied. On the other hand, because the following assumption is far from being fulfilled,

$$C_{Mc,01} + C_{Mc,10} < 20 \tag{27}$$

the McNemar's test is not used in this case [19]. This is due to the small number of observations. All this means that statistically speaking, significance testing does not present the correct assumptions for being applied. Therefore, the significance of the difference in the classifiers performance cannot be confirmed with a statistical test.

IV. CONCLUSIONS

The choice of a correct classifier depends enormously on the accurate evaluation of its performance. In this paper, some useful performance measures are reviewed. This proposal facilitates to obtain additional information about the classifier behavior when the interest of one class (e.g. a particular type of failure) predominates over the other (healthy state). On the other hand, the error estimation methods serve to achieve better predictions according to the data availability and also to accomplish more stable measures regarding the bias-variance trade-off. Besides, the correct use of statistical testing allows confirming the significance of the classifiers performance results. As it can be inferred from the previous points, this insight is useful to compare fairly new proposed classifiers to build diagnosis approaches. Also, this can be used to improve their design criteria (i.e. its optimization according to its true fault indications rather than the accuracy value only, taking into account the higher priority of some rotor states, etc.). A well understanding on the evaluation procedure may motivate the incipient faults diagnosis by focusing on their implications in the multi-objective classification scheme. For example, different degrees of damage in an IM suppose different risks in the maintenance scheduling. This makes necessary to assess the diagnosis of the most critical cases differently, with more severe implications for those situations where the operating motor does not endanger its continuity in the application. A large set of published works in the evaluation of supervised classification approaches in the last years has been analyzed to help readers identify the most suitable contributions to be used for diagnosis purposes of electric machines. Generally speaking, the presented analysis applied to rotor fault diagnosis can also be used for many data-driven fault diagnosis systems by employing classification. These considerations may lead to choosing those classifiers with reduced occurrence of false diagnosis for a more goal-oriented strategy in predictive maintenance. In the last instance, by improving the diagnosis method, it is increased the personnel safety, the continuity in operation of the electrical equipment, and the reduction of costs originated by unneeded maintenance interventions.

REFERENCES

[1] N. E. I. Karabadji, H. Seridi, I. Khelf, N. Azizi, and R. Boulkroune, "Improved decision tree construction based on attribute selection and data sampling for fault diagnosis in rotating machines," *Eng. Appl. Artif. Intell.*, vol. 35, pp. 71–83, Oct. 2014.

[2] R. J. Romero-Troncoso, D. Morinigo-Sotelo, O. Duque-Perez, P. E. Gardel-Sotomayor, R. A. Osornio-Rios, and A. Garcia-Perez, "Early broken rotor bar detection techniques in VSD-fed induction motors at steady-state," in *Proc. 9th IEEE Int. Symp. Diagnostics Electr. Mach., Power Electron. Drives*, Aug. 2013, pp. 105–113.

[3] R. J. Romero-Troncoso *et al.*, "FPGA-based online detection of multiple combined faults in induction motors through information entropy and fuzzy inference," *IEEE Trans. Ind. Electron.*, vol. 58, no. 11, pp. 5263–5270, Nov. 2011.

[4] L. Capocchi, S. Toma, G. A. Capolino, F. Fnaiech, and A. Yazidi, "Wound-rotor induction generator short-circuit fault classification using a new neural network based on digital data," in *Proc. 8th IEEE Symp. Diagnostics Electr. Mach., Power Electron. Drives*, Sep. 2011, pp. 638–644.

[5] C. Delpha, H. Chen, and D. Diallo, "SVM based diagnosis of inverter fed induction machine drive: A new challenge," in *Proc. Ind. Electron. Conf.*, Oct. 2012, pp. 3931–3936.

[6] F. Kadri, S. Drid, F. Djeffal, and L. Chrifi-Alaoui, "Neural classification method in fault detection and diagnosis for voltage source inverter in variable speed drive with induction motor," in *Proc. 8th Int. Conf. Exhibit. Ecol. Vehicles Renew. Energies*, Mar. 2013, pp. 1–5.

[7] D.-M. Yang, "The application of artificial neural networks to the diagnosis of induction motor bearing condition using Hilbert-based bispectral analysis," in *Proc. 5th IEEE Ind. Electron. Appl. Conf.*, Jun. 2010, pp. 1730–1735.

[8] M. D. Prieto, G. Cirrincione, A. G. Espinosa, J. A. Ortega, and H. Henao, "Bearing fault detection by a novel condition-monitoring scheme based on statistical-time features and neural networks," *IEEE Trans. Ind. Electron.*, vol. 30, no. 8, pp. 3398–3407, Aug. 2013.

[9] P. Gardel, D. Morinigo-Sotelo, O. Duque-Perez, M. Perez-Alonso, and L. A. Garcia-Escudero, "Neural network broken bar detection using time domain and current spectrum data," in *Proc. 20th Int. Conf. Electr. Mach.*, Sep. 2012, pp. 2492–2497.

[10] S. S. Moosavi, A. Djerdir, Y. Aït-Amirat, and D. A. Khaburi, "Fault detection in 3-phase traction motor using artificial neural networks," in *Proc. IEEE Transp. Electrific. Conf. Expo*,Jun. 2012, pp. 1–6.

[11] H. Keskes, A. Braham, and Z. Lachiri, "Broken rotor bar diagnosis in induction machines through stationary wavelet packet transform and multiclass wavelet SVM," *Electr. Power Syst. Res.*, vol. 97, pp. 151–157, Apr. 2013.

[12] D. R. Sawitri, D. A. Asfani, M. H. Purnomo, I. K. E. Purnama, and M. Ashari, "Early detection of unbalance voltage in three phase induction motor based on SVM," in *Proc. 9th IEEE Int. Symp. Diagnostics Electr. Mach., Power Electron. Drives*, Aug. 2013, pp. 573–578.

[13] L. Shuang and Y. Fujin, "Fault pattern recognition of bearing based on principal components analysis and support vector machine," in *Proc. 2nd Int. Conf. Intell. Comput. Technol. Autom.*, vol. 2. Oct. 2009, pp. 533–536.

[14] B. Y. Lei, Z. He, Y. Zi, and X. Chen, "New clustering algorithm-based fault diagnosis using compensation distance evaluation technique," *Mech. Syst. Signal Process.*, vol. 22, no. 2, pp. 419–435, Feb. 2008.

[15] A. Soualhi, G. Clerc, and H. Razik, "Detection and diagnosis of faults in induction motor using an improved artificial ant clustering technique," *IEEE Trans. Ind. Electron.*, vol. 60, no. 9, pp. 4053–4062, Sep. 2013.

[16] E. Pacharawongsakda, C. Nattee, and T. Theeramunkong, "Improving Multi-label Classification Using Semi-Supervised Learning and Dimensionality Reduction," In *Trends in Artificial Intelligence, Lecture Notes in Computer Science*, vol. 7458. Springer, Sep. 2012, pp. 423–434, doi: 10.1007/978-3-642-32695-0_38.

[17] A. Soualhi, K. Medjaher, and N. Zerhouni, "Bearing health monitoring based on Hilbert–Huang transform, support vector machine, and regression," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 1, pp. 52–62, Jan. 2015.

[18] D. Ernst, P. Geurts, and L. Wehenkel, "Tree-based batch mode reinforcement learning," *J. Mach. Learn. Res.*, vol. 6, pp. 503–556, Apr. 2005.

[19] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms. A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[20] W. Hu, G. Liu, L. Fu, and H. Zhang, "Research of motor fault diagnosis based on PSO algorithm," in *Proc. 25th Chin. Control Decision Conf.*, May 2013, pp. 4600–4603.

[21] Z. Xu, J. Xuan, T. Shi, B. Wu, and Y. Hu, "Application of a modified fuzzy ARTMAP with feature-weight learning for the fault diagnosis of bearing," *Expert Syst. Appl.*, vol. 36, no. 6, pp. 9961–9968, Aug. 2009.

[22] P. V. J. Rodríguez, M. Negrea, and A. Arkkio, "A simplified scheme for induction motor condition monitoring," *Mech. Syst. Signal Process.*, vol. 22, no. 5, pp. 1216–1236, Jul. 2008.

[23] A. A. Silva, A. M. Bazzi, and S. Gupta, "Fault diagnosis in electric drives using machine learning approaches," in *Proc. IEEE Int. Electr. Mach. Drives Conf.*, May 2013, pp. 722–726.

[24] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artif. Intell. Rev.*, vol. 44, no. 4, pp. 467–508, Jun. 2015.

[25] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[26] I. Martin-Diaz, O. Duque-Perez, R. J. Romero-Troncoso, and D. Morinigo-Sotelo, "Supervised diagnosis of induction motor faults: A proposed methodology for an improved performance evaluation," in *Proc. IEEE SDEMPED*, Sep. 2015, pp. 359–365.

[27] J. P. A. Ioannidis, "Why most published research findings are false," *Public Library Sci. Med.*, vol. 2, no. 8, p. e124, 2005.

[28] C. Drummond, "Machine learning as an experimental science (revised)," in *Proc. AAAI' Workshop Eval. Methods Mach. Learn. I. Amer. Assoc. Artif. Intell.*, Menlo Park, CA, USA, pp. 5–8, 2006.

[29] J. Demsar, "On the appropriateness of statistical tests in machine learning," in *Proc. ICML 3rd Workshop Eval. Methods Mach. Learn., Assoc. Comput. Mach.*, New York, NY, USA: 2008, pp. 1–4.

[30] P. Domingos, "The role of Occam's razor in knowledge discovery," *Data Mining Knowl. Discovery*, vol. 3, no. 4, pp. 409–425, 1999, doi: 10.1023/A:1009868929893.

[31] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.

[32] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Mach. Learn.*, vol. 30, nos. 2–3, pp. 195–215, Feb. 1998.

[33] P. Ducange, B. Lazzerini, and F. Marcelloni, "Multi-objective genetic fuzzy classifiers for imbalanced and cost-sensitive datasets," *Soft Comput.*, vol. 14, no. 7, pp. 713–728, May 2010.

[34] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[35] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, 2006, pp. 233–240.

[36] H. Yu, "Resampling methods: Concepts, applications, and justification," *Practical Assessment, Res. Eval.*, vol. 8, no. 19, pp. 1–23, 2003.

[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd ed. Springer 2008. doi: 10.1007/978-0-387-84858-7.

[38] B. Efron and R. Tibshirani, "Improvements on cross-validation: The 632+ bootstrap method," *J. Amer. Statist. Assoc.*, vol. 92, no. 438, pp. 548–560, Jun. 1997.

[39] R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: http://www.R-project.org/

[40] Terry Therneau, Beth Atkinson and Brian Ripley (2015). *RPART: Recursive Partitioning and Regression Trees. R Package Version 4.1-10*. [Online]. Available: https://CRAN.R-project.org/package=rpart

[41] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F.Leisch (2015). *e1071: Misc Functions of the Departmentof Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package Version 1.6-7*. [Online]. Available: https://CRAN.R-project.org/package=e1071

[42] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: Visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005. [Online]. Available: http://rocr.bioinf.mpi-sb.mpg.de

IGNACIO MARTIN-DIAZ (S'16) received the B.S. degree in industrial engineering from the Escuela de IngenieríasIndustriales, University of Valladolid, Spain, in 2014. He is currently pursuing the Ph.D. degrees with the University of Valladolid, Spain, and the University of Guanajuato, Mexico. He is with the HSPdigital Research Group, Mexico. His current research interests are related to signal processing techniques, monitoring of induction machines, fault diagnosis and detection in electric machines, and machine learning algorithms.

DANIEL MORINIGO-SOTELO (M'04) received the B.S. and Ph.D. degrees in electrical engineering from the University of Valladolid (UVA), Spain, in 1999 and 2006, respectively. He was a Research Collaborator on Electromagnetic Processing of Materials, Light Alloys Division, CIDAUT Foundation, from 2000 to 2015. He is currently with the Research Group, Predictive Maintenance and Testing of Electrical Machines, Department of Electrical Engineering, UVA, and also with the HSPdigital Research Group, Mexico. His current research interests also include condition monitoring of induction machines, optimal electromagnetic design, and heuristic optimization.

OSCAR DUQUE-PEREZ received the B.S. and Ph.D. degrees in electrical engineering from the University of Valladolid (UVA), Spain, in 1992 and 2000, respectively. In 1994, he joined the Escuela Técnica Superior de Ingenieros Industriales, UVA, where he is currently a Full Professor with the Research Group, Predictive Maintenance and Testing of Electrical Machines, Department of Electrical Engineering. His main research fields are power systems reliability, condition monitoring, and heuristic optimization techniques.

RENE DE J. ROMERO-TRONCOSO (M'07–SM'12) received the Ph.D. degree in mechatronics from the Autonomous University of Queretaro, Queretaro, Mexico, in 2004. He is currently a National Researcher level 3 with the Mexican Council of Science and Technology, CONACYT. He is also a Head Professor with the University of Guanajuato and an Invited Researcher with the Autonomous University of Queretaro, Mexico. He has been an Advisor for over 200 theses, an Author of two books on digital systems (in Spanish), and a Co-Author of over 130 technical papers published in international journals and conferences. His fields of interest include hardware signal processing and mechatronics. He was a recipient of the 2004 Asociación Mexicana de Directivos de la InvestigaciónAplicada y el DesarrolloTecnológico Nacional Award on Innovation for his work in applied mechatronics and the 2005 IEEE ReConFig Award for his work in digital systems. He is part of the Editorial Board of Hindawi's *T*he Scientific World Journal and the textit International Journal of Manufacturing Engineering.