Technical Article

Artificial Intelligence Based Predictive Maintenance of Electromagnetic Devices

***Abstract* — In this contribution the authors describe a set of Artificial Intelligence (AI) based signal processing algorithms that are used to establish Predictive Maintenance procedures. A set of applications to real cases in the Railway Systems environments are shown.**

I. Introduction

Predictive Maintenance and Reliability Centered Maintenance approaches are becoming fundamental in basically all technical environments, to reduce costs of operation and increase reliability and safety. This is achieved by optimizing the maintenance process, defining new strategies and algorithms to locate faults, monitoring health conditions of subsystems and estimating residual life of components. Since electric power is driving most of industrial processes and, nowadays, strictly related to mobility, real time monitoring of electromechanical devices and the development of the above mentioned algorithms is central to the main stakeholders (mobility operators, industry, device makers etc.).

The optimal condition to achieve such goals would be to use the lowest number of additional equipment (i.e. sensors and data transmission infrastructure) to record/observe the physical quantities needed to assess the status of the components: in this way the economical impact of the new maintenance procedure would be higher and the number of new devices (characterized by their own failures and maintenance cost) would be low.

The authors have developed a set of algorithms based on different Artificial Intelligence paradigms in order to approach specific needs regarding electrical/electromechanical devices in railway systems. Such algorithms have been developed in the framework of projects financed by the Italian government, having the national railway company as main partner; for this reason the test cases are based on real field measurements, taken during the regular operation of high speed/commuter trains. More specifically, the requested area of interest were the condition monitoring of the induction motors/gearbox block equipping a great number of commuters trains and the condition monitoring of the pantograph-catenary subsystem of high speed trains.

There are many non-invasive techniques to monitor the status of induction motors, and usually they rely on easily measured electrical or mechanical quantities, such as voltage, current, external magnetic field, speed, and vibrations [1]. In particular, the literature on early fault detection in induction motors is mainly focused on the use of stator current measurements, [2] – [4] and vibrations measurements [5], [6]. It is worth to mention that the motors of the regional trains object of the study are not equipped with current and vibrations sensors, that are commonly used in high-speed trains. In contrast, such locomotives are equipped with temperature sensors, that can record the signals during the train operation. Real time temperatures have more rarely been investigated in traction applications, whereas there is a literature dealing with infrared thermography for condition monitoring of induction motors [7], [8].

In [9] – [12] it is underlined that a temperature increase (thermal stress) can have negative effects both on the bearings and on the motor. In particular, a high temperature can deteriorate the bearing lubrication causing an abnormal friction that may eventually lead to a bearing damage and in turn accelerate the ageing process of the winding insulation leading to a winding fault inside the induction motor. In addition, as remarked in [13] and [14] with specific reference to the condition monitoring of bearings, temperature rises can be attributed to several reasons, namely, winding temperature rise, motor operating speed, temperature distribution within the motor, lubricant viscosity, and the amount of lubricant. Accordingly, a rise in the temperature can be caused by different factors and, at the same time, can be used as an indicator of a number of severe faults. In [9] it is said that about 30% of induction motor faults are stator winding fault, and a small percentage of them is directly caused by an initial fault in the insulation. In case the insulation has a defect, then partial discharges are created, which cause overheating that can be detected by looking at the temperatures before a short circuit occurs; on the contrary, if a short circuit suddenly occurs without a previous temperature increase, it means that the insulation presented major issues directly from fabrication, and at this stage nothing can be done in terms of predictive maintenance.

Railway systems based on pantograph catenary collection technique are a vital part of the transportation system of each country in the world. In many countries (Europe and Japan, for instance) the development of high speed trains is the main driving force in the railway industry, but in other countries also the impressive increase of “regular” railway connections has to be considered. In both cases the problem of guaranteeing a high quality (safe and reliable) power collection is fundamental and it is still a hot research topic worldwide.

For this reason the pantograph – catenary system is constantly monitored to evaluate the catenary and contact strip actual status; monitoring is in general performed by visual inspection and periodic contact strip replacement is generally operated by the railway companies with the aim of preventing a fault condition which could lead to serious traffic schedule disruption and consequent economical damage.

There is a vast literature relative to the pantograph – catenary subsystem; in particular in [15] – [19] the mechanical behaviour, in term of vibrations, is monitored by a set of sensors (typically load cells, accelerometers or brag/fiber sensors) located on the track or in the train.

A common approach is to equip the locomotive with either a phototube or a photodiode capable of detecting electric arcing through the ultraviolet emission, [20] – [22], even though it might be not economically convenient to equip all trains with this setup. In [23] the Fourier Transform (FT) is used for the analysis of the current but extracting the proper information from the frequency spectrum is not always an easy task and when wide time windows are used the information regarding the location of the event is lost. New contributions are relative to the use of the sound produced by the electric arc [24] and to the use of optimization algorithms [25] or convolutional neural networks [26].

A noteworthy contribution to the research has been given in the past by the authors for dc railway systems: by the use of the wavelet expansion electric arcs can be detected and located by simply analysing the collected current [27]. This was clearly a great improvement on previous techniques since it avoids the need of a photosensitive device. Now the extensive development of data mining techniques based on AI has opened the possibility of new and more efficient algorithms.

This contribution describes a set of algorithms and shows their application in the above mentioned environments. The results demonstrate that AI based algorithm can be crucial for the creation of predictive maintenance procedures.

II. Algorithms' Descriptions

In this section a basic description of the foundations of different algorithms, developed by the research group, are described. Three of them can be included in the AI area (Neural Networks, Support Vector Machines and Clustering), while the Hotelling Multivariate Control Chart is a statistics based analysis.

One of the main challenges of predictive anomaly detection is that the class distribution of the data is in general unbalanced: observations of the abnormal behaviors are scarce while most of the observations represent nominal behaviours. Under unbalanced class distribution, like in many real scenarios, most of the classification methods perform poorly. A common solution consists in creating a model of the nominal behaviour, and monitoring the deviations from the nominal conditions.

The main characteristics of the proposed algorithms is that they are capable of classifying events when they can be trained on a statistically significant data set, and can be useful for such analysis since no physical knowledge of the phenomenon is requested, which is important in an industrial environment where the only important aspect is the efficiency of the preventive maintenance.

Only a short paragraph is dedicated to Neural Networks since it is the most known paradigm amongst the ones treated in this contribution.

A. Feedforward Neural Networks

Feedforward neural networks (NNs) are a class of universal approximators, as they can approximate arbitrarily well functions from $\mathbb{R}^N$ to $\mathbb{R}$, with a finite number of neurons in a single layer [19] and they are nowadays probably the most common supervised method. NNs are widely used for detection of incipient faults and predictive diagnostics in induction motors, for instance, they are used to detect and classify the faults using vibration signals [5], [6], and stator current signals [4]. In this specific application, we consider one hidden layer networks where a stochastic gradient descent optimization algorithm is used for backpropagation, and where the objective function is the minimization of the mean-square error (MSE). The output of the NN is described by the following expression:

$$F_{NN}(\boldsymbol{x}) = \sum_{i=1}^{H} v_i \varphi(\boldsymbol{w}_i^T \cdot \boldsymbol{x} + b_i) \qquad (1)$$

where $\boldsymbol{x} \in \mathbb{R}^p$ is the input column vector, $\boldsymbol{w_i} \in \mathbb{R}^p$, $b_i \in \mathbb{R}$ and $v_i \in \mathbb{R}$ represent the weight vector, the bias and the output weight of neuron i, respectively, and $\varphi(\cdot)$ is the $\mathbb{R} \longrightarrow \mathbb{R}$ activation function (a non-constant, bounded, and monotonically increasing continuous function). Finally, H is the number of neurons in the hidden layer. A typical problem of NNs, especially when large sets of data are available like in this case, is the so-called overfitting, where roughly speaking the NN learns the data, and not the structure underlying the data, leading to wrong predictions. To avoid this circumstance and to improve the ability of the NN to generalize, we used a validation set as a subset of the training data (not used for training), which stops the training after the MSE in the validation set does not improve for a fixed number of consecutive epochs (max_fail). A peculiar use of four such NNs, as described in the test case, let us obtain good results in terms of malfunction operations.

B. Hotelling Multivariate Control Chart

The Hotelling control chart [29], [30] can be considered as a semi-supervised method, and performs a dimensionality reduction of the multivariate data to a scalar parameter denoted as t^2 statistics, which represents the square of the Mahalanobis distance [31] of the observation vector from the vector containing the mean values of the variables in nominal conditions. As reported in many studies [32] – [34], the t^2 statistics is able to capture the changes in multivariate data, revealing the deviations from the nominal behaviour. In principle, this allows one to select safety thresholds (UCL, upper control limit; LCL, lower control limit) on the t^2 control chart more efficiently than upon the original data. Although such a tool had been originally proposed already in 1947 as a tool for quality control, the Hotelling control chart is still being applied in many process control applications, and can be regarded as a precursor of one-class classification methods (i.e. methods that only model a single class of the data [35]). For these reasons, the Hotelling control chart is widely used for early detection of incipient faults, as an example [2], [3] proposed the use of the Hotelling control chart for incipient fault detection in induction motors, by monitoring the stator current.

The construction of the control chart includes two phases: in the first phase, historical data are analysed and the safety thresholds are computed; phase two corresponds to the monitoring of the real-time process. In phase one, a faultless historic dataset of the process should be defined by experts with the confidence that it represents mainly nominal behaviours of the process. The historic dataset is used to create a statistic of the nominal behaviour, consisting of a nominal mean vector and a covariance matrix. Let the historic dataset of phase one be represented by the matrix $\boldsymbol{X} \in \mathbb{R}^{Nxp}$ containing N observations of nominal states of the process, that consists in row vectors of p variables. We denote the row vectors of $\boldsymbol{X}$ as $\boldsymbol{x}_i \in \mathbb{R}^{1xp}$ where $i = 1, \dots, N$. The sample mean vector $\boldsymbol{\mu} \in \mathbb{R}^{1xp}$ of the data is defined as

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \qquad (2)$$

In order to define the covariance matrix, we need to construct the demeaned (zero-mean) data matrix $\boldsymbol{X_0} \in \mathbb{R}^{Nxp}$

$$\boldsymbol{X_0} = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu} \\ \boldsymbol{x}_2 - \boldsymbol{\mu} \\ \boldsymbol{x}_n - \boldsymbol{\mu} \end{bmatrix} \qquad (3)$$

Then, the covariance matrix $\boldsymbol{C} \in \mathbb{R}^{pxp}$ of the data is defined as

$$\boldsymbol{C} = \frac{1}{N-1}\boldsymbol{X}_0^T \boldsymbol{X}_0 \qquad (4)$$

These parameters represent the nominal behavior of the process, and we assume that $\boldsymbol{C}$ is full rank The scalar t^2 statistics is defined as

$$t^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^T \qquad (5)$$

The t^2 statistics is small when the observation vector **x** represents nominal states, while it increases when the observation vector **x** deviates from the nominal behaviour. In order to define the safety thresholds UCL and LCL of the control chart, in phase one we calculate the mean value μ and standard deviation σ of the t^2 values obtained with the nominal observations $\boldsymbol{x}_i$, for $i = 1, \dots, N$ i.e.:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} t^2(\boldsymbol{x}_i), \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[t^2(\boldsymbol{x}_i) - \mu]^2} \qquad (6)$$

Then we define the safety thresholds as

$$\begin{cases} UCL = \mu + 3\sigma \\ LCL = \max(\mu - 3\sigma, 0) \end{cases} \qquad (7)$$

where the actual concern regards only the upper limit, as in faulty conditions the temperatures exceed the upper limit.

In phase two, we use process statistics and the control limits extracted during phase one in order to detect an anomalous behaviour on new data to be monitored. In particular, during phase two new observation vectors are measured, and the corresponding t^2 values are calculated as in (5). The Hotelling control chart consists in a monitoring tool that plots the t^2 values as consecutive points in time and compares them against the control levels. The process is considered 'out of control', and an anomalous behaviour is detected, when the t^2 values continuously exceed the control limits.

C. Support Vector Machines

The SVM is a supervised classification technique and its aim is to to find the maximum margin hyperplane, where the margin is defined as the distance between the separation hyperplane and the nearest training samples, which are called support vectors as shown in Figure 5.

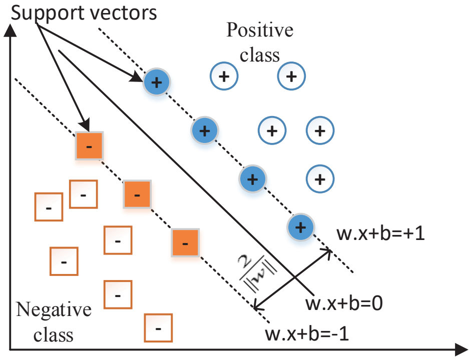

Fig. 1. Support vector classifier for linearly separable data

Classification algorithms of time series in general do not use the raw data signals in the time domain but require a reduced dimensional input vector that represents each time series. To obtain an input dataset, a feature extraction has been performed on the time domain signals, which is based on the calculation of the periodogram.

Given a generic time series $x(k\Delta t)$, where the signal is composed by m time samples spaced by a constant sample time Δt, we can consider, without loss of generality, that the sampling rate is taken as $\Delta t = 1$, and the value of the generic time serie at time k is written as $x(k)$ (at times $k = 1, \dots, m$). The periodogram $X(q)$ at frequencies $q = 1, \dots, m$ is defined as

$$X(q) = \frac{1}{m}\left|\sum_{k=1}^{m} x(k)e^{-j2\pi qk\frac{1}{m}}\right|^2 \qquad (8)$$

Periodogram analysis gives good results in supervised and unsupervised classification of time series data [22], and it is also very easy to compute using an FFT algorithm. In general a reduced number of samples of the periodogram gives a good representation of the time series, in particular a metric based on truncated periodogram in logarithmic scale is proposed in [36]. When the time position k_i of a number of fault events $i = 1, \dots, N$ (in our case detected by using the photosensor signal) is known, for each occurrence we can compute the periodogram of the $2m$ samples time window $[k_i - m, k_i + m - 1]$ of the voltage or current signal. These frequency domain signals are used to identify the arc events. A number of N periodogram signals that are not correlated with arc events are also computed to represent the signal in the case that the arc is not present. The total number of input data signals is then $2N$.

The logarithmic periodogram signals $\log_{10}(X(q))$ are truncated to their first p components, to form the input vectors $\boldsymbol{x_i} \in \mathbb{R}^p$, $i = i, \dots, 2N$, while the target vector components $\boldsymbol{y} \in \mathbb{R}^{2N}$ indicates if the input $\boldsymbol{x_i}$ is related to an arc event or not, correspondingly $y_i = \{-1,1\}$.

The formulation of the SVM classifier used in this work is the soft margin nonlinear classification [37], [38], which is based on the following optimization problem (primal problem):

$$\begin{gathered} \min_{w,b,\xi} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{2N}\xi_i \\ \text{subject to } y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0, i = 1, \dots, 2N \end{gathered} \qquad (9)$$

where the nonlinear transform $\phi(\boldsymbol{x}_i)$ maps $\boldsymbol{x}_i$ into a higher dimensional feature space, and $C \geq 0$ is the regularization parameter. The vector $\boldsymbol{w}$ denotes the normal vector to the optimal separation hyperplane in the transformed space, whereas slack variables ξ_i measure the degree of misclassification of the vectors $\boldsymbol{x}_i$. The decision function is then given by

$$\operatorname{sgn}(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \qquad (10)$$

The solution of the constrained problem in (9) is obtained by the method of Lagrange multipliers, in particular by solving the dual problem which is in the form:

$$\begin{gathered} \min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Q}\boldsymbol{\alpha} - \sum_{i=1}^{2N}\alpha_i \\ \text{subject to } \sum_{i=1}^{2N} y_i\alpha_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, \dots, 2N \end{gathered} \qquad (11)$$

where $\alpha \in \mathbb{R}^{2N}$ is the vector of the Lagrange multipliers, and $\boldsymbol{Q}$ is a positive semidefinite $2N$ by $2N$ matrix $Q_{i,j} = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where k is the kernel function that represents the dot product in the transformed high dimensional space.

The advantages of using the dual form are that the calculation of the transform $\phi(\boldsymbol{x}_i)$ is not needed, as only the calculation of kernel function is required (kernel trick), and that the slack variables ξ_i vanish, with the constant C appearing only as an additional constraint on the Lagrange multipliers.

After problem (11) is solved, only a few α_i will be greater than zero, and the corresponding $\boldsymbol{x}_i$ are the support vectors. Using the primal - dual relationships the optimal $\boldsymbol{w}$ is obtained as:

$$\boldsymbol{w} = \sum_{i=1}^{2N} y_i \alpha_i \phi(\boldsymbol{x_i}) \tag{12}$$

The bias b is calculated as

$$b = \frac{1}{N_{SV}} \sum_{i \in SV} \boldsymbol{w}^T \phi(\boldsymbol{x}_i) - y_i \tag{13}$$

where $SV = \{i: \alpha_i > 0\}$ is the set of the support vectors indices, and N_{SV} is the number of support vectors.

The dot product $\boldsymbol{w}^T \phi(\boldsymbol{x}_j)$ needed for calculating the bias and also decision function (3) is calculated using the kernel trick

$$\begin{aligned} \boldsymbol{w}^T \phi(\boldsymbol{x}_j) &= \sum_{i=1}^{2N} y_i \alpha_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \\ &= \sum_{i=1}^{2N} y_i \alpha_i k(\boldsymbol{x_i}, \boldsymbol{x}_j) \end{aligned} \tag{14}$$

In this way the direct calculation of $\phi(\boldsymbol{x}_i)$ is never required and the classification function (10) becomes

$$\operatorname{sgn}\left(\sum_{i=1}^{2N} y_i \alpha_i k(\boldsymbol{x_i}, \boldsymbol{x}_j) + b\right) \tag{15}$$

The solution of the problem (11) is performed by means of Sequential Minimal Optimization (SMO) [39].

The SVM method presented requires the selection of the kernel, and the parameter C . In this work we use the Gaussian kernel, or radial basis function RBF,

$$k_{RBF}(\boldsymbol{x_i}, \boldsymbol{x}_j) = \exp\left(-\gamma \left\|\boldsymbol{x_i} - \boldsymbol{x}_j\right\|^2\right) \tag{16}$$

which requires the choice of the parameter γ.

D. Clustering Techniques

An increasing interest has recently focused on the clustering and classification of time series, as it reveals to be a significant research area in several fields, such as engineering, physics, economics, finance, medicine, biology, and many others. In general, the analysis of time series requires the use of high dimensionality spaces, and the direct application of existing algorithms for clustering static data leads to reduced performance. In fact clustering algorithms, generally, use the Euclidean distance between the data vectors, which works relatively well in the classification of short time series, with a length of few tens of time samples. However, in this work we consider long time series, where the length is of the order of hundreds or thousands. To overcome the relatively poor performance obtained with Euclidean distance, clustering algorithms for time series in general adopt one of the following two strategies. The first approach modifies the existing algorithms for static data, replacing the distance measure with an appropriate one for time series; the second approach tries to convert time series data into a set of feature vectors of lower dimension, and then use existing algorithms.

A distance function based on the periodogram of the time series is proposed in [36], which compares the proposed metric with many other alternatives, showing the potential of the use of frequency representation for the classification of time series.

In this work, we propose a method that incorporates the periodogram metric in the k-means clustering method [40]. In particular, we consider a truncated periodogram (defined before) so the proposed approach is a combination of the two strategies described above.

Given a generic time series $x_i(k)$, at times $k = 1, \dots, m$, the periodogram $X_i(q)$ at frequencies $q = 1, \dots, m$ is defined as in section IIC. Being the periodogram an estimation of the power spectral density of the time series, it makes sense to use the logarithm of the periodogram $\tilde{X}_i(q) = \log_{10}(X_i(q))$

The metric proposed in [36] is based on the logarithmic periodogram:

$$d_{LP}(x_i, x_j) = \sqrt{\sum_{q=1}^{\lfloor m/2 \rfloor} \left[\tilde{X}_i(q) - \tilde{X}_j(q)\right]^2} \tag{17}$$

where $\lfloor m/2 \rfloor$ is the largest integer less or equal to $\lfloor m/2 \rfloor$.

In this work we consider a truncation of the logarithmic periodogram $\tilde{X}_i(q)$ to its first d values, $q = 1, \dots, d$, where $d \ll m$. In this way, each time series $x_i(k)$ is converted to a d-dimensional vector $\boldsymbol{y}_i = [\tilde{X}_i(1), \tilde{X}_i(2), \dots \tilde{X}_i(d)]$.

To cluster the converted dataset $\boldsymbol{y}_i = 1, \dots, n$ we use the k-means clustering [40], which aims to partition the n data vectors $\boldsymbol{y}_i$ into c clusters sets, $\{S_1, S_2, \dots, S_c\}$, in order to minimize the within-cluster sum of squares:

$$\min_{S_i} \sum_{i=1}^{c} \sum_{y_i \in S_i} \|\boldsymbol{y}_i - \boldsymbol{v}_i\|^2 \tag{18}$$

where $\boldsymbol{v}_i$ is the mean of the points in the cluster S_i. The k-means algorithm defines a heuristic strategy that uses an iterative refinement method to reach the goal in (18). After defining a set of initial means $\boldsymbol{v}_i$, $i = 1, \dots, c$, the iterative technique proceeds by alternating between the following two steps:

- Expectation step: Assign each point vector $\boldsymbol{y}_j$, $j = 1, \dots, n$ to the cluster S_j with the nearest mean vector $\boldsymbol{v}_i$
$$S_i = \left\{\boldsymbol{y}_j: \left\|\boldsymbol{y}_j - \boldsymbol{v}_i\right\|^2 \le \left\|\boldsymbol{y}_j - \boldsymbol{v}_k\right\|^2, k = 1 \dots c\right\} \tag{19}$$
- Maximization step: Calculate the new mean vectors by the centroids of the points in each cluster
$$\boldsymbol{v}_i = \frac{1}{|S|} \sum_{y_j \in S_i} \boldsymbol{y}_j \tag{20}$$

When the assignments no longer change, the algorithm has converged to a local minimum of (18). A commonly used initialization method for defining initial means is the random partition, which first randomly assigns a cluster to each point in the expectation step, and proceeds to the maximization step. In the following the random partition is used for initial means. In general, there is no guarantee that the algorithm will converge to the global optimum, and the result may depend on the initial clusters. As the algorithm is usually very fast, it is common to run it multiple times with different starting conditions. In our experiments, we performed 200 replicates with random initial clusters, and the results of the run with the smaller within-cluster sum of squares (18) is selected.

For selecting the number of clusters c we use the internal validity measure defined by the Dunn index. Given a partition of the data in c sets $\{S_1, S_2, \dots, S_c\}$ the Dunn index is defined as the ratio between the minimum intercluster distance $\delta(S_i, S_j)$ and the maximum cluster size Δ_i. Various definition exist for the intercluster distance $\delta(S_i, S_j)$ and the cluster size Δ_i, and we use the following ones:

$$\Delta_i = \frac{1}{|S_i|} \sum_{y_j \in S_i} \|\boldsymbol{y}_j - \boldsymbol{v}_i\| \tag{21}$$

which is the mean value of the distance of all the points from the centroid, whereas

$$\delta(S_i, S_j) = \|\boldsymbol{v}_j - \boldsymbol{v}_j\| \tag{22}$$

which represents the distance between the centroids of the clusters i and j. The Dunn index DI_c is then defined as

$$DI_c = \frac{\min\limits_{k \neq j} \delta(S_k, S_j)}{\max\limits_{i=1\cdots c} \Delta_i} \tag{23}$$

A good choice of the number of clusters c is given by the partition that has the higher value of the Dunn index DI_c, varying c between 2 and a maximum value defined by the user.

III. Application to Railway Systems Induction Motors

A. Available data description

The induction motors object of the study equip one of the most common locomotive running throughout Italy, and operating on regular commuter service. Each locomotive is equipped with four induction motors, all having (i) squirrel cage rotors; (ii) double start stators; (iii) 4 poles; (iv) a maximum power of 895 kW, and a (v) rated voltage of 1090 V.

In addition, the locomotive is equipped with two inverters, where each one of them powers one of the two stator windings of the four motors, connected all-together in parallel. During the normal operation of the train, sensors are used to regularly monitor the temperature of the motor windings, the temperature of the gearbox lubricant (both acquired with standard PT100 Resistance Temperature Detectors), and the train speed (acquired through a pulse generator). The pulse generator has a maximum operating frequency of 30 kHz, and the temperatures are recorded with a sampling time of 3 minutes. More in detail, the temperature probes are located in the stator laminated core, so that the sensitive element can reach one of the stator tooth and could reach a temperature very close to the winding temperature; as for the probes in the gearbox they are directly immersed in oil

For redundancy reasons, two temperature sensors are used on each motor and gearbox, thus leading to a total of 16 temperatures that are measured for each locomotive every 3 minutes.

Some examples of recorded data are provided in Figures 2 and 3. In particular, Figure 2 shows a sequence of 1500 samples for a regular operation of a train, with temperatures recorded for the four engines. On the other hand, Figure 3 show a sequence of 1500 samples that ends with a fault, as can be seen more clearly towards the end of Figure 4, where the temperature of one motor becomes ultimately too large.

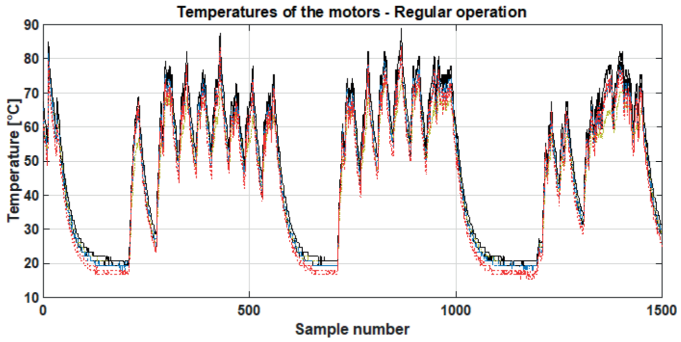

Fig. 2. Example of the temperatures recorded for the motor in a normal operation

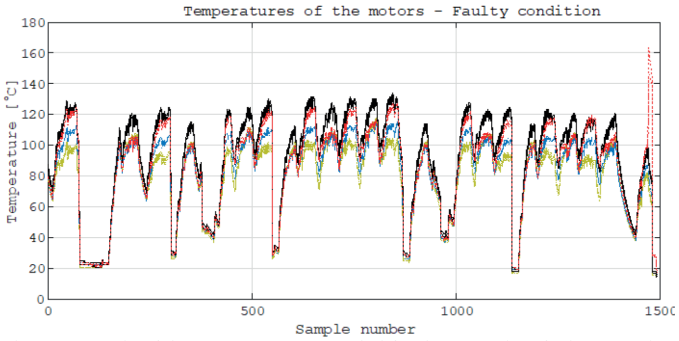

Fig. 3. Example of the temperatures recorded for the motor in a faulty operation

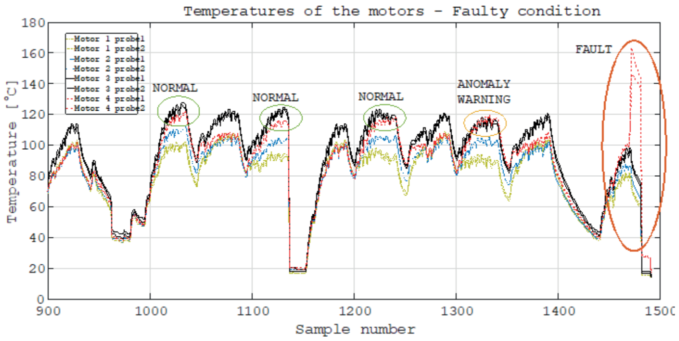

Fig. 4. Visual inspection of the temperatures recorded for the motor when a fault occurs

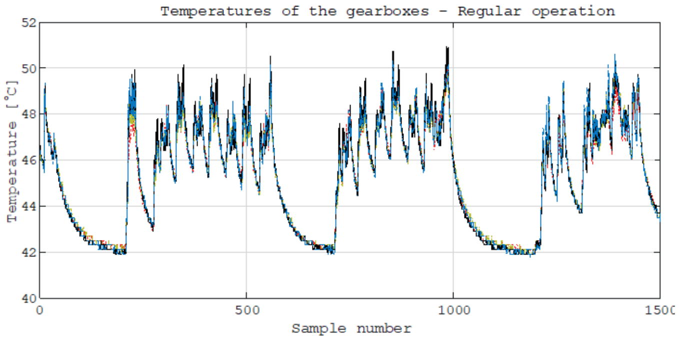

Fig. 5. Example of the temperatures recorded for the gearbox in a normal operation

Figure 5 shows the temperatures measured in the gearbox.

A number of useful remarks can be made by observing the data:

- the temperatures have some periodic patterns. This is due to the fact that temperatures increase during a trip of the train and decrease again when the train stops before the next trip. Also, sometimes trains have long pauses (e.g., some trains do not travel at night time), and consequently all temperatures converge to a value close to the environmental temperature;
- in general, it is possible to observe that the temperatures of the motors are greater than those of the gearboxes;

- in some cases all sensors provide very similar values of the temperature, while in other circumstances they read more different values. While this fact is not correlated with the chance of a fault, still the very different variance of the vector of read temperatures from train to train complicates the prediction of a fault in practice;
- in some cases some sensors provide evidently wrong reads (i.e., negative temperatures). This could occur for single values (i.e., where more likely an error occurred in the data transmission process) or for longer sequences of values (i.e., more likely a fault in a single sensors). In our analysis we have cleaned the data by neglecting (single) values (of single sensors) that are out of the historical range of temperatures. In the case of long sequences of values out of range, the conclusion that the sensor itself is malfunctioning could be taken.

B. Application of the Hotelling Multivariate Control Chart

We now briefly discuss how we implemented it in our specific application. The observation vector is represented by the vector of the 16 temperatures measured simultaneously. In particular, we have used the initial 600 samples (i.e., 30 hours) of normal operation of the train to calculate the nominal values of the t^2, i.e., in terms of its average value μ and its standard deviation σ as in (6). These values, as well as the control limits defined in (7), are characteristic of a specific train. Note that the specific values depend on the train, on its typical route, and also on some specific installation parameters (e.g., motors and sensors). In our experience, we have found out that it is very important to continuously update the values of μ and σ to take into account physiological variations of the nominal parameters, for instance due to different environmental temperatures. At the same time, the parameters cannot be updated too frequently to avoid including possible incipient faulty conditions into the computation of the safety thresholds. Accordingly, we shift the window of 600 samples every 50 new samples (i.e., every 2.5 hours), when the new safety thresholds of (7) are duly recomputed.

C. Application of the Feed Forward Neural Network

Referring to the general FFNN the following choices have been performed, based on the authors' experience:

- we have chosen a sigmoidal function as an activation function, we have determined the optimal number of neurons by using a three-fold cross-validation (which is defined as the ratio between the number of correctly identified points and the total number of points in test data, and averaging over 3-folds);
- we set the validation set as a randomly selected set of 20% the size of the training set;
- we set the value of max_fail equal to 15, and similarly to the Hotelling solution, we used a training set of 600 samples, representing nominal operation.

As further data preprocessing it is worth to mention that the Matlab implementation of the FFNN automatically applies a mapping of input and output data to the range [–1, 1].

In our specific application, we decided to use four NNs in parallel, where each one of them had the temperatures of three motors as an input, $p = 12$, and the average temperature of the four sensors of the remaining motor as an output. The rationale of this choice is that under the assumption that the temperature signals remain more or less the same (as a whole), then by knowing the temperatures of three motors, one may learn how to predict the temperatures of the fourth one. However, such a pattern breaks when one fault occurs. In particular, one motor starts heating, and the other three motors lose their ability to predict its temperature.

As a result of the cross-validation analysis, we obtained that the optimal number of neurons in the hidden layer H is between 3 and 5, depending on the particular run and particular set of input– output signals. In the monitoring stage, new observations are measured and given as input to the four trained NNs. The output of each NN is compared with the average of the measured temperatures of the motor to be predicted, and an absolute error, AE, is calculated and compared to an upper threshold. The threshold in this case is defined as three standard deviations of the AE calculated in the training phase.

D. Results

With both methodologies, and as typical with most fault predictions approaches, it is important to decide whether we are interested in receiving many alarms (which might include false alarms as well), or whether we wish to receive an alarm only when the algorithm is pretty confident that a fault has actually occurred. In this specific case, we are interested in being very conservative when giving an alarm, as a false alarm is also very expensive from the point of view of the train company (in fact, the train company might want to stop the train while running, with a number of inconveniences for the passengers). Thus, we are specifically interested in minimizing the chance of having false alarms. However, the counterpart of being conservative is that a fault may be recognised with some delay (i.e., some extra time is required to make sure that a failure has actually occurred and it is not a false alarm).

Figure 6 shows an example of the Hotelling statistics in case of a fault. It is evident that the t^2 statistics frequently exceeds the UCL threshold (horizontal dashed line). To avoid false positives (i.e., fault alarms when not required), we decided to set the alarm when a sequence of 20 consecutive samples is found to continuously exceed the Threshold

Here a faulty occurs towards the end (the vertical thick black line indicates that a too large temperature has been achieved). The vertical red line corresponds to the instant of time when Hotelling predicts the fault and recommends to stop the train.

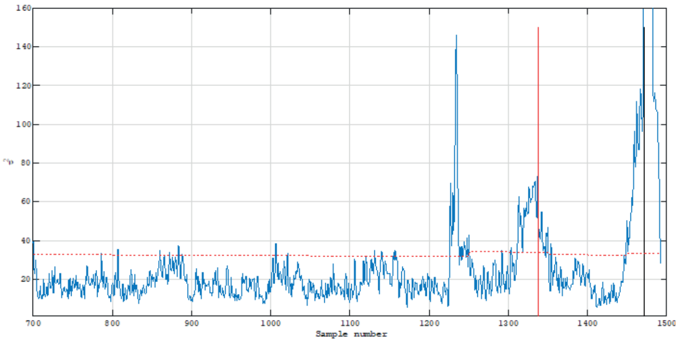

Fig. 6. Hotelling analysis result of a faulty test case.

On the other hand, NNs do not seem to ever provide false positives, and in general they do not require special care in tuning particular parameters. This is a great advantage, as some tuning procedures (e.g., the previous choice of 20 consecutive samples out-of-bounds before Hotelling recognises a faulty condition in practice) may be regarded as empirical. Fig, 7 shows an example: the vertical thick black line indicates that a too large temperature has been reached (fault), while the vertical red line corresponds to the instant of time when the NN predicts the fault.

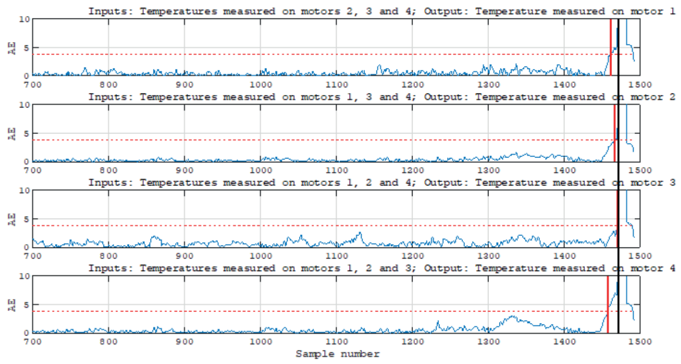

Fig. 7. NN analysis of a faulty case.

Table I show the performances of the algorithms for 10 different trains, showing that the result of FFNN are more stable in terms of time from early prediction to fault occurrence (ranging from 33 to 42 minutes) with respect to the early prediction times for Hotelling (from 6 to 402 minutes). So, while Hotelling is in general less reliable, it was able to predict the fault almost 7 hours in advance in one case.

TABLE I. PERFORMANCE OF THE TWO ALGORITHMS IN THE EARLY DETECTION OF THE FAULT.

Performance/Method	Hotelling	Neural Networks
False positives	0	0
False negatives	0	0
Shortest early prediction of a fault (minutes)	6	33
Largest early prediction of a fault (minutes)	402	42
Ability to assess the exact motor where the fault has occurred	no	yes

IV. APPLICATION TO PANTOGRAPH CATENARY SUBSYSTEM

A. Available data description

The data available for the analysis are relative to 6 test runs of a 25kV a.c. high speed trains, operated on regular passengers railway tracks. The trains are equipped with voltage and current recording instruments (which are always present on high speed trains), and two phototubes revealing the presence of the electric arcs. The data are sampled either at 5kHz or at 20kHz and for each test run the data available are:

- Voltage
- Current
- Train velocity
- Phototube output

In each test run the above described quantities have been recorded for approximately 25 minutes. Figure 8 shows the typical velocity profile of a test run and the envelope of the current collected by the pantograph: it is evident that the train is accelerated at the beginning of the run and decelerated at the end of the run while the velocity is kept approximately constant for about 15 minutes. During this time different values of the current are used to simulate different operating conditions.

Figure 9 shows a portion of the phototube signal in presence of arcs, and the corresponding recorded current: a simple visual analysis of the current in Figure 9b would not lead to any conclusion; in addition, as recalled in the introduction more complex time-frequency techniques do not give extremely satisfactory results.

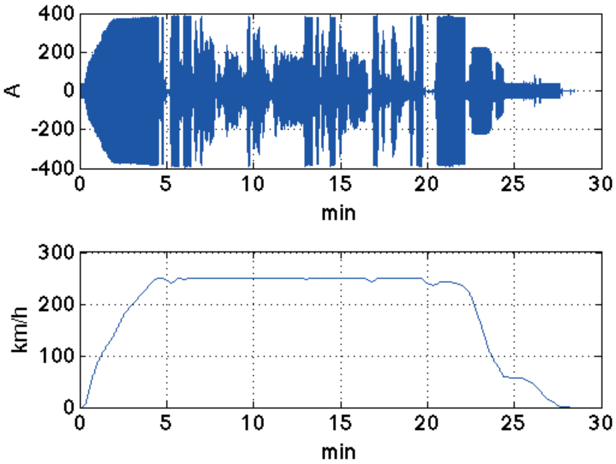

Figure 8: Velocity and current profile of a test run.

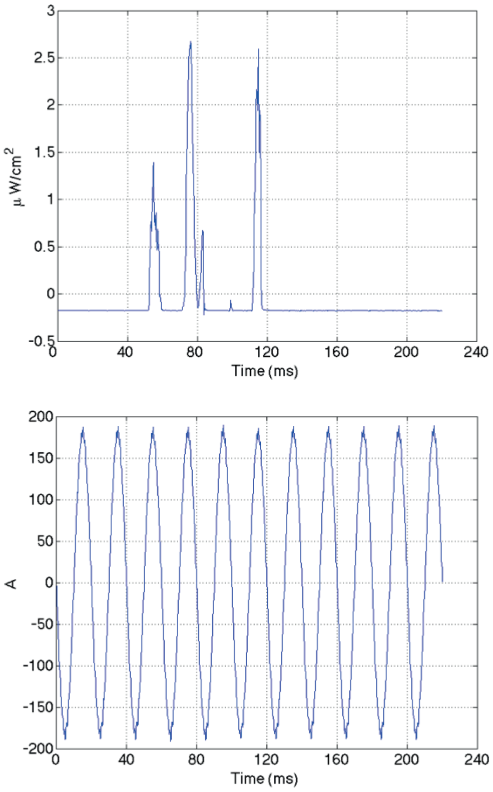

Figure 9: Phototube output (a) and corresponding collected current (b)

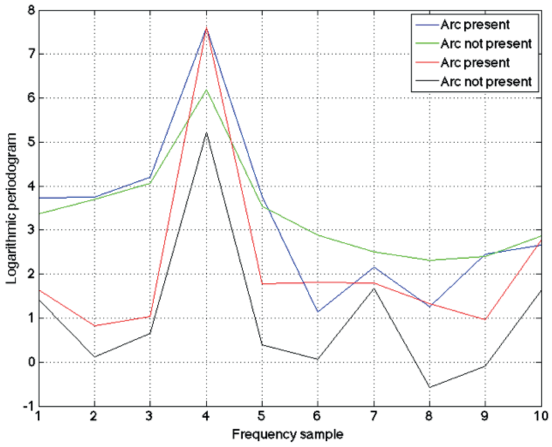

Figure 10: Logarithmic periodogram (frequency sampling f0 = 12.5 Hz)

Figure 10 shows the logarithmic periodogram of different current signals characterized by presence and absence of arcs: it is evident by analysing this figure that a direct analysis of frequency domain data can be extremely difficult.

It is noteworthy to mention that according to the regulation EN 50317 (Railway applications - Current collection systems - Requirements for and validation of measurements of the dynamic interaction between pantograph and overhead contact line), "for the output, only arcs longer than a defined duration shall be analyzed. This duration depends of the problem which has to be investigated. A common value is 5 ms, when investigating current collection quality". Arcs longer than 5 ms can be related both to catenary and to pantograph defects/wear, while shorter arcs might anyway indicate contact strip wear, for this reason all the arcs have been considered in this work since they anyway indicate the actual status of the collection system. Consequently, a variable threshold is used to identify the arcs and the threshold is obtained by means of biasing a moving average of the photosensor signal. Arcs (and its time coordinate) are identified when the signal of one of the two photosensors exceeds the threshold.

B. Application of SVM

In the first case study, the photosensor currents and voltage signals are sampled at *5kHz* and a total number of around $7 \cdot 10^6$ points are recorded for each signal; As a first step the photosensors signals are analyzed in order to identify the arc events using the variable threshold method described above, where the moving average window is set to 6000 points, and the bias is selected in order to avoid the background noise. As a result, a total number of *N=22039* arcs are identified using both phototubes, hence only the 0.3% of the recorded signal is affected by the presence of electric arcs. For each identified arc $i = 1 \dots N$ we save its time position index k_i; in addition, a number *N* of time positions k_i for $i = N + 1 \dots 2N$ where the arc is not present are randomly selected from the signal. Therefore, the SVM for classification will be trained using *2N=44078* points, with two balanced classes.

The scope of the classification is to detect the arc presence from the currents and voltage signals: once the SVM is trained with known data, it is capable to assign an input time series (voltage and/or current) to one of the two classes (arc present or not present) without the need of the photosensor.

Due to the big amount of data available from direct measurements (one voltage signal V_1 and two currents signals I_1, I_2), we need to determine which signals are more correlated to the arc occurrences. Therefore as a preliminary analysis we train a SVM using different combinations of the input signals in order to determine which combination gives higher accuracy results.

As described in section II.C we associate each occurrence $i = 1 \dots 2N$ to the correspondent time series of *2m* samples of the currents and voltage signals around the time positions k_i. For each time series we calculate its logarithmic periodogram, and we truncate it to the first *p* components in order to obtain the SVM inputs. For this preliminary analysis we use *m=50* and *p=6*. Note that with this choice each time series has a length of *20ms*, which corresponds to one period of the main signal. In this analysis we use base parameters for the SVM, with a Gaussian kernel, in particular *C=1*,$\gamma = 1/p$.

Table II shows the results of the 3-fold cross-validation using different input signals. From row #1 to row #4 the input vector is the truncated logarithmic periodogram of *2m* samples of the indicated quantity; the sum of the two currents $I_1 + I_2$ is the signal which gives more accurate results. Rows #5 to row #8 show the results of the analysis performed by concatenating two different signals; in this case the input vector is the concatenation of the logarithmic truncated periodograms of the signals, and in this case $\gamma = 1/2p$. The use of the combination of voltage and the sum of the currents gives the best result in our analysis (more than 86% of the events are correctly evidenced).

In the last column we use the concatenation of the three signals (voltage and the two currents) obtaining the higher dimension of input vector (3p) leading to a value of $\gamma = 1/3p$.

TABLE II. RESULTS OF PRELIMINARY ANALYSIS.

Signal	3-fold CV
V_1	79.4236%
I_1	79.7276%
I_2	78.8723%
$I_1 + I_2$	80.3061%
$[V_1, I_1]$	85.3925%
$[V_1, I_2]$	86.0368%
$[V_1, I_1 + I_2]$	86.2614%
I_1, I_2	83.6865%
$[V_1, I_1, I_2]$	84.3671%

Following the previous results we search for the best SVM parameters using the best input configuration $[V_1, I_1 + I_2]$.

As performance index we use now a 5-fold CV, and we perform a grid-search approach [41] to obtain the best (C, γ) pattern for different choices of the following parameters:

- temporal window m = 50, 100, 150, 200;
- logarithmic periodogram truncation (frequency window): *p = 6, 12;*
- sigmoidal or radial basis function.

TABLE III. GRID SEARCH FOR BEST ACCURACY

Kernel	Frequency window *p*	Time window *m*	5-fold CV
RBF	6	50	87.3025
RBF	6	100	90.0481
RBF	6	150	90.8196
RBF	6	200	88.4597
RBF	12	50	87.6882
RBF	12	100	90.5473
RBF	12	150	92.6576
RBF	12	200	88.9589
SIG	6	50	83.7400
SIG	6	100	88.2328
SIG	6	150	89.3220
SIG	6	200	85.3283
SIG	12	50	85.8956
SIG	12	100	89.7304
SIG	12	150	90.5019
SIG	12	200	85.8049

Various approaches have been proposed in literature to determine the SVM parameters and some analytical procedures exist in particular for SVM regression methods [42]. For classification problems the selection of the parameter pair (C, γ) is a hard task which has also been approached using evolutionary optimization methods [43]. One of the most robust, efficient and well accepted techniques is the two step grid search approach described in [41], which consists in calculating the CV accuracy first in a coarse grid of exponentially growing values of (C, γ), and then, a second grid search is performed on a finer grid in the region where the better results have been found in the first step. In the following of this work, all the reported SVM results are obtained using the two step grid search approach which for instance can be easily parallelized.

Table III shows the 5-fold CV obtained in the different cases described above; we can observe that the RBF kernel in general outperforms the sigmoidal kernel, and that the 5-fold CV increases for increasing *m* from 50 to 150 but it decreases at 200, hence the best time window is of *2m=300* points, corresponding to three mains periods. Increasing the frequency window p from 6 to 12 gives in general a small improvement of the 5-fold CV. The best configuration is highlighted in Table III. For this choice the input of the SVM has dimension *2p=24*, and it is worth to note that in general the SVM algorithm is not significantly affected by the course of dimensionality, regarding the input dimension. So choosing *p=12* gives slightly better accuracy without significantly increasing the computational costs related to the smaller frequency window. For the best configuration highlighted in the table the SVM parameters found by using grid search are the following: *C=4871*, $\gamma = 8.6317 \cdot 10^{-5}$. It is important to note that, with a sampling frequency of 5kHz and a window of 300 points, the frequency resolution of the periodogram is 16.67 Hz . So the third harmonic is the mains period of 50Hz, and with 12 points of the periodogram we look up to 183.3 Hz.

The obtained accuracy of around 92% means that the SVM, once properly trained with the voltage and current input and phototube results, is capable of correctly classifying the 92% of time windows, detecting the presence or absence of an arc from the analysis of voltage and currents. The remaining 8% is related to false positives and false negatives. This percentage is an extremely good result from an industrial point of view, making the proposed method a potentially powerful instrument.

The other 6 test cases are relative to data sampled at 20 kHz; in all these cases the selected values are *M=1200* and *p=48:* Table IV shows the results in terms of 5-fold CV.

TABLE IV. RESULTS RELATIVE TO ALL DATASETS

Dataset#	Max CV(%)
2	90.6705
3	90.5795
4	95.3295
5	94.1932
6	87.4318
7	89.4432

The procedure gives good results, which means an accuracy of around 90% in the arc location.

C. Application of clustering

The time domain signals that have been processed with the clustering algorithm are voltage and current, while the phototube output and the train velocity are used as a validation to test the quality of the clustering analysis, i.e. to check if different clusters are really related to different arcing conditions.

As first result the authors have verified that the method works better with the currents signals than with the voltage signal, a result that is physically consistent with the nature of the arc, even though also the voltage presents characteristic behaviour in presence of the arcing phenomena.

When applying the clustering algorithm to the current data, no preprocessing technique has been used since any action (for instance filtering for noise reduction) could modify the signal and alter the pattern relative to the presence of electric arcs. In addition, the main goal is to develop a procedure which could be used any time new current data are acquired, hence a lower number of steps needed to run the procedure means easiness in its application

In order to keep the possibility to detect the arcs and locate them in time, the authors have decided to consider time windows of three periods, i.e. 60 ms, which at 20 kHz sample rate lead to a number of 1200 samples. So each recorded run (in particular the recorded current) has been divided into sections of m=1200 points and all the sections of a single run have been clustered with the proposed algorithm.

Consequently, the procedure clusters the currents or voltages logarithmic periodogram of each time window into different groups whose elements have similar characteristics. Due to the availability of the phototube data it is then possible to a-posteriori relate the different clusters to the arc length and magnitude. Based on this result a predictive maintenance procedure could be established with the availability of only currents and/or voltage data.

The dimensionality reduction has been performed by truncating the logarithmic periodogram from 600 frequency points down to *d*=10. Based on the authors' experience and on performed test, a further reduction would lead to information loss, hence worse performances of the clustering algorithm.

The above mentioned choices lead to the following problem dimension: for each run a number of about $n = 30 \cdot 10^3$ signal sections has been represented by their logarithmic periodogram of dimension *d*=10.

Due to the sampling time, truncation of the periodogram at the first 10 frequency samples means to limit our analysis up to frequencies of 166 *Hz*. Based on numerous tests an increase of *d* does not lead to any performance increase. Considering the presence of harmonics introduced by the traction equipment switching, which are generally of the order of 1 *kHz*, with this particular choice they do not affect the proposed technique. The same thing holds for the filtering process performed by the current transformer: such equipment has in general a cut-off frequency which is higher than 166 *Hz*.

As for the number of clusters, the main goal of the procedure would be to detect the at least the absence and presence of an arc (two clusters). By choosing only two clusters, though, we would prevent the procedure to identify additional clusters which might be of relevant meaning (i.e. arcs of different magnitude etc.). At the same time choosing a larger number of clusters might lead to slow convergence of the procedure and the existence of similar data associated to different clusters.

For each run we applied the k-means algorithm to find a partition with $c = 2,3,\ldots,20$ clusters. Calculating the Dunn index as in (23) for each partition always gives a clear indication that the best number of cluster should be selected as *c=4*, so in the following analysis we consider four clusters.

As an example, the result of the clustering procedure for a single run (run #1) is reported in Figure 11. The x axis now shows the number of sections as described before (each section contains three periods of the current waveform, i.e. 60 ms and 1200 time samples); in the top graph each different colour represent a cluster; the middle graphs shows the velocity of the train, coincident with the profile in Figure 8. In this case, since data are grouped in sections, at each point the velocity is calculated as the mean velocity of each 60 ms window.

Following the same principle, the phototube signal shown in the bottom graph, shows the maximum value of the sum of the two phototube signals in each section (i.e. 60 ms window). This could be apparently seen as a lost of information, but as a matter of fact a resolution of three main periods is enough for such application: the algorithm detects the presence of arcs every 60 ms. The authors have run the procedure also with time windows of 20 ms with basically no improvements of the accuracy of the technique.

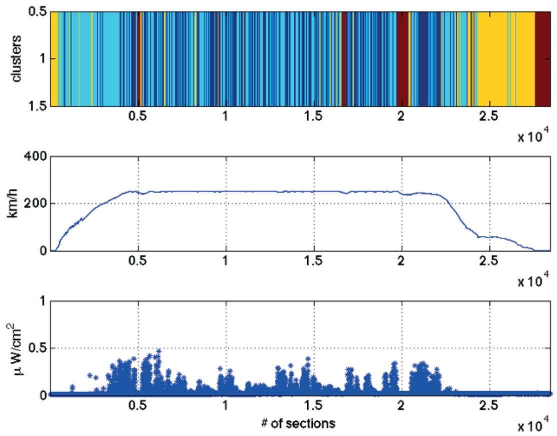

Figure 11. Result of the clustering procedure for run #1 (clusters, velocity and phototube)

In the top graph of Figure 11 different colours represent the four different clusters and it is evident how the temporal development of the run is divided according to the presence (or absence) of electric arcs. Further analysing Figure 11 we can say that the red and cyan cluster are related to the absence of arcs, while blue and yellow are related to the presence of arcs.

In Figure 12 the pantograph output has been replaced with the collected current (in particular, as done before, each point is the maximum value of the current in the 60 ms time window). We can observe that the red and cyan cluster (no arcs) correspond to lower values of currents, with the red in particular seems to be related to current values very close to zero which occur at the end of the run or during the run when the engines are disconnected.

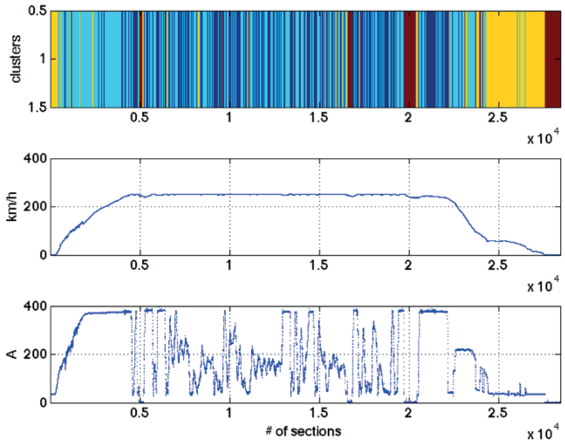

Figure 12. Result of the clustering procedure for run #1 (clusters, velocity and current)

The clustering results are clearer in Figure 13, in which the sections ordering do not follow time but they are sorted according to the clusters. Subgraph #2 shows the clusters, while subgraphs # 3 and #4 respectively show the electric arcs and velocity.

It is now evident that red and yellow cluster are related to the absence of electric arcs while the blue and cyan are related to the presence of arcs. In particular, observing Figure 13, we can notice that the blue cluster is related to arcs of higher magnitude while the cyan cluster is related to lower magnitude arcs. This is an important result, showing that by clustering current data it is possible to detect the presence of electric arcs and have an indication of their magnitude, besides localizing them in time.

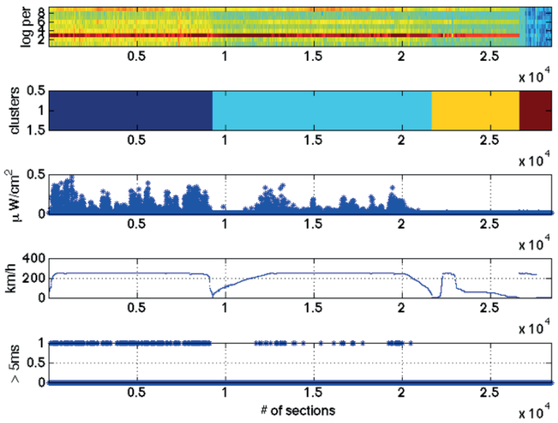

Figure 13. Result of the clustering procedure for run #1 (ordered)

In order to better investigate the difference between the two clusters associated to the presence of electric arcs, the bottom graph of Figure 13 shows the Boolean value relative to the arc length: the value is 1 if the arc has a duration longer than 5ms while it is 0 if the arc is shorter than 5 ms.

Analyzing again Figure 13 it is possible to see that most of the arcs with duration longer than 5 ms are concentrated in the blue cluster, which is another important result.

As for the top graph we can see how different clusters correspond to different logarithmic periodograms which the clustering algorithm is able to group.

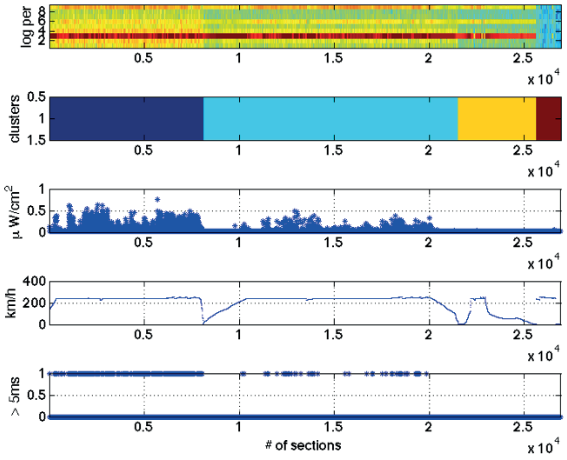

Figure 14. Result of the clustering procedure for run #2 (ordered)

The authors have run the same algorithm for each available run and in Figure 8 the results for run #2 are reported: the algorithm again groups the current data in 4 different clusters. The results show the same trend: arcing events are divided into two separate clusters, one related to higher magnitude arcs (in which most of the arcs of duration > 5ms are included) and the other one related to arcs of lower magnitude, and two clusters related to the absence of arcs, in which one of them is characterized by a zero current condition.

In order to generalize the results obtained and evaluate the possibility of an extended use of the algorithm, the authors have investigated how the clusters relative to different runs, found by the clustering algorithm, differ from each other.

Figure 15 shows the four centroids (relative to each cluster) for the 6 runs analysed; it is evident that the four centroids (each one composed by 10 points of the logarithmic periodogram) are well separated one from the other and are practically coincident.

This means that the results obtained above can be generalized: each of the four clusters is a footprint relative to the arc characteristics and is independent on the run, which means that it is independent of the of the track.

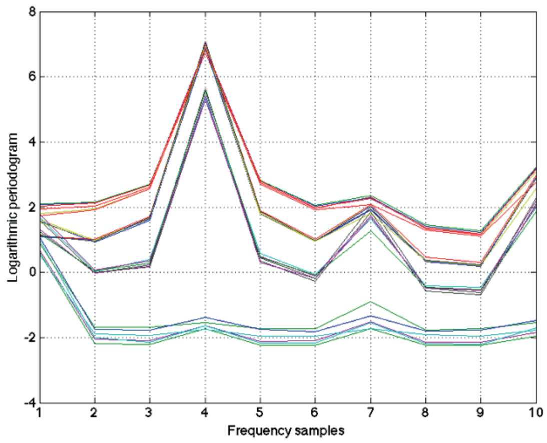

Figure 15. Centroids of the clusters for each run.

Based on this conclusion it is possible to compare the results of different runs: Figures 16 and 17 show the four clusters and the relative arcs for each run. Superimposed to the arcs magnitude there is the information relative to arcs having a duration > 5m (labelled with a circle of value 1). It is evident from the figures that:

a) the clustering algorithm is capable of separating the events detecting the presence of arcs from its absence;

b) the time windows characterized by the presence of an electric arc are separated into 2 clusters, depending on the arc magnitude (as detected by the phototube);

c) most of the arcs of a duration longer than 5 ms are grouped into the same cluster, even though there is a direct relation between the arc length and arc magnitude.

d) different clusters related to the absence of arc events, are related to different current magnitudes;

e) there is an extremely low number of arcs, as detected by the phototube, belonging to cluster #3 (in runs #4 to #6): they can be either low magnitude arcs that do not produce any effect on the train current or results originated by inaccuracy of the measurements.

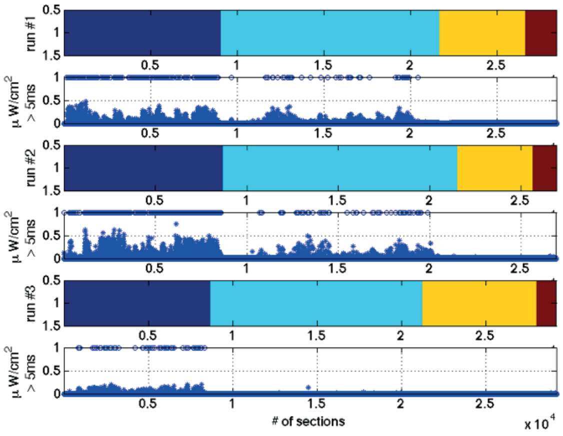

Figure 16. Results generalization for runs #1 to #3

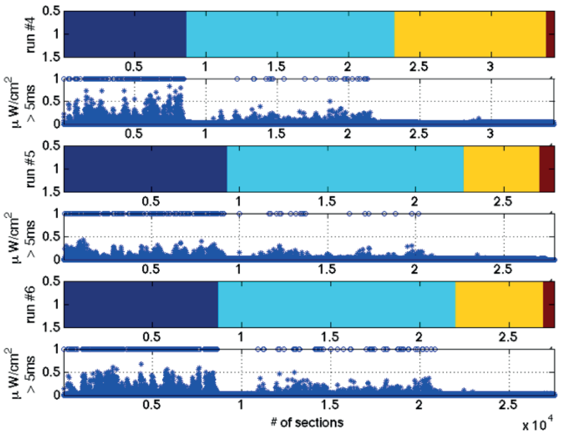

Figure 17. Results generalization for runs #4 to #6

As a general comment on these results, we can say that all the arcs of length > 5ms are localized in the first two clusters (as it was clear also from figures 16 and 17); in particular the percentage of such arcs which are in cluster #2 (instead of cluster #1) is around 6.5% (with the exception of Run #3 in which they are all in cluster #1). In the authors opinion this is a good results since the procedure is able to averagely classify the 93.5% of longer arcs in a single cluster, and the remaining 6.5% are classified in a second cluster which is anyway related to arc presence. Clusters #3 and #4 are relative to the absence of electric arcs, and in cluster #4 no current is measured.

V. Final comments and conclusions

This contribution shows the application of a set of AI based algorithm to the analysis of time series coming from measurements of physical quantities of electrical/electromechanical devices. The main rationale of the research is to develop modern and efficient algorithm capable of predicting incipient fault conditions, in order to implement the so-called predictive maintenance, identified as one of the key points by many manufacturers. The results show that a black box approach (with no deterministic simulation of the device behaviour either in nominal or faulty condition) is compatible with AI based techniques, and the commercial availability of computational power makes such algorithm usable not only in research environments but also in real life environments.

VI. References

[1] Henao, H., Capolino, G.-A., Fernandez Cabanas, M., Filippetti, F., Bruzzese, C., Strangas, E., Pusca, R., Estima, J., Riera-Guasp, M., Hedayati-Kia, S.: "Trends in Fault Diagnosis for Elec- trical Machines: A Review of Diagnostic Techniques", *IEEE Industrial Electronics Magazine*, 2014, 8, (2), pp. 31 - 42

[2] Garcia-Escudero, L.A., Duque-Perez, O., Morinigo-Sotelo, D., Perez-Alonso, M.: "Robust condition monitoring for early detection of broken rotor bars in induction motors", *Expert Systems with Applications*, 2011, 38, (3), pp. 2653 – 2660

[3] Garcia-Escudero, L. A., Duque-Perez, O., Fernandez-Temprano, M., Morinigo-Sotelo, D.: "Robust detection of incipient faults in VSI-fed induction motors using quality control charts", *IEEE Transactions on Industry Applications*, 2016, PP,(99), pp.1 - 1

[4] Ghate, V. N., Dudul, S. V.: "Cascade Neural-Network-Based Fault Classifier for Three-Phase Induction Motor", *IEEE Transactions on Industrial Electronics*, 2011, 58, (5), pp. 1555 - 1563

[5] Su, H., Chong, K. T.: "Induction Machine Condition Monitoring Using Neural Network Modeling", *IEEE Transactions on Industrial Electronics*, 2007, 54, (1), pp. 241 - 249

[6] Prieto, M.D., Cirrincione, G., Espinosa, A. G., Ortega, J. A., Henao, H.: “Bearing Fault Detection by a Novel Condition-Monitoring Scheme Based on Statistical-Time Features and Neural Networks”, *IEEE Transactions on Industrial Electronics*, 2013, 60, (8), pp. 3398 - 3407

[7] Garcia-Ramirez, A. G., Morales-Hernandez, L. A., Osornio-Rios, R. A., Benitez-Rangel, J. P., Garcia-Perez, A., Romero-Troncoso R.: “Fault detection in induction motors and the impact on the kinematic chain through thermographic analysis”, *Electric Power Systems Research*, 2014, 114, pp. 1 - 9

[8] Picazo-Rodenas, M.J., Royo, R., Antonino-Daviu, J., Roger-Folch, J.: “Use of the infrared data for heating curve computation in induction motors: Application to fault diagnosis”, *Engineering Failure Analysis*, 2013, 35, (15), pp. 178-192

[9] Singh, G. K, Al Kazzaz, S. A. S.: “Induction machine drive condition monitoring and diagnostic research - a survey”, *Electric Power Systems Research*, 64 (2), pp. 146 - 168

[10] Bellini, A., Filippetti, F., Tassoni, C., Capolino, G.-A.: “Advances in diagnostic techniques for induction machines”, *IEEE Transactions on Industrial Electronics*, 2008, 55, (12), pp. 4109 - 4126

[11] Grubic, S., Aller, J.M., Lu, B., Habetler, T.G.: “A Survey on Testing and Monitoring Methods for Stator Insulation Systems of Low-Voltage Induction Machines Focusing on Turn Insulation Problems”, *IEEE Transactions on Industrial Electronics*, 2008, 55, (12), pp. 4127 - 4136

[12] Zhang, P., Du, Y., Habetler, T.G., Lu, B.: “A survey of condition monitoring and protection methods for medium-voltage induction motors”, *IEEE Transactions on Industrial Applications*, 2011, 47, (1), pp. 34 - 46

[13] Zhou, W., Habetler, T.G., Harley, R.G.: “Bearing Condition Monitoring Methods for Electric Machines: a General Review”, *Proceedings of the IEEE International Symposium on Diagnostics for Electric Machines*, Power Electronics and Drives, Cracow, Poland, 2007, pp. 3 – 6

[14] Maru, B., Zotos, P.A.: “Anti-Friction Bearing Temperature Rise for NEMA Frame Motors”, *IEEE Transactions on Industrial Applications*, 2011, 25, (5), pp. 883 – 888

[15] Collina, A., Fossati, F., Papi, M., Resta, M.: “Impact of overhead line irregularity on current collection and diagnostics based on the measurement of pantograph dynamics”, *Proceedings of the Institution of Mechanical Engineers*, Part F: Journal of Rail and Rapid Transit, vol. 221, no. 4, pp. 547-559, 2007

[16] Elia, M., Diana, G., Bocciolone, M., Bruni, S., Cheli, F., Collina, A., Resta, F.: “Condition monitoring of the railway line and overhead equipment through onboard train measurements – an Italian experience”, *Proceedings of the IET Conference on Railway Condition Monitoring*, RCM 2006, pp. 102 – 107

[17] T. Usuda, T., Ikeda, M., Yamashita, Y.: “Prediction of contact wire wear in high speed railways”, *Proceedings of the 9th World Congress on Railway Research*, 2011, pp. 1 – 10

[18] Daadbin, A., Rosinski, J.: “Development, testing and implementation of the Pantograph Damage Assessment System (PANDAS)”, *Computers in Railways XII*, WIT Press, 2010, pp. 573 – 578

[19] Jutard, M., Fitaire, M., Le Duc, E.: “Moyens d'étude des arcs de rupture du contact pantographe-cateénaire”, *Revue Générale des Chemin de Fer*, vol. 108, no. 11, pp. 5 - 15, 1989

[20] Bruno, O., Landi, A., Papi, M., Sani, L.: “Phototube sensor for monitoring the quality of current collection on overhead electrified railways”, *Proceedings of the Institution of Mechanical Engineers*, Part F: Journal of Rail and Rapid Transit, vol. 215, no. 3, pp. 231-241, 2001

[21] Landi, A., Menconi, L., Sani, L.: “Hough transform and thermo-vision for monitoring pantograph-catenary system”, *Proceedings of the Institution of Mechanical Engineers*, Part F: Journal of Rail and Rapid Transit, vol. 220, no. 4, pp. 435-447, 2006

[22] Östlund, S., Gustafsson, A., Buhrkall, L., Skoglund, M.: “Condition Monitoring of Pantograph Contact Strip”, *Proceedings of 3rd Railway Condition Monitoring Conference*, Derby UK, June 2008

[23] Huang, H. H., Chen, T. H.: “Development of method for assessing the current collection performance of the overhead conductor rail systems used in electric railways”. *Proceedings of the Institution of Mechanical Engineers*, Part F: Journal of Rail and Rapid Transit, vol. 222, no. 2, pp. 159-168, 2008

[24] Wei, W., Liang, C., Yang Z., Xu, P., Yan, X., Gao, G., Wu, G.: “A novel method for detecting the pantograph–catenary arc based on the arc sound characteristics”, *Proc IMechE* Part F: J Rail and Rapid Transit 2019, Vol. 233(5) 506–515

[25] Qu, Z., Yuan, S., Chi, R., Chang, L., Zhao, L.: “Genetic Optimization Method of Pantograph and Catenary Comprehensive Monitor Status Prediction Model Based on Adadelta Deep Neural Network”, *IEEE Access*, vol. 7, Feb. 2019, pp. 23210 - 23221

[26] Shizea, H., Yachana, Z., Zhang, M., Xiaoxueb, H. “Arc detection and recognition in pantograph–catenary system based on convolutional neural network”, *Information Sciences*, vol. 501, 2019, pp. 363-376

[27] Barmada, S., Landi, A., Papi, M., Sani, L.: “Wavelet multi-resolution analysis for monitoring the occurrence of arcing on overhead electrified railways”, *Proc. Instn. Mech. Engrs*. vol. 217 part. F: J. Rail and Rapid Transit, 2003, pp. 177 - 187

[28] Haykin, S.: “Neural Networks: A comprehensive foundation”, Prentice Hall, Upper Saddle River, NJ, 1994

[29] Hotelling, H.: “Multivariate quality control – illustrated by the air testing of sample bombsights”, *Techniques of statistical analysis* (McGraw-Hill, New York, 1947), pp. 111– 184

[30] Sun, Y., Kamel, M.S., Wong, A.K.C., et al.: “Cost-sensitive boosting for classification of imbalanced data”, *Pattern Recognit*., 2007, 40, (2), pp. 3358– 3378

[31] De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L.: “The Mahalanobis distance”, *Chemometr. Intell. Lab. Syst*., 2000, 50, (1), p. 1–18

[32] Aparisi, F., de Luna, M.A.: “The design of the multivariate synthetic-T2 control chart”, *Commun. Stat. Theory Methods*, 2009, 38, (2), pp. 173–192

[33] Aparisi, F., Avendaño, G., Sanz, J.: “Techniques to interpret T2 control chart signals”, IIE Trans., 2006, 38, (8), pp. 647–657

[34] Mason, R. L., Chou, Y.M., Young, J.C.: “Applying Hotelling's T statistic to batch processes”, *J. Qual. Technol*., 2001, 33, (4), pp. 466–479

[35] Moya, M. M., Hush, D. R.: “Network constraints and multi-objective optimization for one-class classification”, *Neural Netw*., 1996, 9, (3), pp. 463– 474

[36] Caiado, J., Crato, N., Peña, D.: “A periodogram-based metric for time series classification”, *Comput. Stat. and Data Analysis*, vol 50, 2006, pp. 2668–2684

[37] Cortes, C., Vapnik, V.: “Support-vector network”, *Machine Learning*, vol. 20, pp. 273-297, 1995

[38] Boser, B. E., Guyon, I., Vapnik, V.: “A training algorithm for optimal margin classifiers”, *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ACM Press, pp. 144-152, 1992

[39] Platt, J. C.: “Fast training of support vector machines using sequential minimal optimization”, *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MA, 1998. MIT Press

[40] Hartigan, J. A., Manchek A. W.: “Algorithm AS 136: A k-means clustering algorithm”, *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 1979, 28 (1), 100-108

[41] Hsu, C. W., Chang, C. C., Lin, C. J.: “A practical guide to support vector classification”, *Tech. rep., Department of Computer Science*, National Taiwan University, 2003

[42] Cherkassky, V., Ma, Y.: “Practical selection of SVM parameters and noise estimation for SVM regression”, *Neural networks*, 2004 - Elsevier, Volume 17, Issue 1, 2004, Pages 113–126

[43] Friedrichs, F., Igel, C.: “Evolutionary tuning of multiple SVM parameters”, *Neurocomputing,* vol. 64, Mar 2005, pp. 107–117

Authors Name and Affiliation

Sami Barmada, Emanuele Crisostomi, Nunzia Fontana, Marco Raugi, Rocco Rizzo, Dimitri Thomopulos, Mauro Tucci
DESTEC, University of Pisa, Italy,
+39 050 2217312, sami.barmada@unipi.it