Sensors **2011**, *11*, 5695-5715; doi:10.3390/s110605695

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Information Theory Filters for Wavelet Packet Coefficient Selection with Application to Corrosion Type Identification from Acoustic Emission Signals

Gert Van Dijck *,† and Marc M. Van Hulle

Computational Neuroscience Research Group, Katholieke Universiteit Leuven, Herestraat 49, B-3000 Leuven, Belgium; E-Mail: marc.vanhulle@med.kuleuven.be

* Author to whom correspondence should be addressed; E-Mail: gert.vandijck@med.kuleuven.be; Tel.: +32-16-330-428; Fax: +32-16-345-960.

† Previous address: Materials Performance and Non-Destructive Evaluation (NDT), Department of Metallurgy and Materials, Katholieke Universiteit Leuven, Kasteelpark Arenberg 44, B-3001 Heverlee, Belgium.

Received: 8 April 2011; in revised form: 9 May 2011 / Accepted: 23 May 2011 / Published: 27 May 2011

Abstract: The damage caused by corrosion in chemical process installations can lead to unexpected plant shutdowns and the leakage of potentially toxic chemicals into the environment. When subjected to corrosion, structural changes in the material occur, leading to energy releases as acoustic waves. This acoustic activity can in turn be used for corrosion monitoring, and even for predicting the type of corrosion. Here we apply wavelet packet decomposition to extract features from acoustic emission signals. We then use the extracted wavelet packet coefficients for distinguishing between the most important types of corrosion processes in the chemical process industry: uniform corrosion, pitting and stress corrosion cracking. The local discriminant basis selection algorithm can be considered as a standard for the selection of the most discriminative wavelet coefficients. However, it does not take the statistical dependencies between wavelet coefficients into account. We show that, when these dependencies are ignored, a lower accuracy is obtained in predicting the corrosion type. We compare several mutual information filters to take these dependencies into account in order to arrive at a more accurate prediction.

Keywords: acoustic emission; chemical process industry; corrosion monitoring; feature subset selection; information theory; mutual information; Wavelet Packet Transform

1. Introduction

1.1. Corrosion Monitoring

A large part—25 to 40%—of the costs related to corrosion can be saved by the use of appropriate corrosion monitoring and control systems [1]. Corrosion monitoring provides feedback to operators about the state of the plant, information that in principle can be used for reducing the costs due to corrosion [1]. Direct costs can be avoided thanks to the increased reliability of the plant, avoidance of the disruption of the supply of products, decreased loss of capital and avoidance of lawsuits against companies (e.g., due to pollution caused by leaks of the installations), among other factors. Indirect costs can be equally important as these costs have an impact on the society and the environment. In some sectors, damage due to corrosion can be tolerated, but in the chemical, petrochemical and nuclear sectors, corrosion damage can be catastrophic, even resulting in the loss of lives and irreversible environmental damage.

Regular practice in the chemical process industry consists of periodic inspections of the plant, e.g., every 3 months, every 6 months or every year [2]. A recurring problem with such periodic inspections is that one can overlook the active damage that occurs in the plant; furthermore, immediately after inspection, the damage can continue to grow until the next periodic inspection is scheduled. Clearly, such cases should be avoided. A solution is offered by continuous monitoring procedures using corrosion monitoring systems. Different techniques are available for corrosion detection and monitoring in the chemical process industry [2,3]. In this research, we detect the most important types of corrosion in the chemical process industry from acoustic emission signals that are emitted during the corrosion process. Chemical reactions, as occurring during corrosion, emit acoustic activity [4,5] as well as the microscopic damage and fracture processes resulting from corrosion [6]. The acoustic emission technique has the advantage that it is low cost and allows for a continuous, on-line monitoring so that the damage can be detected as soon as it occurs [3].

1.2. Importance to Distinguish between Different Corrosion Types

The most frequent corrosion processes in the chemical process industry are: uniform corrosion (or general corrosion), pitting and stress corrosion cracking (SCC) [1,2]. It may also be possible that no corrosion process is active during the measurement. Therefore, we consider, in addition to the mentioned types, the absence of corrosion also needs to be discriminated.

There are at least two important reasons why researchers and industrial experts should be able to distinguish between different types of corrosion. Firstly, pitting and SCC are more harmful types of corrosion compared to uniform corrosion. Uniform corrosion reduces the thickness of the material relatively uniformly, hence taking a long time before holes are formed in the material. On the other hand, pitting causes pits and SCC causes cracks which can grow much faster, puncturing the material. This may lead to unexpected leaks in chemical plants. Therefore, occurrence of pitting and SCC Acoustic Emission (AE) events should advance the inspection of the installation.

Secondly, the discrimination between different corrosion processes should be performed prior to the quantitative analysis of correlating acoustic emission activity to the corrosion rate. In Seah *et al.* [7] a quantitative analysis has shown that the count rate (defined by the authors as the total number of

threshold crossings of AE signals per unit area of the exposed part of the metal sample and per unit time) is correlated with the rate of corrosion measured by means of the weight loss of the metal sample. A quantitative relation between the number of AE events and the number of pits in pitting as well with the pitted area and volume was established in Mazille *et al.* [8]. In stress corrosion cracking, a relationship between AE parameters (counts change per unit time and energy change per unit time) and the corrosion speed (change of crack length per unit time) has been established [9]. This shows that for several corrosion processes one can estimate the corrosion speed from AE parameters, although one should first link an AE event to the corresponding corrosion process. Erroneously relating AE events originating from pitting to SCC leads to a poor estimate of the corrosion speed of SCC and *vice versa*.

1.3. Wavelet Packet Feature Extraction and Selection from Acoustic Emission

Although future successes in corrosion prevention will still depend on selecting and developing more corrosion resistant materials, it is expected that the main progress in corrosion prevention will be achieved with better information-processing strategies and the development of more efficient monitoring tools that support corrosion control programs [10]. Feature extraction, feature subset selection, and classifier choice and design are all information-processing strategies that should be explored in the design of better corrosion monitoring systems.

Features to characterize the acoustic emission activity have often been obtained in the time-amplitude domain [2,5,11], the frequency domain [2,5,12], or the time-frequency domain using the Continuous Wavelet Transform (CWT) [13,14], the Discrete Wavelet Transform (DWT) [14] or the Wavelet Packet Transform (WPT) [15]. The process of constructing informative features that can help to discriminate between different classes is not trivial, but some generic approaches are available [16]. One of those approaches is to consider basis functions that can be used to extract features. A library of basis functions can be obtained from the Wavelet Packet Transform [17-19].

A challenge that arises after the extraction of wavelet coefficients with a Wavelet Packet Transform is the selection of a basis that is optimal in some sense, or the selection of a few coefficients for signal compression or pattern recognition purposes [18,20-22]. One of the most established algorithms to select wavelet coefficients for the prediction of a target variable, the corrosion type in this case, is the local discriminant basis (LDB) algorithm [20-22]. In previous research, we pointed to some disadvantages of the LDB algorithm [15]; in particular, the statistical dependency between wavelet coefficients, since it is leading to a lower prediction accuracy, but we did not come up at the time with a remedy to overcome it.

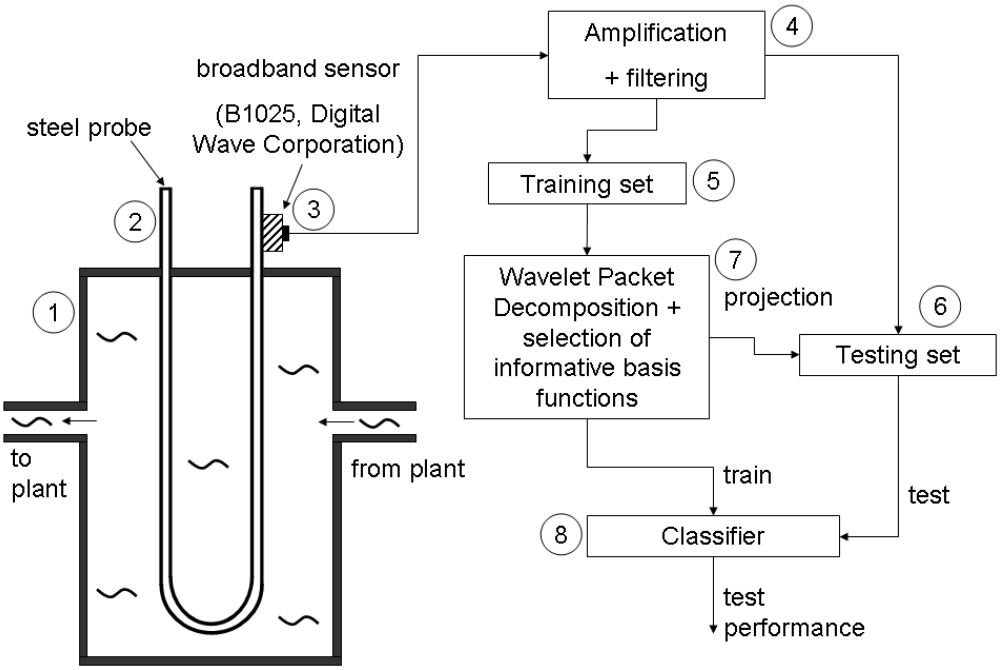
In the research reported in this article, we contribute to the selection of the most informative basis functions, from a library of wavelet packets, to distinguish between different types of corrosion, using information-theoretic criteria. We use the mutual information [23] to guide the search for informative basis functions by taking into account the statistical dependencies between the wavelet coefficients. The advantage of using the mutual information is that it easily enables us to take dependencies between features into account, *i.e.*, the wavelet coefficients in our case [24]. Moreover, the behavior of mutual information in feature selection is well-understood [24].

2. Materials and Methods

2.1. Signal Acquisition

This section describes the experimental set-up to obtain the acoustic emission signals. A U-shaped steel sample is shown in Figure 1.

Figure 1. Processing stages for making predictions of the corrosion type. A steel probe (2) is inserted in a bypass (1) of the chemical process plant and is therefore exposed to the same environmental conditions as the installation. Acoustic events are captured by means of a broadband sensor (3). Subsequently AE signals are amplified and filtered (4). In order to obtain a fair validation of the system, the acquired signals are split into a training (5) and testing set (6). Features are extracted from the training signals by means of a Wavelet Packet Decomposition (7). A classifier (8) is trained based on the selected wavelet coefficients of the training set. Testing signals are projected onto the selected basis functions. Subsequently, the wavelet coefficients of the testing signals are used to test the overall performance of the system.



The probe is designed such that the corrosion process occurring in the probe is representative for that in the plant [2]. Therefore, the probe is made of the same type of steel as the plant and the probe is exposed to the same environmental conditions, that is the same corrosive medium, temperature and pressure. This is represented in Figure 1 by means of the input flow that arrives from the plant and the output flow that is guided back to the plant, hence, forming a bypass of the process plant. Measuring the corrosion with a reference probe is based on some important considerations. The probe is relatively small: approximately 30 cm in height. This means that dampening of the waves when they propagate over such small distances is small. On the other hand when performing measurements on the large installation itself, AE waves may have dampened out before they reach a sensor when there is no

sensor in the neighborhood of the AE source. This would call for a dense network of AE sensors, leading to a more complex and expensive set-up. Moreover, due to the large difference in distances that waves could have travelled, AE events can be deformed to different degrees e.g., due to dispersion. This deformation will hamper the recognition of the type of corrosion from the waveforms. Thirdly, installations are often exposed to external sources that can create AE events: e.g., mechanical vibrations, rain drops, *etc.* These sources may be confounded with AE events originating from corrosion events.

The damage that occurs on the probe is captured by means of piezoelectric sensors attached to the corroding probe. In order to guarantee a good acoustical transfer from the probe to the sensor, a 'high vacuum' grease (DOW Corning®) is applied between the sensor and the probe. The sensors used here are broadband sensors (B1025, Digital Wave Corporation) [2]. These sensors have a guaranteed frequency bandwidth from 50 kHz to 2 MHz and can be used in a temperature range from −50 °C to 100 °C. Subsequently, the signals are amplified with an amplification factor of approximately 40 dB. The signals are then bandpass filtered between 50 kHz–2 MHz, because outside this range the sensor does not guarantee reliable information. Signals are sampled at 20 MHz or 25 MHz, both sampling rates are safely higher than the Nyquist sampling rate of 4 MHz for signals up to 2 MHz. Before computing the wavelet transform, signals are resampled to the same sampling rate (25 MHz) if they were sampled at 20 MHz.

2.2. Experimental Conditions

Two types of steel that belong to the most often used construction materials in the chemical process industry [1] are considered: carbon steel and stainless steel. The carbon steel considered here is German Material Number 1.0038, called S235JRG2 (DIN EN 10025) or RSt 37-2 (DIN 17100). The stainless steel considered here is German Material Number 1.4541, called X6CrNiTi18-10 (DIN EN 10088-2) and similar to AISI 321. The chemical composition of the two considered steel types can be found in [25]. All materials and experimental conditions are summarized in Table 1, together with the number of different experiments for the material-environment combinations (the environment is the combination of a corrosive medium and a temperature). The total number of time series obtained from these experiments is indicated between parentheses. The signals for each experiment were often collected over several days to obtain a representative set of signals. The acoustic emission data set contains 197 time series of "absence of corrosion" (indicating that no corrosion was active during these experiments), 194 time series of uniform corrosion, 214 time series of pitting and 205 time series of SCC. The time series have been assigned a corrosion class label by an expert [2] based on a visual inspection of the damage to the probe, the experimental conditions, and the inspection of the acoustic emission signals [2]. Each time series consists of "N" = 1,024 samples.

The different mechanisms that lead to the emission of acoustic events have been treated extensively in [2,6,15]. In Figure 2, we show some examples of different acoustic signals that were captured during different corrosion processes. Acoustic signals in the uniform corrosion experiments are characterized by a continuous-type acoustic emission signal [2,15], see also Figure 2(c,d). Localized forms of corrosion, such as pitting and stress corrosion cracking, lead to a burst-type acoustic activity, see Figure 2(e,f) and Figure 2(g,h) respectively.

Table 1. The steel types, the corrosive medium and the number of different experiments considered. The data was obtained from [2].

Type of corrosion	Material	Corrosive medium + conditions	Number of experiments (number of time series)	Total number of experiments per class (number of time series)
Absence of corrosion	1.0038	NaOH 20 weight% + NaCl 3 weight% 80 ℃	1 (99)	4 (197)
	1.4541	$CaCl_2$ 40 weight% 85 ℃	3 (98)	
Uniform corrosion	1.0038	H_3PO_4 10 weight% $T_{environment}$	6 (194)	6 (194)
Pitting	1.4541	brackish water + $FeCl_3$ 1 weight% 45 ℃	9 (214)	9 (214)
Stress corrosion cracking	1.0038	$Ca(NO_3)_2$ 60 weight% 105 ℃	1 (147)	10 (205)
	1.4541	$CaCl_2$ 40 weight% 85 ℃	9 (58)	

3. Wavelet Packet Decomposition

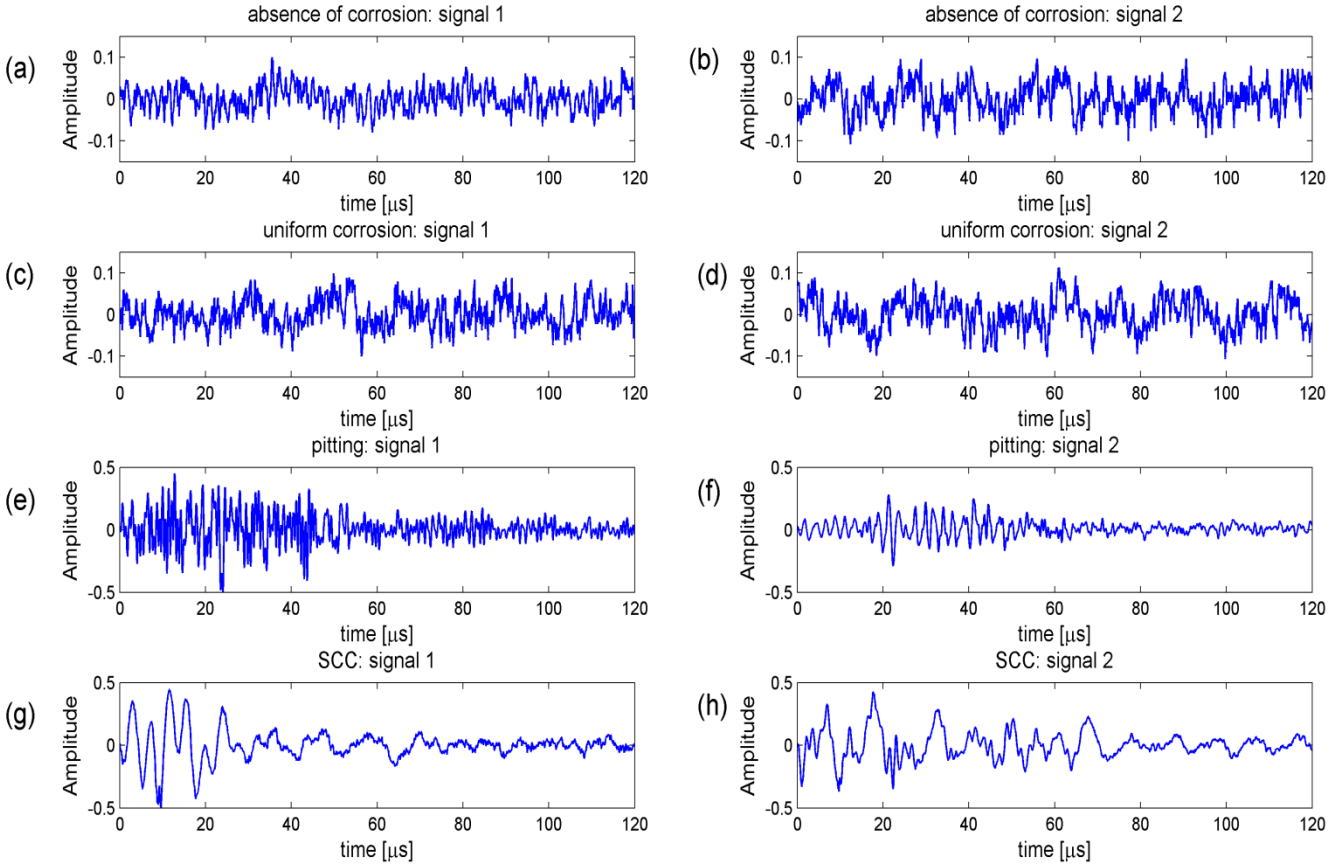
The basic approach for constructing features is to compute a number of general statistical parameters from time series such as the median, the mean, the standard deviation and higher-order moments. However, when restricting oneself to a limited number of parameters in advance, important information may be lost due to the implicit assumptions behind these parameters, e.g., the mean and standard deviation are only sufficient to characterize signals that consist of independent and identically distributed (i.i.d.) Gaussian noise.

A more thorough approach is to extract the wavelet coefficients from a wavelet packet decomposition (WPD) [26,27]. Wavelet packet decompositions offer a library of templates that have many desirable properties. First of all, WPD's are founded on a solid mathematical theory [27] that allows one to represent the signals in a new basis. The decomposition in a new wavelet packet basis guarantees that no 'information' is lost as the original signals can always be reconstructed from the new basis. Secondly, the templates in a wavelet packet decomposition can be interpreted in terms of frequencies and bandwidths [27]. Thirdly, wavelet packet decompositions are more flexible than the discrete wavelet transform and the Fourier transform. This means that the basis functions that are used in a discrete wavelet transform (DWT) are also available in the wavelet packet decomposition. We will use the wavelet coefficients, obtained from a wavelet packet decomposition, as the constructed features.

3.1. Wavelet Packet Decomposition Basics

The reader acquainted with wavelet packet decompositions may skip this section, which introduces the background to feature extraction from wavelet packet decompositions. This background is needed in order to understand the feature selection procedures in Sections 3.2 and 4. We will use the terminology of template and basis function interchangeably. Strictly speaking, a template is a more general terminology, because it does not need to be part of a basis.

Figure 2. Example signals of different corrosion types. The example of the absence of corrosion in **(a)** was captured from stainless steel in $CaCl_2$ 40 weight% at 85 ℃ environment. The example of the absence of corrosion in **(b)** was captured from carbon steel NaOH 20 weight% + NaCl 3 weight% at 80 ℃ environment. The examples in **(c)** and **(d)** are from continuous emissions during uniform corrosion of carbon steel in H_3PO_4 10 weight% at environment temperature. The signals in **(e)** and **(f)** are burst emission pitting signals captured from stainless steel in brackish water + $FeCl_3$ 1 weight% at 45 ℃ environment. In **(g)** a SCC burst emission signal was captured from stainless steel in $CaCl_2$ 40 weight% at 85 ℃ environment; **(h)** SCC burst emission signal was captured from carbon steel $Ca(NO_3)_2$ 60 weight% at 105 ℃ environment.



We represent a single time series by means of a sequence of observations x(t): x(0), x(1), … x(N-1), where 't' refers to the time index and 'N' is the number of samples. The time series x(t) can be considered as being sampled from an 'N' dimensional distribution defined over an N dimensional variable X(t): X(0), X(1), … X(N-1), we write this 'N' dimensional variable in short hand notation as $\mathbf{X}_{0:N-1}$.

Features are computed from a wavelet packet decomposition by computing the inner product between the templates and the time series (using a continuous representation, for the ease of notation):

$$\gamma_{i,j,k} = \langle x(t), \Psi_i^j(t - 2^i k) \rangle = \int_{-\infty}^{+\infty} x(t) \Psi_i^j(t - 2^i k)\, dt \quad (1)$$

A feature, in this case a wavelet coefficient, in the wavelet packet decomposition needs to be specified by the scale index 'i', frequency index 'j' and time index 'k'. The coefficient $\gamma_{i,j,k}$ can be considered as quantifying the similarity, by means of the inner product, between time series x(t) and wavelet function $\Psi_i^j(t-2^ik)$ at position 2^ik in time. The parameter 'i' is the scale index and causes a dilation (commonly called a 'stretching') of the wavelet function $\Psi^j(t)$ by a factor 2^i:

$$\Psi_i^j(t) = \frac{1}{\sqrt{2^i}}\Psi^j\left(\frac{t}{2^i}\right) \tag{2}$$

It is the parameter 'j' that determines the shape of the template. If we choose the 12-tap Coiflet filter [20], we obtain the first 16 different templates $\Psi^0(t)$, $\Psi^1(t)$, $\Psi^2(t)$,… $\Psi^{15}(t)$ shown in Figure 3. This 12-tap Coiflet filter has been consistently used in the experiments in Section 5. The construction of these basis functions can be found in text books [27]. The shapes of these basis functions also motivate the use of wavelet packet decompositions in our application. With an appropriate scaling and time shift some of the basis functions in Figure 3 resemble the AE bursts in Figure 2 (e–h). Choosing the appropriate template, the scaling factor and the time shift is the task of the feature selection procedure in Section 4.

In Figure 4, we show a graphical representation of the different subspaces that are obtained in a wavelet packet decomposition. In the discrete wavelet transform, the only nodes in the tree that are considered are W_1^1, W_2^1, W_3^1, W_4^1 and W_4^0 these subspaces are shaded in grey.

Figure 3. Templates (wavelet packets) corresponding to the 12-tap Coiflet filter.

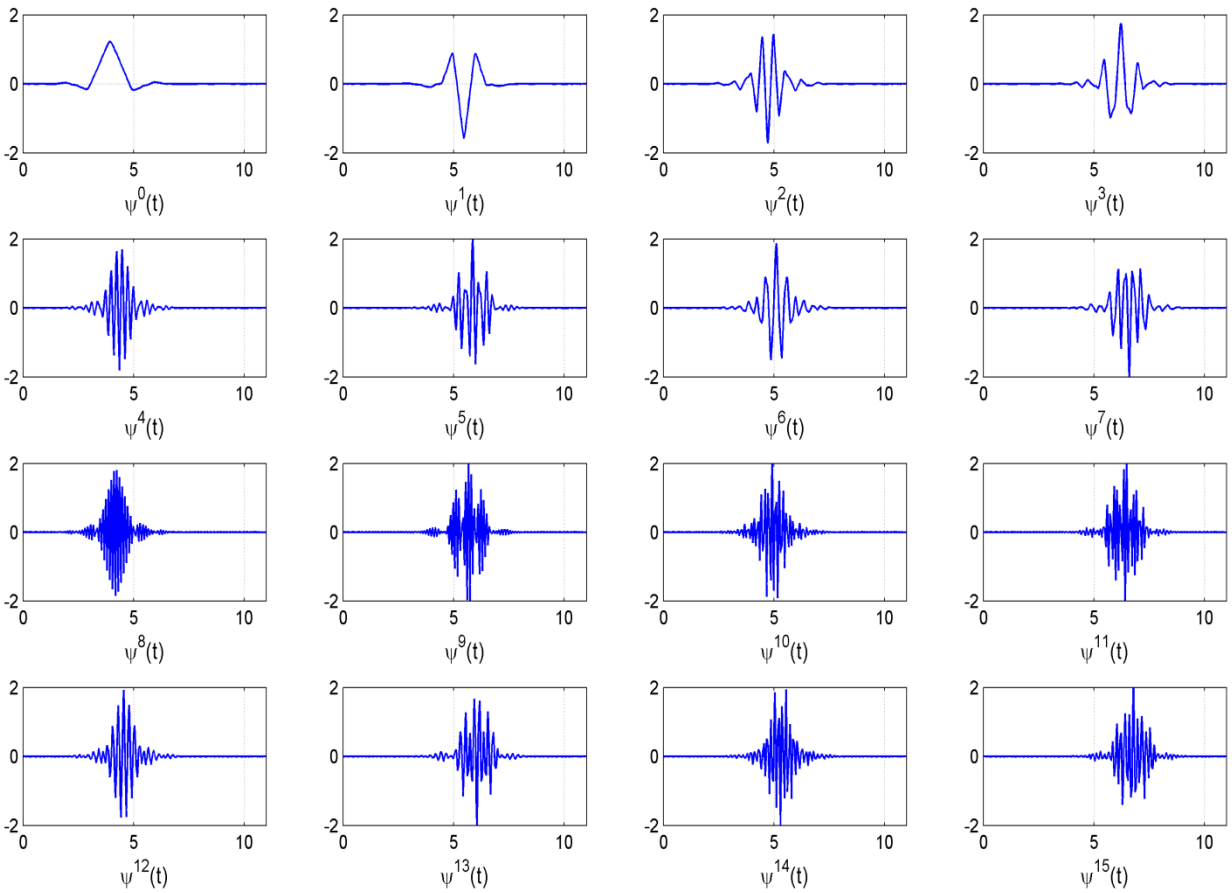
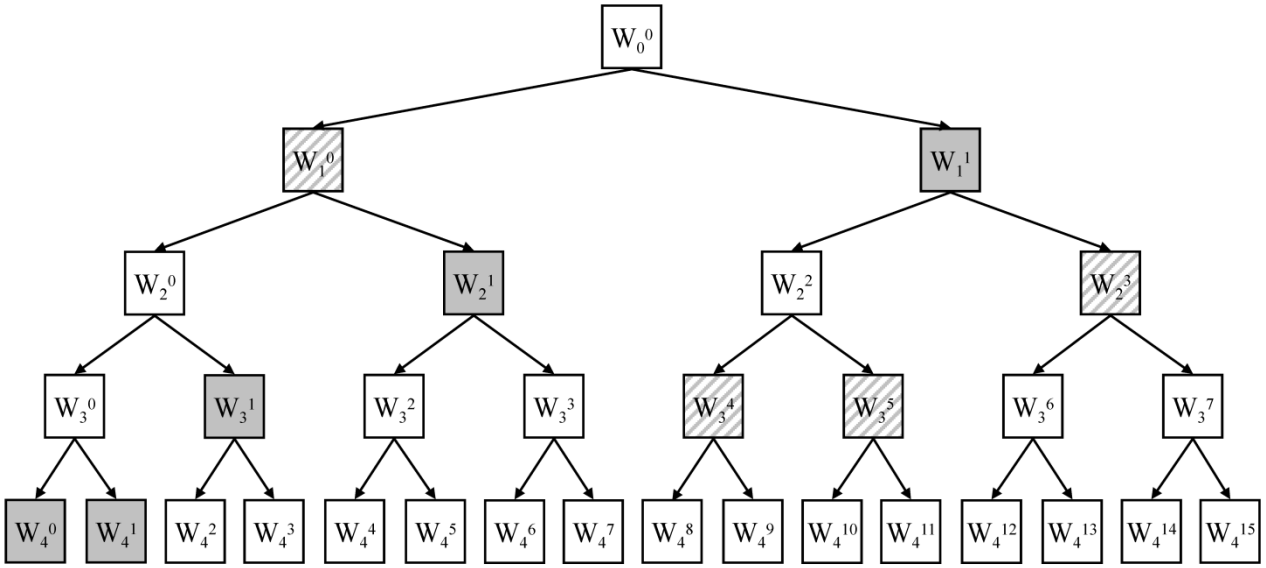


Figure 4. Library of wavelet packet functions. Different subspaces are represented by W_i^j. Index 'i' is the scale index, index 'j' is the frequency index. The depth 'I' of this tree is equal to 4. Every subtree within this tree, where each node has either 0 or 2 children, is called an admissible tree. Two admissible trees are emphasized, one shaded in grey and one marked with diagonals.



The first four subspaces are spanned by $\{\psi_1^1(t-2k)\}_{k\in\mathbb{Z}}$, $\{\psi_2^1(t-2^2k)\}_{k\in\mathbb{Z}}$, $\{\psi_3^1(t-2^3k)\}_{k\in\mathbb{Z}}$, and $\{\psi_4^1(t-2^4k)\}_{k\in\mathbb{Z}}$ respectively. Subspace $W_4{}^0$ is spanned by $\{\psi_4^0(t-2^4k)\}_{k\in\mathbb{Z}}$. So in the discrete wavelet transform, the signals are only analyzed by means of the time translated functions of $\Psi_4^0(t)$. Note that $\Psi_0^0(t)$ is called the scaling function, shown as the first template in Figure 3, and the dilated and time translated functions of $\Psi_0^1(t)$ (the latter is called the mother wavelet function and is shown as the second template in the top row of Figure 3). In Figure 4, only two bases are shown: the gray shaded basis corresponds with the discrete wavelet transform, the basis marked with diagonals is chosen arbitrarily and is one of the possible bases in the wavelet packet decomposition. The basis marked with diagonals puts more emphasis on a finer analysis of the higher frequency part of the signals.

Retaining any binary tree in Figure 4, where each node has either 0 or 2 children, leads to an orthonormal basis for finite energy functions, denoted as $x(t) \in L^2(\mathbb{R})$:

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt < \infty \tag{3}$$

Such a tree is called an admissible tree. If the leaves of this tree are denoted by $\{i_l, j_l\}_{1 \le l \le L}$ the orthonormal system can be written as:

$$W_0^0 = \oplus_{l=1}^{L} W_{i_l}^{j_l} \tag{4}$$

This means that the space $W_0{}^0$, which is able to represent the input space of the time series, can be decomposed into orthonormal subspaces $W_{i_l}^{j_l}$.

It should be noted that a full wavelet packet decomposition yields too many features. In cases where one can assume that the exact time location 'k' of the template is of no importance, one can, e.g.,

consider an average or the energy of wavelet coefficients over time for each possible combination of the scale index 'i' and the frequency index 'j'. This will lead to fewer features to be selected from. Here, we will consider the full complexity of the problem, when the exact time location of the template can be of importance, and consider all coefficients from a full wavelet packet decomposition as selectable.

A full wavelet packet decomposition leads to $N \times (\log_2 N + 1)$ features. This can be seen as follows. From Figure 4, it can be noted that the number of subspaces at a certain scale 'i' is determined by the scale index 'i'. The number of subspaces at scale 'i' is equal to 2^i. Therefore the frequency index 'j' at a certain scale 'i' will be an integer from $[0, 2^i - 1]$, indicating the starting position of the subspace at scale 'i'.

As can be seen from Equation (1), at scale 'i', the inner products are computed at discrete time instants $2^i k$. Therefore, at scale 0, we obtain 'N' (length of the signals) coefficients: $\gamma_{0,0,0}, \ldots \gamma_{0,0,N-1}$. At the next scale, 'i' = 1, we obtain N/2 coefficients in each subspace *i.e.*, $\gamma_{1,0,0}, \ldots \gamma_{1,0,N/2-1}$ and $\gamma_{1,1,0}, \ldots \gamma_{1,1,N/2-1}$.

At the highest frequency resolution 'i' = $\log_2 N$, and we obtain coefficients: $\gamma_{\log N,0,0}, \ldots \gamma_{\log N,N-1,0}$. Hence, at each scale, there are 'N' coefficients, and in total there are $\log_2 N+1$ different scale levels. This leads overall to $N \times (\log_2 N + 1)$ different coefficients to select from. The variables that can be associated with the coefficients $\gamma_{i,j,k}$ are further denoted by capitals $\Gamma_{i,j,k}$.

3.2. Local Discriminant Basis

In this section, we consider the selection of the most discriminative basis functions $\Psi_i^j(t - 2^i k)$ in order to make a prediction about the target variable '*y*' (*i.e.*, the corrosion class). The target variable is a class variable taking values 1 … *#C*, where *#C* is the total number of classes. An outline of the Local Discriminant Basis algorithm [22] is provided. We assume that we are given a set of training signals x_j and, for each one of them, we are given the associated target class c_j:$\{(x_j,c_j)\}$.

Step 0: Expand each training signal into a time-frequency dictionary D: this involves the computation of all coefficients $\gamma_{i,j,k}$ for each training signal, and assumes that we choose a particular conjugate mirror filter [27] in advance, which will define the templates.

Step 1: Estimate the class conditional probability density functions $\hat{p}^y(\Gamma_{i,j,k})$ (PDF's) for each wavelet coefficient variable, $\Gamma_{i,j,k}$, in the dictionary. Superscript '*y*' refers to the class label, with $y = 1, 2, \ldots \#C$ and *#C* is the total number of classes. These PDF's were estimated by means of the averaged shifted histograms method (ASH) as in Saito *et al.* [22].

Step 2: For each wavelet coefficient variable, $\Gamma_{i,j,k}$, compute the discriminant measure $\delta_{i,j,k}$. The computational cost of this procedure is $O((N+1)\log_2 N)$. Many discriminant measures can be used in practice. We use the symmetric relative entropy, Equation (5), as in Saito *et al.* [22]. The relative entropy for $\Gamma_{i,j,k}$ between two classes, $y = 1$ and 2, can be computed as [23]:

$$D\left(\hat{p}^1(\Gamma_{i,j,k}), \hat{p}^2(\Gamma_{i,j,k})\right) \triangleq \int \hat{p}^1(\gamma_{i,j,k}) \log \frac{\hat{p}^1(\gamma_{i,j,k})}{\hat{p}^2(\gamma_{i,j,k})} d\gamma_{i,j,k} \quad (5)$$

Because this discriminant measure is not symmetric, a symmetric version is obtained as:

$$\delta_{i,j,k} = D^s\left(\hat{p}^1\left(\Gamma_{i,j,k}\right), \hat{p}^2\left(\Gamma_{i,j,k}\right)\right) = D^s\left(\hat{p}^1\left(\Gamma_{i,j,k}\right), \hat{p}^2\left(\Gamma_{i,j,k}\right)\right) + D^s\left(\hat{p}^2\left(\Gamma_{i,j,k}\right), \hat{p}^1\left(\Gamma_{i,j,k}\right)\right) \quad (6)$$

When more than two classes are considered, $\delta_{i,j,k}$, is defined as the sum over all (*#C*.(*#C* − 1))/2 pairs of different classes as:

$$D^S_{Pair}\left(\hat{p}^1\left(\Gamma_{i,j,k}\right), \hat{p}^2\left(\Gamma_{i,j,k}\right), \ldots, \hat{p}^c\left(\Gamma_{i,j,k}\right)\right) = \sum_{m=1}^{\#C-1} \sum_{n=m+1}^{\#C} D^s\left(\hat{p}^m\left(\Gamma_{i,j,k}\right), \hat{p}^n\left(\Gamma_{i,j,k}\right)\right) \quad (7)$$

Step 3: Evaluate the discriminant power of each basis $B \in D$ (the dictionary) and obtain the best basis Ψ for which the discriminant power is maximal:

$$\Psi = \arg\max_{B \in D} \sum_{(i,j,k) \in B} \delta_{i,j,k} \quad (8)$$

Hence, one searches for the indices (*i*,*j*,*k*) such that the associated basis functions form a basis B. This corresponds also with the search for an admissible tree in Figure 4, with the largest discriminant power.

Step 4: Select '*m*' basis functions, $\Psi_i^j(t - 2^i k)$, from Ψ corresponding to the '*m*' largest $\delta_{i,j,k}$. The number of basis functions '*m*' to be retained is not determined in Saito *et al.* [22]. Therefore, we perform experiments for '*m*' ranging from 1 to 50 basis functions.

Step 5: Construct classifiers using the '*m*' coefficients, $\gamma_{i,j,k}$. Experiments with different classifiers are performed in Section 5.

In Step 3, the algorithm searches a basis Ψ for which the discriminant power is maximal. However, the total discriminant power in Step 3 is computed as the sum of the discriminant measures of each of the coefficients in a basis *B*: $\sum_{(i,j,k) \in B} \delta_{i,j,k}$.

The additive property of the discriminant powers of coefficients in a basis leads to a very rapid search for the basis with the highest discriminant power. It is easily seen that an optimal basis can be found in O(*N*) comparisons, with '*N*' the length of the signal, see Mallat [27]. However, as we showed before [15], the sum of the discriminant measures in Equation (8) does not necessary reflect the joint discriminant power, *i.e.*, taking the joint probability distribution of the wavelet coefficients into account. It will only be the case when the wavelet coefficients are class conditional independent [15]. When some wavelet coefficients are highly correlated, they may capture essentially the same information and, hence, the joint discriminant power is not simply a sum of the marginal discriminant measures. The consequence is that the accuracy in classification prediction may increase at a much slower rate compared to the case when the dependencies between the coefficients are taken into account. This is exactly what we will show in Section 5. So far we did not present a solution to take the dependencies into account. In Section 4, we present information theoretic filter feature selection approaches to serve this purpose.

4. Information Theory Filter Feature Selection Approaches

The feature selection procedures based on the mutual information are called filter approaches, due to the fact that the classifier used in the prediction is not involved in the selection of the features [28].

An alternative approach is the wrapper approach [28] in which the classification algorithm is involved in the selection of the features. The wrapper approach is often computationally more expensive, but may lead to a higher classification accuracy. A follow-up paper that combines a wrapper approach and a filter approach in a so called hybrid filter-wrapper approach is in preparation. The reason to use mutual information here is that it is a well-established criterion for taking dependencies between variables or features into account [24]. The high dimensional mutual information between a feature vector **F** and class variable C can be defined as:

$$MI(F;C) = \sum_{c=1}^{\#C} \int_{F} P(f,c)\, log_2 \left(\frac{P(f,c)}{P(f)P(c)} \right) df \tag{9}$$

We perform a sequential forward search (SFS) over all wavelet coefficients using a mutual information criterion. In the SFS, we start with the empty feature set S = {∅} as the selected coefficients so far and the whole dictionary D = $\{\Gamma_{i,j,k}\}$, with $0 \leq i \leq \log_2 N$, $0 \leq j \leq 2^i - 1$ and $0 \leq k \leq N/(2^i) - 1$, as the available feature set. In each iteration of the SFS, the variable $\Gamma'_{i,j,k}$, which achieves the highest value of the mutual information criterion, taking into account the previously selected features, is selected. S is updated in each iteration as: $S = S \cup \Gamma'_{i,j,k}$ and the dictionary is updated as $D = D\backslash\{\Gamma'_{i,j,k}\}$. Three different mutual information criteria were compared for the SFS filter: a density-based method (Section 4.1), a distance-based method (Section 4.2) and a relevance-redundancy method (Section 4.3).

4.1. Parzen Window Density (MI Parzen)

The estimation of the mutual information by means of a Parzen window density estimator was proposed in [29]. This is a probability density based mutual information estimator. If a Gaussian window function is used, the mutual information is estimated as (a hat is used to indicate an estimator):

$$\widehat{MI}(F;C) = \hat{H}(C) - \hat{H}(C|F) \text{ with}$$

$$\hat{H}(C|F) = -\sum_{j=0}^{n} \frac{1}{n} \sum_{c=1}^{\#C} \hat{p}(c|f_j) \log_2 \hat{p}(c|f_j) \tag{10}$$

$$\hat{p}(c|f) = \frac{\sum_{j \in I_c} \exp\left(-(f - f_j)^T \Sigma^{-1} (f - f_j)/2h^2\right)}{\sum_{k=1}^{C} \sum_{j \in I_k} \exp\left(-(f - f_j)^T \Sigma^{-1} (f - f_i)/2h^2\right)} \tag{11}$$

The functional H(.) is the entropy [23]. Further, I_k is the set of indices of data points which belong to class "k", $\mathbf{f}_j$ is the feature vector of the j'th training data point and #C is the total number of classes. The covariance matrix $\boldsymbol{\Sigma}$ is estimated as the full sample covariance matrix. The parameter "h" is set to a default value as suggested in the experiments in [29]: $h = 1/\log_2(n)$, where "n" is the sample size of the training set. This estimator is referred to as "MI Parzen".

4.2. K-Nearest Neighbors (MI knn)

Instead of estimating the probability density functions, the mutual information between a discrete class variables and a feature vector **F** can be estimated based on the pairwise distances between data

points. We presented such an approach for feature selection, in case of a discrete target variable, in [30]. The mutual information estimator relies on the Kozachenko-Leonenko entropy estimator [31] of the differential entropy:

$$\widehat{H}(F|c) = -\psi(k) + \psi(n_c) + \ln(c_d) + \frac{d}{n_c}\sum_{i \in I_c} \ln\left(\varepsilon_c(i,k)\right) \quad (12)$$

which is plugged into:

$$\widehat{MI}(F;C) = \widehat{H}(F) - \sum_{c=1}^{\#C} \widehat{H}\,(F|c)\hat{p}(c) \quad (13)$$

In Equation (12), ψ(.) is the psi-function, "n_c" the number of training data points in class "c", ε_c(i,k) is twice the distance from the i'th data point in class "c" to its k'th neighbor in class "c" in the training set, "d" the dimensionality of the data points and "c_d" the volume of the d-dimensional unit ball. We used the Euclidean distance between data points, in this case "c_d" = $\pi^{d/2}/\Gamma(1 + d/2)$, with Γ(.) the gamma-function.

The unconditional entropy $\widehat{H}(\mathbf{F})$ in Equation (13) can be estimated similarly as the conditional entropy in Equation (12), but with "n_c" replaced with the total number of training points "n" and ε_c(i,k) replaced by ε(i,k), *i.e.*, twice the distance from data point "i" to its "k" nearest neighbor when all training data points from all classes are merged into one set. The prior probabilities $\hat{p}(c)$ are estimated as the number of training points in class "c" divided by the total number of training points as follows: n_c/n. In the experiments, the number "k" of nearest neighbors was set equal to 6. This estimator is referred to as "MI knn".

4.3. Relevance-Redundancy Approach

Relevance-redundancy approaches select features that are highly relevant with respect to the class variable, but penalize a feature if it is redundant with respect to previously selected features. These approaches often use mutual information to estimate both the relevance and the redundancy. Suppose that F_i is a candidate feature to be selected and that S is the set of already selected features; a relevance-redundancy criterion based on the normalized mutual information [32] is then obtained as:

$$Crit_S(F_i, C) = MI(F_i; C) - \frac{1}{|S|}\sum_{F_S \in S} \frac{1}{min\{H(F_i), H(F_S)\}} MI(F_i; F_S) \quad (14)$$

where |S| is the size of the set of already selected features. Note that, as opposed to Equations (11) and (13), here only the lower dimensional $MI(F_i;C)$ and $MI(F_i;F_s)$ are required. Note that the normalization in Equation (14) is achieved by dividing $MI(F_i;F_s)$ through $\min\{H(F_i), H(F_s)\}$. The ratio $\frac{1}{\min\{H(F_i),H(F_s)\}} MI(F_i; F_s)$ will be a value between 0 and 1, because $MI(F_i;F_s)$ is always smaller or equal to the minimum of $H(F_i)$ and $H(F_S)$, hence, this ratio is called the normalized mutual information [32]. In Equation (14), the mutual information $MI(F_i;C)$ quantifies the relevance of feature F_i with respect to the target variable 'C', it will be large when F_i is highly relevant. The term $\frac{1}{|S|}\sum_{F_S \in S} \frac{1}{\min\{H(F_i),H(F_s)\}} MI(F_i; F_s)$ quantifies the redundancy of F_i with the already selected features $F_s \in S$. When F_i and F_s are strongly dependent, or correlated in a more stricter sense,

$\frac{1}{\min\{H(F_i),H(F_s)\}} MI(F_i;F_s)$ will be large, hence the relevance term in Equation (14) $MI(F_i;C)$ will be penalized. This allows features that are less relevant, but have a very low redundancy with the already selected features, to be included.

In the computation of the normalized mutual information, the features were first discretized into 3 states [33]: values of $F_i < \mu(F_i) - (\sigma(F_i))/2$ were set to state 0, $\mu(F_i) - (\sigma(F_i))/2 \leq F_i \leq \mu(F_i) + (\sigma(F_i))/2$ were set to state 1 and values of $F_i > \mu(F_i) + (\sigma(F_i))/2$ were set to state 2. Note that $\mu(F_i)$ and $\sigma(F_i)$ are, respectively, the mean and standard deviation of F_i. The mutual information was then computed from the contingency tables of the discretized features, *i.e.*, from the co-occurrences of the states of different features.

5. Results and Discussion

We tested four different popular classifiers to predict the different corrosion types:

- k-nearest neighbor (knn): the Euclidean distance is used with "k" set to 3, see Section 4.5.4 in [34] for a reference on k-nearest neighbor classification;
- decision tree J48 (WEKA's implementation of C4.5) from WEKA package 3.4.1 [35], we used the default values from the WEKA package, *i.e.*, the minimum number of instances per leaf (-M) equal to 2 and the confidence factor for pruning (-C) is equal to 0.25, see Section 8.4.2 in [34] for a reference on decision trees;
- Gaussian Mixture Model (GMM): the number of Gaussians per class is taken equal to 1 in the experiments and hence each class is modeled as a multivariate Gaussian distribution (see, e.g., McLachlan and Peel [36] for a reference on Gaussian mixture modeling);
- naïve Bayes classifier (NB) from WEKA package 3.4.1 [35] with kernel estimation (-K) for modeling numeric attributes, see Section 2.12 in Duda *et al.* [34].

In the validation of the different algorithms, we performed a 10-fold cross-validation [37]. This implies that 10 different training sets and 10 different testing sets are considered and that each data point is used once as test data in the validation. We compute the test classification performances on the sets that have not been considered in the selection of the wavelet coefficients nor in the training of the classifiers to avoid overfitting [37]. We let 'm', the number of selected wavelet coefficients, range from 1 to 50 coefficients. The test classification accuracies for the knn, decision tree, Gaussian mixture model and naïve Bayes classifiers are shown in Figures 5, 6, 7 and 8 respectively.

We stopped feature selection after 50 features have been selected, as can be observed from Figures 5 to 8 the testing performances of the different feature selection algorithms have leveled off at that moment. In practice, one can use a stopping rule to determine automatically how many features should be retained. This can be achieved as follows. The data is split into three parts: a training set, a validation set and a testing set. The feature selection can be stopped when the performance on the validation set does not increase further using the training set to train the machine learning algorithm. The final performance is then obtained on the testing set using the training and validation set to train the machine learning algorithm. This can be iterated in a cross-validation procedure, so that all data have been used for testing once. Note that the computational cost of feature selection algorithms will increase, because an additional validation step is included.

Figure 5. Evolution of the accuracy of the k-nearest neighbor classifier (k = 3) as a function of the number of wavelet coefficients selected with the LDB algorithm and the mutual information filter algorithms. The horizontal line indicates the accuracy when all 1,024 samples are used (no FSS).

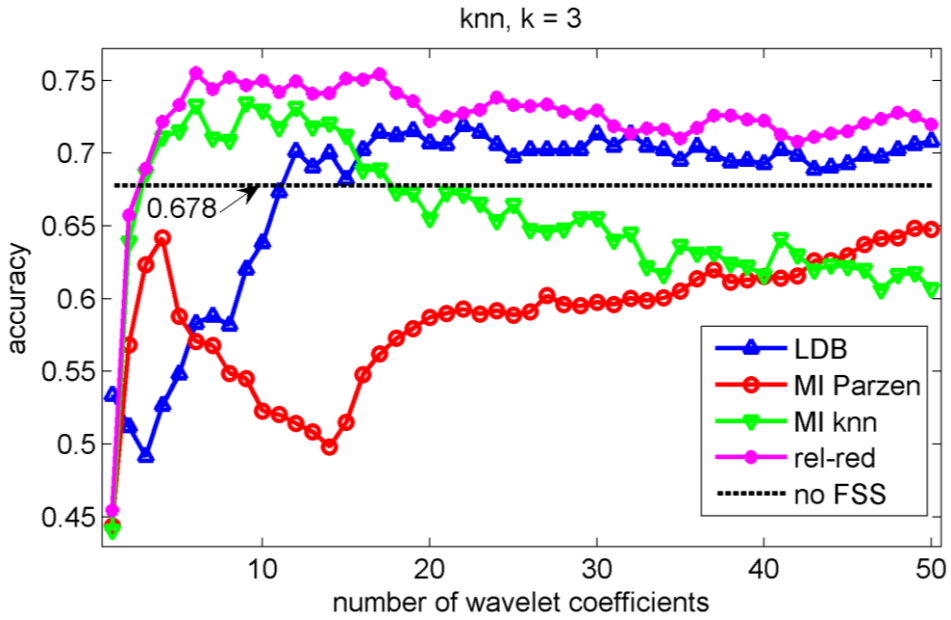


Figure 6. Evolution of the accuracy of the decision tree J48 classifier as a function of the number of wavelet coefficients selected with the LDB algorithm and the mutual information filter algorithms. The horizontal line indicates the accuracy when all 1,024 samples are used.

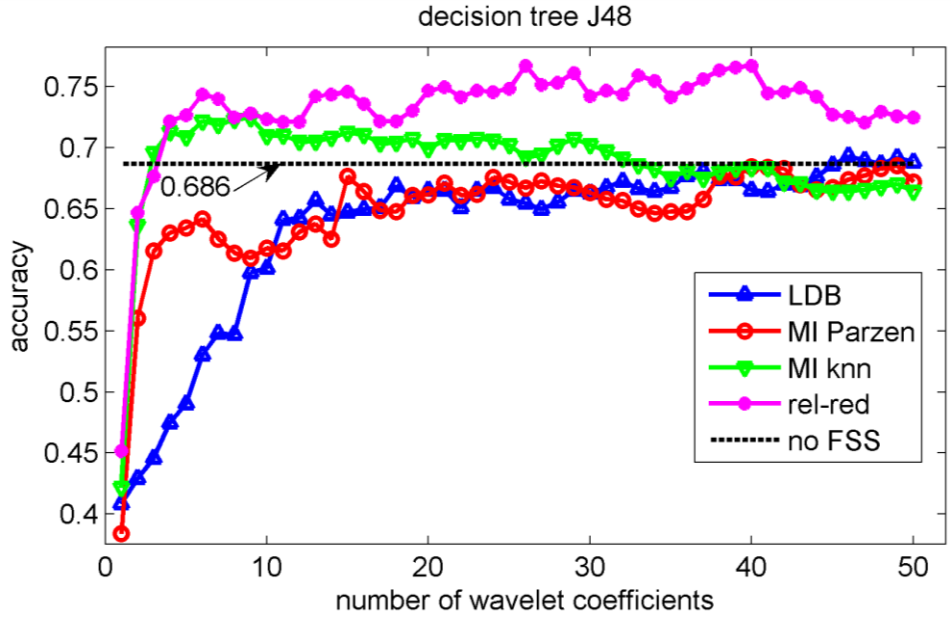


Figure 7. Evolution of the accuracy of the Gaussian mixture model as a function of the number of wavelet coefficients selected with the LDB algorithm and the mutual information filter algorithms. The horizontal line indicates the accuracy when the 1,024 samples were sub-sampled with a factor 15 to avoid numerical problems in the estimation of the parameters of the model. This subsampling was performed by taking the first time sample and then every 15th sample.

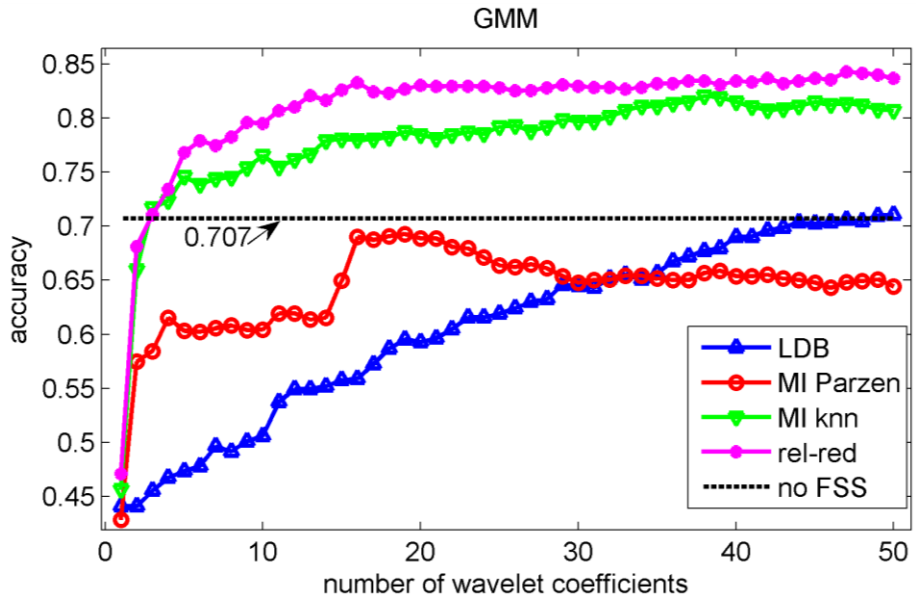
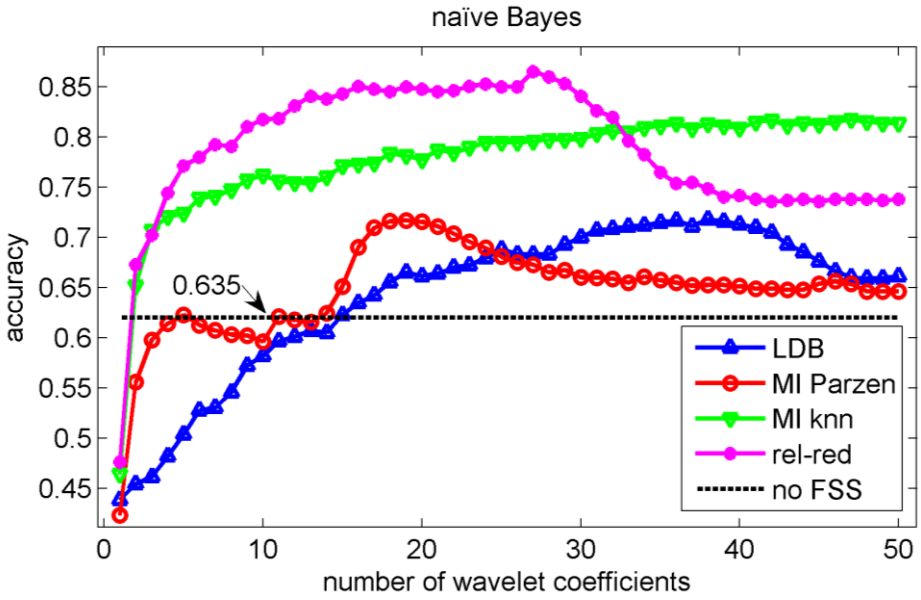


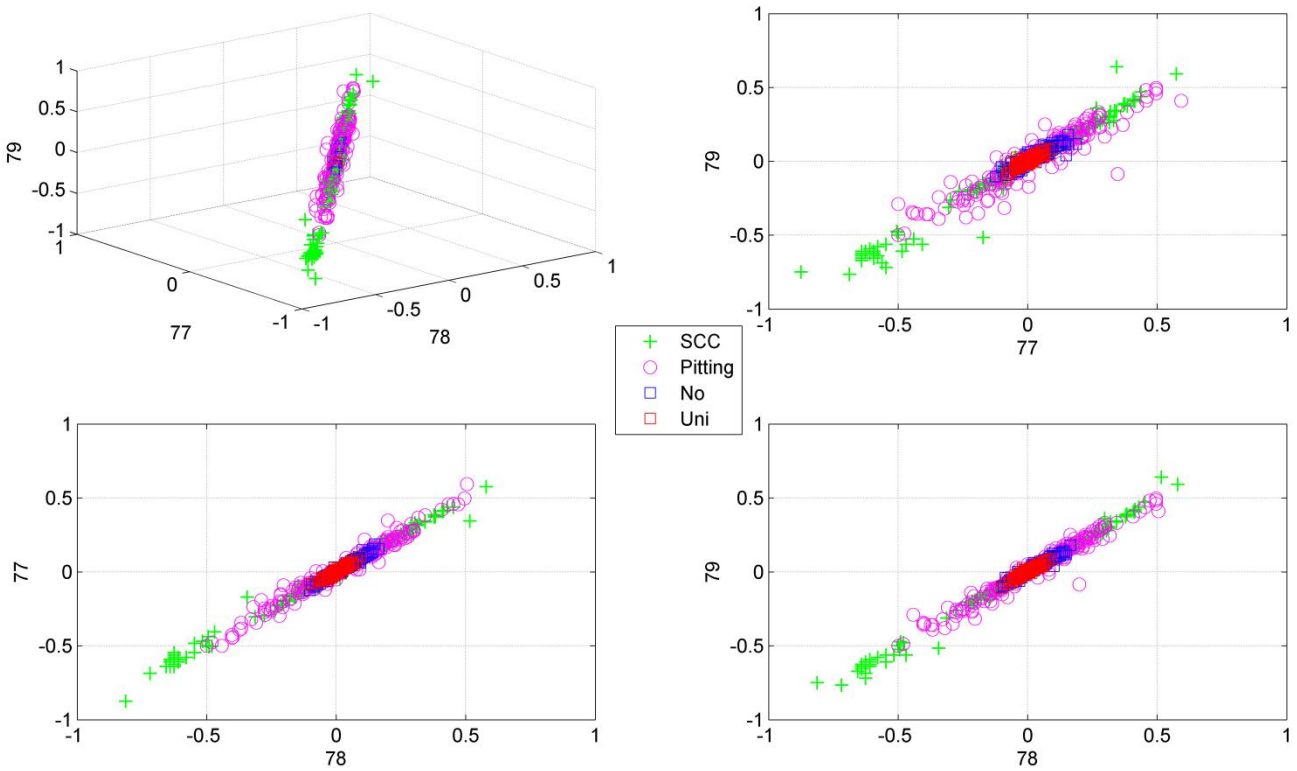
Figure 8. Evolution of the accuracy of naïve Bayes classifier as a function of the number of wavelet coefficients selected with the LDB algorithm and the mutual information filter algorithms. The horizontal line indicates the accuracy when all 1,024 samples are used.



Note the slower increase in accuracy for the LDB algorithm compared to the mutual information approaches that can be observed in Figures 5 to 8. This is related to the fact that the LDB algorithm ignores dependencies between the wavelet coefficients. In fact, the selected wavelet coefficients are

highly redundant. In each of the training folds of the 10 fold cross-validation, the local discriminant basis selection algorithm selected subspace $\mathbf{W}_0^0$ as the most discriminative basis. Although the coefficients in this subspace provide discriminative information between SCC (largest values), pitting (intermediate values) and uniform corrosion + absence of corrosion (these two classes have the smallest values), the LDB algorithm was misled by the high dependencies that are present in subspace $\mathbf{W}_0^0$. Indeed, in the scatter plot of Figure 9, it can be seen that the first three features, which occurred most often as a triplet in the 10 training sets of the 10 fold cross-validation, are in fact highly dependent. Each one of the three coefficients provides about the same discriminative power, so adding up their discriminative powers to obtain the joint discriminative power is misleading. The highest accuracy achieved with the LDB algorithm is obtained for the k-nearest neighbor classifier using 22 wavelet coefficients: 71.9%.

Figure 9. Scatter plots of the first 3 coefficients that were selected most often by the local discriminant basis algorithm (LDB) as a triplet in the 10 training sets of the 10 fold cross-validation. These are the coefficients $\gamma_{0,0,77}$, $\gamma_{0,0,78}$ and $\gamma_{0,0,79}$ in subspace $\mathbf{W}_0^0$. These scatter plots illustrate that the first three selected coefficients are highly redundant.



Comparison of Figures 5, 6, 7 and 8 reveals that the relevance-redundancy criterion for wavelet coefficient selection results in the highest classification accuracies. In fact, it is almost always better, no matter how many wavelet coefficients are selected, and no matter which classifier is chosen. The MI knn approach can be regarded as second best, because it is almost always better than the LDB algorithm and the MI Parzen approach in case of the decision tree, Gaussian mixture model and naïve Bayes classifiers. Note also that the performance of the relevance-redundancy approach is higher than the case when no feature subset selection (no FSS) is applied. Indeed, e.g., in Figure 8 the performance

of the relevance-redundancy approach is higher than the 'no FSS' approach as soon as two features have been selected. In case no feature subset selection is applied, the whole signal, *i.e.*, all 1,024 time samples for each signal, are used to train the classification algorithms and to perform the predictions. The observation that a subset of features may lead to higher classification accuracies compared to the whole signal can be related to the 'curse of dimensionality' [34]. A part of the explanation lies in the fact that when using more features, more parameters need to be estimated for the classification algorithms based on the same finite training sample size. These parameters can only be estimated with limited accuracy, and this in turn increases the classification error. Furthermore, when using all 1,024 time samples possibly many noisy samples are included which could corrupt the prediction accuracy. One of the purposes of feature selection is to select those features from which good predictions can be generated, and ignore the noisy ones.

The classification accuracies do not reveal the structure of the errors made in the identification of the corrosion types. Therefore, we computed the confusion matrix. We concentrate on the highest accuracy we could achieve: this is obtained in Figure 8 with the naïve Bayes classifier when 27 wavelet coefficients are used. The accuracy is equal to 86.4% which is obviously much higher than could be obtained with LDB algorithm (71.9%).

The columns in the confusion matrix shown in Table 2 correspond with the known corrosion types, the rows are the predicted corrosion types using the naïve Bayes classifier. The pitting column e.g., in Table 2, shows that of all 214 pitting signals, eight are identified wrongly as absence of corrosion, 199 are identified correctly as pitting and seven are identified wrongly as SCC. This leads to a high sensitivity for pitting: 199/(8 + 7 + 199) × 100% = 93.0%. SCC can also be identified with high sensitivity: of all 205 SCC signals, six are identified wrongly as absence of corrosion, one wrongly as pitting and 198 are identified correctly. This leads to a sensitivity for SCC equal to: 198/(198 + 1 + 6) × 100% = 96.6%. Absence of corrosion and uniform corrosion are more easily (mutually) confused: the sensitivity for absence of corrosion is 73.1% and for uniform corrosion 82.0%. Note that signals from absence of corrosion and uniform corrosion are both of continuous-type emission and that their signatures in Figure 2(a–d) are hard to distinguish. It is important to note that the most harmful types of corrosion can be identified accurately, whereas the confusion between absence of corrosion and uniform corrosion is less problematic.

Table 2. Confusion matrix for the naïve Bayes classifier using 27 wavelet coefficients. The numbers are obtained using all 10 test folds from the 10 fold cross-validation.

	Absence of corrosion	**Uniform corrosion**	**Pitting**	**Stress corrosion cracking (SCC)**
Absence of corrosion (predicted)	**144**	35	8	6
Uniform corrosion (predicted)	52	**159**	0	0
Pitting (predicted)	1	0	**199**	1
Stress corrosion cracking (predicted)	0	0	7	**198**

Finally, we note that the approach presented in this paper is generally applicable to acoustic events originating from different steel types. However, the resistance of steel towards a particular type of corrosion is influenced largely by its alloyed elements: chromium, manganese, molybdenum, nickel and nitrogen [10]. Hence, besides the acoustic activity also the steel type is indicative for the type of corrosion that is occurring. The steel type could be used as an additional discrete input variable that the machine learning algorithm can use to predict the corrosion type. Alternatively, one could use the chemical composition as an additional set of continuous input variables. However, the machine learning algorithm would require a large number of different steel types to be used in order to infer the corrosion type from the chemical composition together with the acoustic emission signals.

6. Conclusions

We have used the acoustic emission technique, a non-destructive testing technique, to identify different types of corrosion that occur most often in the chemical process industry. As stated in the introduction, one of the main progresses in corrosion prevention can be achieved with better information-processing strategies and the development of more efficient monitoring tools that support corrosion control programs [10]. A large progress in corrosion identification was achieved here by exploiting more advanced information-processing strategies. When the raw acoustic signals were used, the maximal accuracy achieved was rather disappointing: 70.7% (see Figure 7). A small improvement in accuracy, up to 71.9%, was achieved by using the local discriminant basis algorithm (LDB) when features are extracted with a wavelet packet decomposition. However, we noted that the LDB algorithm selected wavelet coefficients that may be highly redundant (see Figure 9). Mutual information allows us to exclude wavelet coefficients that are redundant, and this leads to a large improvement in accuracy: 86.4% using the normalized mutual information criterion and a naïve Bayes classifier. The largest confusion was observed between absence of corrosion and uniform corrosion. The most harmful corrosion types pitting and stress corrosion cracking could be indentified each with a very high sensitivity.

Acknowledgements

The authors are grateful to N. Saito, University of California, Davis, USA, for providing the local discriminant basis selection algorithm and to M. Winkelmans for providing data from corrosion experiments. We are also grateful to M. Wevers for offering the opportunity to work on the problem of corrosion identification using the acoustic emission technique. GVD is supported by the CREA Financing (CREA/07/027) program of the K.U.Leuven. MMVH is supported by research grants received from the Excellence Financing program (EF 2005), the Belgian Fund for Scientific Research—Flanders (G.0588.09), the Interuniversity Attraction Poles Programme—Belgian Science Policy (IUAP P6/054), the Flemish Regional Ministry of Education (Belgium) (GOA 10/019), and the European Commission (IST-2007-217077). This work used the HPC (high-performance computing infrastructure) of the K.U.Leuven.

References

1. Kane, R.D. A new approach to corrosion monitoring. *Chem. Eng.* **2007**, *114*, 34-41.
2. Winkelmans, M. Fusion of Non-Destructive Testing Techniques for Corrosion Monitoring in Chemical Process Plants. Ph.D. Thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2004.
3. Brongers, M.P.H.; Tubens, I. *Corrosion Cost and Preventive Strategies in the United States, Appendix V, Chemical, Petrochemical and Pharmaceutical*; U.S. Department of Transportation, Federal Highway Administration: Washington, DC, USA, 2001; Available online: http://www.corrosioncost.com/pdf/chem.pdf (accessed on 6 April 2011).
4. Wade, A.P. Acoustic emission: Is industry listening? *Chemometr. Intell. Lab. Syst.* **1990**, *8*, 305-310.
5. Soulsbury, K.A.; Wade, A.P.; Sibbald, D.B. A rules-based approach to classification of chemical acoustic emission signals. *Chemometr. Intell. Lab. Syst.* **1992**, *15*, 87-105.
6. Yuyama, S. Fundamental aspects of acoustic emission applications to the problems caused by corrosion. In *Corrosion Monitoring in Industrial Plants Using Non-Destructive Testing and Electrochemical Methods*; Moran, G.C., Labine, P., Eds.; American Society for Testing and Materials: Philadelphia, PA, USA, 1986; pp. 43-74.
7. Seah, K.H.W.; Lim, K.B.; Chew, C.H.; Teoh, S.H. The correlation of acoustic emission with the rate of corrosion. *Corros. Sci.* **1993**, *34*, 1707-1713.
8. Mazille, H.; Rothéa, R.; Tronel, C. An acoustic emission technique for monitoring pitting corrosion of austenitic stainless steels. *Corros. Sci.* **1995**, *37*, 1365-1375.
9. Shaikh, H.; Amirthalingam, R.; Anita, T.; Sivaibharasai, N.; Jaykumar, T.; Manohar, P.; Khatak, H.S. Evaluation of stress corrosion cracking phenomenon in an AISI type 316LN stainless steel using acoustic emission technique. *Corros. Sci.* **2007**, *49*, 740-765.
10. Roberge, P.R. *Handbook of Corrosion Engineering*; McGraw-Hill: New York, NY, USA, 2000; p. 11.
11. Huang, M.; Jiang, L.; Liaw, P.K.; Brooks, C.R.; Seeley, R.; Klarstorm, D.L. Using acoustic emission in fatigue and fracture materials research. *JOM* **1998**, *50*, 1-14.
12. Surgeon, M.; Wevers, M. Modal analysis of acoustic emission signals from CFRP laminates. *NDT E Int*. **1999**, *32*, 311-322.
13. Suzuki, H.; Kinjo, T.; Hayashi, Y.; Takemoto, M.; Ono, K. Wavelet transform of acoustic emission signals. *JAE* **1996**, *14*, 69-84.
14. Marec, A.; Thomas, J.-H.; El Guerjouma, R. Damage characterization of polymer-based composite materials: Multivariable analysis and wavelet transform for clustering acoustic emission data. *Mech. Syst. Signal Process.* **2008**, *22*, 1441-1464.
15. Van Dijck, G.; Wevers, M.; Van Hulle, M.M. Wavelet packet decomposition for the identification of corrosion type from acoustic emission signals. *Int. J. Wavelets Multiresolut. Inf. Process.* **2009**, *7*, 513-534.
16. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157-1182.
17. Wickerhauser, M.V. INRIA lectures on wavelet packet algorithms. In *Proceedings of Ondelettes et Paquets d'Ondes*; Roquencourt, France, 17–21 June 1991; pp. 31-99.

18. Coifman, R.R.; Wickerhauser, M.V. Entropy-based algorithms for best basis selection. *IEEE Trans. Inform. Theory* **1992**, *38*, 713-718.
19. Mallat, S. *A Wavelet Tour of Signal Processing*; Academic Press: New York, NY, USA, 1998.
20. Saito, N.; Coifman, R.R. Local discriminant bases and their applications. *J. Math. Imaging Vis.* **1995**, *5*, 337-358.
21. Saito, N.; Coifman, R.R. Geological information extraction from acoustic well-logging waveforms using time-frequency wavelets. *Geophysics* **1997**, *62*, 1921-1930.
22. Saito, N.; Coifman, R.R.; Geshwind, F.B.; Warner, F. Discriminant feature extraction using empirical probability density estimation and a local bases library. *Pattern Recogn.* **2002**, *35*, 2841-2852.
23. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2006.
24. Van Dijck, G.; Van Hulle, M.M. Increasing and decreasing returns and losses in mutual information feature subset selection. *Entropy* **2010**, *12*, 2144-2170.
25. Wegst, C.W. *Stahlschlüssel*, 19th ed.; Verlag Stahlschlüssel Wegst GmbH: Marbach, Germany, 2001.
26. Mallat, S. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674-693.
27. Mallat, S. *A Wavelet Tour of Signal Processing*, 2nd ed.; Academic Press: San Diego, CA, USA, 1999.
28. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273-324.
29. Kwak, N.; Choi, C.-H. Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1667-1671.
30. Van Dijck, G.; Van Hulle, M.M. Speeding up feature subset selection through mutual information relevance filtering. In *Proceedings of ECML/PKDD 2007*, Warsaw, Poland, 17–21 September 2007; pp. 277-287.
31. Kozachenko, L.F.; Leonenko, N.N. On statistical estimation of entropy of random vector. *Probl. Inform. Transm.* **1987**, *23*, 95-101.
32. Estévez, P.A.; Tesmer, M.; Perez, C.A.; Zurada, J.M. Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* **2009**, *20*, 189-201.
33. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *Int. J. Data Min. Bioinform.* **2005**, *3*, 185-205.
34. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; John Wiley & Sons: New York, NY, USA, 2001.
35. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I.H. The WEKA data mining software: An update. *SIGKDD Explorations* **2009**, *11*, 10-18.
36. McLachlan, G.; Peel, D. *Finite Mixture Models*; John Wiley & Sons: New York, NY, USA, 2000.
37. Mitchell, T. *Machine Learning*; McGraw Hill: New York, NY, USA, 1997; p. 112.