Equipment diagnostics based on comparison of past abnormal behaviors using a big data platform

Carole MAI [1], and Robin CHEVALIER[2]

[1,2] *EDF – R&D Recherche et Développement, Chatou Cedex, France*
carole.mai@edf.fr, robin.chevalier@supelec.fr

ABSTRACT

Nuclear monitoring systems provide an increasing amount of real-time measurements to analyse equipment condition. Storage, retrieval, and effective use of information must therefore be made as efficient as possible in order to achieve reliability goals and additional reductions in operating costs. Combining them with other data sources like equipment characteristics, maintenance logs and previous expertise reports will help make better diagnostics while minimizing the time dedicated to an analysis hence leading to better maintenance decisions. This paper presents an overview of the project under study at EDF to develop a case based reasoning platform enabling experts to efficiently retrieve similarities between past events and the current situation. Several technical barriers are encountered such as the comparison of multidimensional time series, of textual information and the extraction of signal features. The problem is here framed as a pattern classification problem where the classes correspond to equipment faults. Similarity criteria have been defined and evaluated against nuclear power plants data for the diagnosis of abnormal patterns on critical equipment. The classification results are compared with service and expertise reports using clustering and classification algorithms. The prospect of this diagnosis is to support the adjustment of maintenance schedule by estimating the remaining useful life of critical equipment.

1. INTRODUCTION

In the face of competition from other energy sources, research on nuclear power plants aims to increase the equipment reliability to ensure safety and security, while improving performance by reducing operating costs. The balance between these objectives is partly ensured by optimizing maintenance decisions. On the one hand, this optimization helps reduce outage duration by limiting equipment inspection or unnecessary replacement of equipment – required by the *calendar based maintenance* policy. On the other hand, a smarter scheduling of equipment replacement helps to avoid unplanned outage caused by safety monitoring alarms – consequence of the *reactive maintenance* policy.

To mitigate these issues, research on a third maintenance policy, called *condition based maintenance*, grows rapidly (Jardine, Lin, & Banjevic, 2006). It consists in, first, detecting incipient failures using monitoring systems and then, identifying the resulting maintenance tasks that should eventually be undertaken. This process can be broken down in four main tasks which are Fault detection, Fault diagnosis, Fault prognosis and Maintenance optimisation.

Besides the benefits of maintenance decisions optimization, the rapid technological progress of data management solutions opens up new prospects by making a better use of the available data – sensors measurements, equipment characteristics, maintenance logs, previous expertise, etc. These storage systems indeed remove some technical barriers by easing access to data, potentially heterogeneous in nature, and offering scope for further analyses by combining these data sources. These analyses help to extend the equipment knowledge which lead to potential applications on monitoring and maintenance optimization, by supporting the experts in their diagnostic, prognostic and maintenance schedule.

In this context, this study focuses on fault diagnosis – which can in turn support prognosis and maintenance decision. The diagnostic process of an incipient failure that occurs on an equipment can be compared to that realized by a doctor examining a patient. The first step of the process consists in the health assessment whose objective is to state precisely the symptoms of the potential equipment failure. For this purpose, the doctor combines information coming from the regular medical check-ups – respectively the sensors monitoring the state of the equipment in real-time – with other additional measurements driven by the suspected cause of the failure. The characterization of the study case through the identification of the symptoms constitutes the key part of the diagnostic process. This diagnostic will entirely rely on the measurements realized to assess the equipment status. These measurements, combined with information regarding the history of the equipment and previous maintenance operations, form the inputs of the diagnostic process.

For the diagnostic task, two approaches can be taken, alternatively or jointly. The first approach is to use the experts' knowledge of monitored equipment – which relies on different sources, namely the manufacturer's information, physical or thermodynamical models, simulations, etc. The second one uses the operating experience by comparing the current situation against previous abnormalities – which occurred on similar equipment.

The proposed approach for diagnostic is to build a structured database of previous abnormalities. Granting this database with a metric, the comparison of situations will follow the case-based reasoning approach. This approach will be defined and evaluated against nuclear power plant equipment. Statistical methods will help to refine the diagnostic and experts' knowledge of equipment

by adjusting the model – both the structure and the metric. Therefore, the problem is here framed as a pattern classification problem in which adapting the model reduces the misclassification and therefore leads to a better diagnostic.

This paper is organized as follows. Section 2 introduces the diagnostic process and some background while Section 3 presents the problem for a chosen EDF application case and states the approach followed. The building up of the model will be developed in Section 4 followed by its evaluation and refinement in Section 5. Drawbacks and perspectives of the research will be discussed in the conclusion.

2. BACKGROUND

Monitoring systems provide a huge amount of physical measurements regarding the equipment status and its components. Thanks to the extensive instrumentation of equipment, in addition to safety alarms, fault detection methods are implemented for early detection of abnormal behaviors. These detection methods use the sensors time series to detect a faulty behavior. Potential abnormalities are then analyzed by experts to assess the severity of the potential damage and subsequently needed maintenance operations. Case-Based reasoning methods assist experts in finding the cause behind the current situation, and providing guidance on maintenance decision-making and planning.

2.1. Fault Detection

A fault is defined as a 'condition of a machine that occurs when one of its components degrades or exhibits abnormal behavior, which may lead to the failure of the machine' (ISO 13372:2012(E/F), 2012). Monitoring systems provide real time information to operators about the current equipment status.

The fault detection task has already been thoroughly investigated and tools are deployed at EDF for early detection of abnormal behavior on any monitored equipment (refer to statistical methods as the one presented in (Todorov, Feller, & Chevalier, 2015)). One algorithm in use on monitored systems is based on a clustering approach applied to the sensors measurements (ECM: Evolving Clustering Method (Song, 2001)). This supervised method distinguishes between the normal and the abnormal behaviors of multidimensional time series. When the residual – relative deviation of signals from their referenced normal behavior – reaches a threshold, warnings are triggered which alerts the operators of a potential fault detection, as illustrated by the Figure 1.

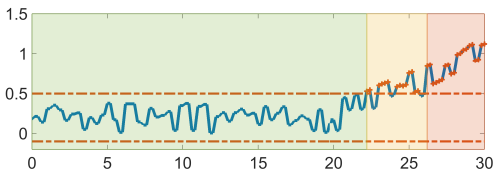


Figure 1. Global residual evolution which triggers the warning signal

Note 1: This figure illustrate the evolution of a global residual signal. Three areas are presented. The first one (in green) corresponds to a normal behavior. In the second area (in yellow), few points are in error. The persistence of these errors leads, in the third area (in red), to trigger the warning signal.

Following this detection, every anomaly must be extracted and analyzed by experts to diagnose the encountered situation. Information used for the diagnostic are monitoring data and additional information on the equipment. Regarding the monitoring data, this extraction concerns only sequences of a subset of sensors, which is considered relevant for the monitored equipment.

2.2. Case-Based Reasoning

Once a potential failure is detected, it has to be diagnosed. Current internal research works tend to refine this diagnostic by extending the equipment knowledge in order to support experts' decisions.

The proposed method is to follow the Case-Based Reasoning (CBR) approach to determine the cause of the current faulty behavior. This method aims to diagnose a situation by taking advantage of past experience as detailed in (Aamodt & Plaza, 1994). This approach mimics the diagnostic process manually realized by the experts. It could enrich the current equipment knowledge by being used in addition to other methods already in place such as model-based or data driven methods.

Facing a potential abnormal behavior of an equipment, the purpose of this approach is to support experts in their diagnosis by finding the most similar situation that occurred in the past.

Actually, CBR is a broader approach that includes four steps:

- Retrieve: retrieve the closest previous situation(s) stored in the case-base. A case consists of all essential data regarding the event, including the event characteristics, associated expertises and the proposed solution;
- Reuse: analyze the closest case or cases in order to make a diagnosis and adapt their solution to the current situation;
- Revise: test the solution on the current situation and, if necessary, revise;
- Retain: store the resulting experience as a new case in the case-base.

Each step is directly part of the diagnostic process: comparing a new potential equipment failure to previous situations, using the most similar referenced situations to diagnose the current one, adapting the maintenance and corrective actions and finally storing the newly diagnosed case to enrich the case-base.

Thus, the CBR process is a way of structuring the past experience regarding abnormal situations. The underlying difficulties are with the definition of:

1. A standardized case structure;
2. A similarity distance between two cases.

These two issues are fundamentally related. They both need to be tackled during the model construction, and will play a significant role on the diagnostic results.

2.3. Synthesis

To sum-up, monitoring data are analyzed in real-time for early detection of abnormal behaviors. When a faulty behavior occurs, it is transmitted to the CBR platform. This system extracts essential information needed to structure the current case according to the standardized structure. Then, this faulty behavior is compared to past events already stored in the database and find the closest cases. Additional information stored in the case-base, concerning the closest cases, helps to identify the cause of the current

situation, the maintenance or corrective actions, and assess the Remaining-Useful Life (RUL) of the monitored equipment, i.e. assess the remaining time before an alarm, a failure or an outage are likely to occur (refer to Figure 2).

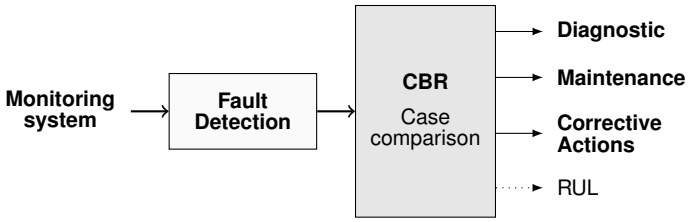


Figure 2. Diagnostic process using CBR

Finally, the newly diagnosed case is stored in the case-base to refine the model. This storage includes first the current case structure, secondly all the helpful information needed in face of a new situation, such as expertises, maintenance, evolution and more generally all information, decisions and actions related to this failure.

3. PROBLEM STATEMENT

In this paper, the CBR approach has been defined and evaluated for nuclear power plant equipment for the diagnosis of abnormal patterns.

3.1. Case-study

EDF is a French electric utility which operates a diverse portfolio of 140+ gigawatts of generation capacity both inside and outside France. Its 58 nuclear reactors represent 54% of its generation capacity.

In this study, the monitored equipment is the reactor coolant pump (called RCP), more specifically the seals #1, ensuring the thermal barrier to the reactor coolant system (Figure 3). This critical equipment plays an essential role in nuclear power plants for both safety and efficiency reasons. Early fault detection tools warn the operators when a potential fault occurs while safety alarms lead to unplanned outages. In this study, the early warning signals trigger the CBR process.

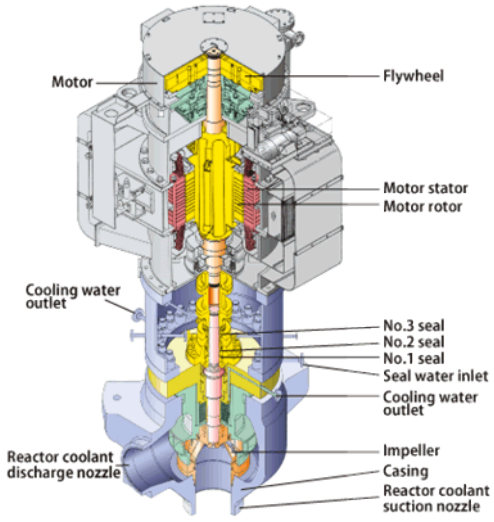


Figure 3. Reactor coolant pump [www.mhi-global.com]

The pumps seals ensure the sealing of the moto-pump at the interface between moving shaft and fixed part. The seal #1 is a dynamic seal with a controlled leakage flow. It is subject to different degradations that affect its behavior. These potential degradations depend on many influencing factors such as the technology and the endured stress.

The evolution of the seal degradation is, among others, monitored by flow, pressure and temperature sensors. Experts' knowledge of equipment helps to select the relevant features for fault diagnostics.

As a result, data needed for the diagnostic are both a selected subset of monitoring data and influencing factors – additional information concerning the equipment. Combining these data sources enriches the knowledge on the situation under study.

3.2. Available cases

A database of past events, that occurred in the last 15 years on the RCPs, has been set up using expertise reports and histories of monitored data but it remains unstructured. It includes all expertises associated to the different situations and the involved monitoring data. This database contains less than a hundred cases – an amount which is not significant enough to efficiently use statistical methods. Artificial data will therefore be generated for the refinement of the model.

Originally, this case-base was not structured enough for the case comparison. Cases encompass heterogeneous data in nature such as:

- Time series coming from monitoring data;
- Equipment one-off measurements;
- Pictures, equipment sketches, etc.;
- Texts from maintenance decisions, equipment characteristics, etc.

To ease the case comparison, a standardized case structure has been established with the support of experts.

This structure consists in a fair view of a given situation, preserving its fundamental characteristics. It encompasses all essential information for the case comparison, based on both influencing factors and monitoring data, such as the measurements time series, feature extracted from them, the equipment characteristics and the warning signals. Only these case structures will be used during the case comparison.

3.3. Protocol

To sum-up, the proposed approach for diagnostic automation is to properly establish a model defined by state variables – the case structure – and a system metric – a similarity measure. This model will be used at the end to diagnose a new potential failure through a classification method. Initially established by experts, it will be refined using both real and artificial data with clustering and optimization methods.

This protocol is summed up in the following schema (Figure 4).

4. MODEL CONSTRUCTION

As previously stated, available data are heterogeneous. The model construction first consists in identifying essential information to

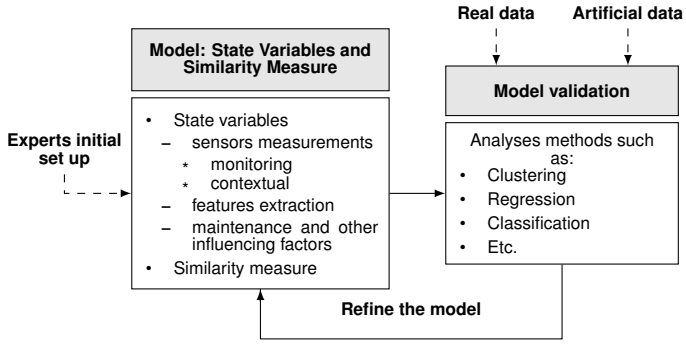


Figure 4. General protocol

build a standardized case structure. State variables are then automatically extracted from all registered situations to obtain a structured case-base. Parallel to this case structure, the second aspect of the model construction is to define a similarity measure for the case comparison.

4.1. State variables

State variables are computed using both monitoring data and influencing factors.

4.1.1. Preprocessing

The variables included in the case structure are numerical or textual. Regarding the textual data, manual or semi-automatic preprocessing has been applied to standardize available information, correct the eventual mistakes and fill the missing slots. This task of reconciling data has been achieved with the expert' support. The resulting meta-data can be sorted in two categories which are the machine identity (e.g. power plant, pump, technologies of the seal) and the fault context (e.g. dates, operating regime, number of pumps affected).

Regarding the numerical data, several preprocessing tasks have been achieved:

1. **Data cleaning**: removing all abnormal values by filtering signals according to a quality indicator;
2. **Data selection**: filtering according to the operating states using the reactor fluid pressure and thermal power sensors;
3. **Data sub-sampling and compression** (applied to time series): reducing the volume in order to speed up the processing and focus on long-term evolutions;
4. **Data smoothing** (applied to time series): reduce the noise and its influence on the subsequent comparisons;
5. **Features extraction** (applied to time series (see 4.1.2)): the selection of features has been tackled by the experts, and their extraction is automated;
6. **Normalization & Synchronizing** (applied to time series (see 4.2.3)): this preprocessing helps in the time series comparison.

This fifth step – Features extraction – heavily relies on the application and requires particular attention when designing it. Regarding the seal #1 application case, the most impacted signal when a fault occurs is the seal leakage flow measurement. Its evolution can be classified in four categories which are:

- **Stable**: healthy case, where the leakage flow remains almost constant;
- **Stable irregular**: abnormal case, the leakage flow fluctuates rapidly in a limited range;
- **Upward drift**: abnormal case, large rise of the leakage flow for a period of several weeks to several months, potentially with steps;
- **Downward drift**: abnormal case, large drop of the leakage flow for a period of several weeks to several months, potentially with steps.

4.1.2. Case structure

The resulting structure is close to the following (Table 1) – confidential information being concealed.

Table 1. Standardized case structure for the EDF application case

Machine identity			
Plant	Plant n°1	Seal type	techno α
Affected pump	3	Glass technology	techno β
Plant type	1300	Sliding ring techno	techno γ
Context			
Start Date	15/11/1997	Operating regime	Nominal Level
End Date	22/01/1998	Nb pump concerned	2
Alarm Level LF	Yes	Alarm Level LT	No
Alarm residual LF	Yes	Alarm Residual LT	No
Time series			
Leak flow	Leak temperature	Injection flow	Injection temp
Pressure	Power		
Influence parameters			
Injection flow-IF	Drop, normal level	Injection Temp IT	Stable, high level
Correlation LF-IF	-92%	Correlation LF-IT	70%
Correlation LT-IF	-89%	Correlation LT-IT	93%
Linear Reg LF=f(IF)	(-0.54, 1706, 0.85)	Linear Reg LF=f(IF)	(1.25, -5.6, 0.87)
Identified symptoms			
Evolution LFJ1	Upwart drift	Drift Amount	XXX L/h
Mean level	XXX L/h	Noise	15 L/h (std)
Evolution speed	Fast: 63.9 days	Step(s)	No
Evolution LTJ1	Stable	Standard Deviation	1.3°C
Mean level	55.9°C		
Additional information			
Validated Fault	BDGJ1 drilled	Suspected Fault	BDG/DLI
File references	0e621293-8d49-4e31-8c1c-99fe14f6016a		

The textual data are used to fill the two first categories of the structure – *Machine identity* and *Context*. The model includes time series coming from the monitoring system: leak flow, leak temperature, injection flow, injection temperature, pressure and speed. Then, features are extracted from these measurements time series to fill the *Influence parameters* and the *Identified symptoms* categories. Finally, the last section provides some additional information and links to related documents about the corrective actions, diagnostic, etc. This last category is only used once the closest situation is found.

Features extracted from the time series consist in: correlations between interesting signals, regression parameters and evolution indicators. These computed indicators describe the global evolution, the drift amount, the mean level, the noise, the detection of step, etc. As previously mentioned, the selection of the interesting features extracted is governed by the experienced abnormalities.

4.2. Similarity measure

The similarity measure used to compare two situations must be designed to deal with heterogeneous data. To ease the study, this global similarity measure has been set up as a weighted measure – sum of similarity measures. These weights have been initialized by experts and need to be adjusted.

To avoid initial predominance of a factor on another, every local similarity measure s_i of the i^{th} factor F_i, has been set in the interval $[0, 1]$ such that:

$$\begin{array}{rccc} s_i: & F_i \times F_i & \rightarrow & [0,1] \\ & (x,y) & \mapsto & s_i(x,y) \end{array} \tag{1}$$

4.2.1. Categorical data

The key characteristic of categorical data is that values are not inherently ordered. The categorical similarity measure is applied to textual data and some numerical information that cannot necessarily be meaningfully ordered such as the plant type. A wide variety of similarity measures of categorical data exists but in this study, only the simplest one has been used: the overlap measure (Stanfill & Waltz, 1986).

This similarity measure s_i between two elements x and y belonging to the i^{th} factor F_i is defined by:

$$\forall (x,y) \in F_i \times F_i, \quad s_i(x,y) = \begin{cases} 1, & \text{if } x = y \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Thanks to the reconciliation of textual data, the different categories are standardized which eases the construction of a categorical similarity measure.

4.2.2. One-off numerical data

The simplest way to make the similarity measure between one-off numerical data take values in $[0, 1]$ is to normalize a distance. This definition requires the knowledge of the limits of this distance measure that should be mapped to the similarity measure; in addition, this resulting similarity is sensitive to outliers causing distortions. The solution adopted here is to introduce saturation effects (Lesot, Rifqi, & Benhadda, 2008). Given d a distance measure such as a Minkowski distance, this similarity measure s_i^M between two elements x and y belonging to the i^{th} factor F_i and based on the saturation limit M is defined by:

$$\forall (x,y) \in F_i \times F_i, \quad s_i^M(x,y) = max\left(\frac{M - d(x,y)}{M}, 0\right) \tag{3}$$

M is a fixed parameter chosen according to the range of the compared signals. This parameter is an addition to the set of weights in the refinement.

4.2.3. Time series

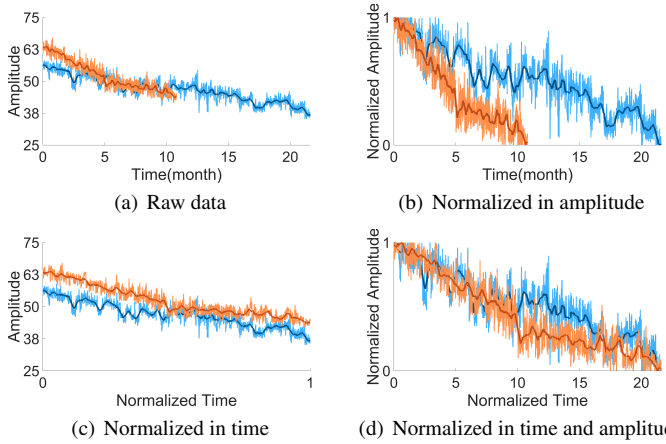
The comparison of time series has been applied to multidimensional time series – the leakage flow, the leakage temperature, the injection flow and the injection temperature – but the computation is expensive. These comparisons have been made on both normal and residual data, using alternatively normalized and unnormalized signals in time and amplitude. The similarity measures adopted for the time series comparison are based on the Euclidean and the Dynamic Time Warping distances (Wang et al., 2012).

1. *Residual data*

The residual data comes from the fault detection model already deployed. Their time series correspond to the gap between the expected behavior and the real behavior. In a first approximation, these residual data can be computed by subtracting the mean value just before the fault detection.

2. *Normalization*

The following figure (Figure 5) illustrates the normalizations in time and amplitude. Initially, the two time series do not necessarily have the same length nor the same range. Although the resulting curves seem very close when normalizing in both amplitude and time, the physical meaning of these normalization is uncertain. To avoid arbitrary eliminating some of them, these four comparison methods were used in the global similarity measure with four respective weights. The weights adjustment will reduce the influence, if not disregard, the less meaningful measures. Note that for the multidimensional time series comparison, these normalizations are applied to each time series involved in the comparison.



(a) Raw data (b) Normalized in amplitude

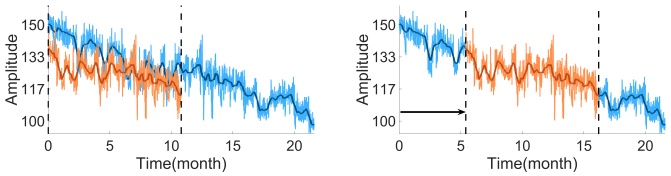
(c) Normalized in time (d) Normalized in time and amplitude

Figure 5. Time series normalization for the comparison

Note 5: This figure illustrated the additional preprocessing methods applied for the comparison here on leak flow data (normalizations). Two time series are illustrated, with each time both raw data and preprocessed time series (cleaned, filtered, compressed, smoothed). Note that the amplitude values have been modified in order to respect the confidentiality. Note 5(a): Raw data with their respective preprocessed time series represented by the smoothed curves. Note 5(b): Signals normalized in amplitude – $y_{N_A} = \dfrac{y - y_{min}}{y_{max} - y_{min}}$. Note 5(c): Signals normalized in time – unchanged for $y_{N_T}^{(1)}$ while $\forall t, y_{N_T}^{(2)}\left(t\dfrac{\Delta t_1}{\Delta t_2}\right) = y^{(2)}(t)$. Note 5(d): Signals normalized in both time and amplitude.

3. *Synchronization*

In the case when signals do not have the same length, synchronization help in the comparison. The objective is to find the best matching position. For this purpose, the method used here is a sliding window with the Euclidean distance to find the position corresponding to the smallest residual. Again, this synchronization has been made on all four signals involved in the comparison at the same time to find the best global position. The following figures illustrate this process for a 1D synchronization (Figure 6).



(a) Raw data (b) Synchronized time series

Figure 6. Time series synchronization

4. *Distance measure*

For the comparison of time series, the similarity measure has been created based on a distance measure mapped to the adequate range, by using normalization and saturation. The distance measures applied in this study are the Euclidean distance and the Dynamic Time Warping. The first one is a lock-step distance while the second one is elastic. By including both of them with a proper weight in the global similarity measure, the adjustment of the set of weights will provide a assessment of their impact.

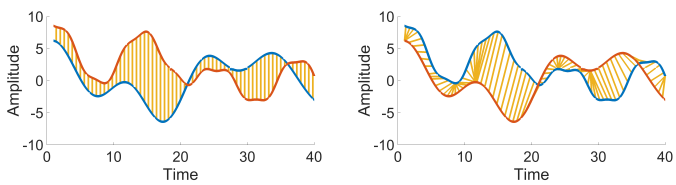
The Euclidean is a particular Minkowski distance, also called L^2. The distance between two signals x and y of the same length n is given by:

$$\forall(x,y), \quad d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

The Dynamic Time Warping is an elastic measure based on a lock-step distance. The resulting distance is the solution to a constrained optimization problem that dynamically computes the optimal match between two sequences. In this study, the deformation has been curbed to maintain the integrity of the subsequences. When comparing $x_{(1..N)}$to $y_{(1..M)}$, calling $p_{(1..L)}$ the followed path vector of (N, M)-warping path – i.e. $\forall l, p_l = (n_l, m_l)$ where n and m are the index of the path followed – the constraints imposed are the following (Senin, 2008):

- Boundary condition: $p_1 = (1,1)$ and $p_l = (N,M)$
- Monotonicity condition: $n_1 \leq n_2 \leq ... \leq n_L$ and $m_1 \leq m_2 \leq ... \leq m_L$
- Step size condition: $p_{l+1} - p_l \in \{(1,0),(0,1),(1,1)\} \forall l \in [\![1..L-1]\!]$
- Deformation limitation condition: $\exists s$ such that $|n_l - m_l| < s(l)$.

The following figure (Figure 7) illustrates these two kinds of comparisons on artificial signals:



(a) Euclidean distance (lock-step) (b) DTW (Elastic)

Figure 7. Euclidean and DTW illustrations

Note 7: This figure illustrate the lock-step and elastic measure where time series have equal length. The elastic distance allows space-time distortions.

Unlike the Euclidean distance, the DTW allows the comparison of sequences of different lengths. Nevertheless, in this study, every distance computation has been operated on subsequences of equal length. The final distance result is then divided by this subsequence length in order to obtain comparable results.

5. *Penalty*

For original sequences which are not of the same length after the preprocessing, a penalty has been set. This cost is applied when the synchronization lead to compare a subsequence of length l_{sub} less than both of the original sequences length (l_1, l_2). This cost parameter penalizes additional points according to the following rule:

$$cost = \alpha_{cost} \left(min\left(l_1, l_2\right) - l_{sub}\right)^2 \tag{5}$$

A cost parameter α_{cost} has been introduced. This parameter is an addition to others in the refinement.

6. *Time series similarity measure*

As a result, the time series similarity measure applied in this study is a combination of weighted sub-similarities. The resulting distances (d_i) are mapped to similarity measures (s_i) by applying normalization and saturation (M). Thus, the similarity measure s_i^M between two elements x and y belonging to the i^{th} factor F_i and based on the limit M is defined by:

$$\forall(x,y), \quad s_i(x,y) = max\left(\frac{M - d_i(x,y) + cost}{M}, 0\right) \tag{6}$$

As with the one-off measure, the saturation parameter has to be adjusted during the refinement.

Note that the multidimensional signals comparison has been carried out by defining similarity measures for, on the one hand, computed data and features extraction and on the other hand, preprocessed time series.

4.2.4. Global similarity measure

In the global comparison, some categorical factors may be set to be exclusive such as the operating regime. The resulting similarity measure between two cases c_1 and c_2 composed in factors $(x_1, .., x_n)$ and $(y_1, .., y_n)$ respectively is defined by:

$$S(c_1,c_2) = \frac{\prod_{i \in F_i^{(C_e)}} s^{C_e}(x,y)}{W} \left(\sum_{i \in F^{(C)}} \alpha_i^C s^C(x,y) + \sum_{i \in F^{(O)}} \alpha_i^O s_i^O(x,y) + \sum_{i \in F^{(TS)}} \alpha_i^{TS} s_i^{TS}(x,y) \right) \tag{7}$$

where:

$F_i^{(C_e)}$	is the i^{th} factor associated to exclusive categorical measurements;
α_i^C	is the categorical measurement weight associated to factor $F_i^{(C)}$;
α_i^O	is the one-off measurement weight associated to factor $F_i^{(O)}$;
α_i^{TS}	is the time series weight associated to factor $F_i^{(TS)}$;
$s_i^{C_e}$	is the local similarity measure of exclusive categorical data;
s_i^C	is the local similarity measure of categorical data;
s_i^O	is the local similarity measure of one-off data;
s_i^{TS}	is the local similarity measure of time series;
W	is the total weight $W = \sum_{i \in F_C} \alpha_i^C + \sum_{i \in F_O} \alpha_i^O + \sum_{i \in F_{TS}} \alpha_i^{TS}$;

Note that some local similarity measures depends on the factor (i.e. on type of measure compared) since they have been obtained using a saturation parameter which relies on the range of the signals compared.

5. VALIDATION AND REFINEMENT

First, a set of weights defined with the expertise support will be evaluated against the 100 fault cases of the case-base. Then, a second approach aims to optimize these weights using a genetic algorithm on artificial data.

5.1. Validation of the model defined with experts

For the validation of the set of weights adjusted with experts, the resulting similarity measure has been compared to an arbitrary one. For this purpose, both similarity measures have been used to cluster the 100 referenced fault cases. The set of clusters obtained with each measure will be compared with the case categories as assessed by the experts. The method used for the clustering is the agglomerative hierarchical clustering tree based on the Ward distance between clusters since it provides multiple levels of granularity (Everitt, Landau, Leese, & Stahl, 2011).

Ideally, these case categories would correspond to the diagnosed fault. Unfortunately, due to missing or incomplete information regarding the seal #1 faults, this evaluation has been limited to the four generic leak flow categories.

5.1.1. Set of weights established with the experts

The set of weights has been established according to the following rules:

- The operating regime of the cases and their seal type are set to be exclusive criteria – if they differ, the similarity is set to zero;
- The most important categories of the case structure are the *Identified symptoms*, followed by the *Time series* (relation of two-thirds):
  - For the *Identified symptoms* category, the greatest weights are set for the drift amount, its duration and the signal stability;
  - For the *Time series* category, the preponderant weight is set for the unnormalized comparison, followed at a distance by the amplitude and time normalization. As suggested, the combination of normalizations is more or less disregarded.

5.1.2. Evaluation of the clustering results

The evaluation has been realized by comparing the results obtained, regarding the leak flow, using an arbitrary set of weights to those using the refined set. The resulting clustering trees are presented on dendograms showing the dissimilarity measures of the clusters at multiple levels of granularity (Figures 8 and 9):

1. *Using an arbitrary set of weights (Figure 8)*:
Judging only from the leak flow generic categories, many cases are not well classified. The elements under the mention 'cases not recognized' are a mix of cases of all categories including upward drift and downward drift.

2. *Using the experts' support to adjust the weights (Figure 9)*:
By adjusting the weights, the classification is significantly improved. On this dendrogram of the complete tree, the first differentiation is on the operating regime between nominal transient on the left hand side and operation state on the right hand side. Then, in the nominal operation state category, the distinction between downward drift, stable and irregular and upward drift is well established.

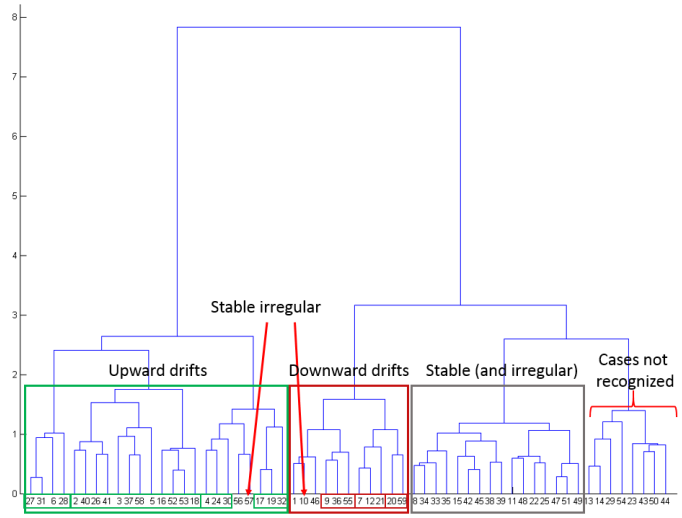


Figure 8. Clustering tree with an arbitrary set of weights

Note 8: y-axis represents dissimilarity measure and x-axis shows case number rearranged.

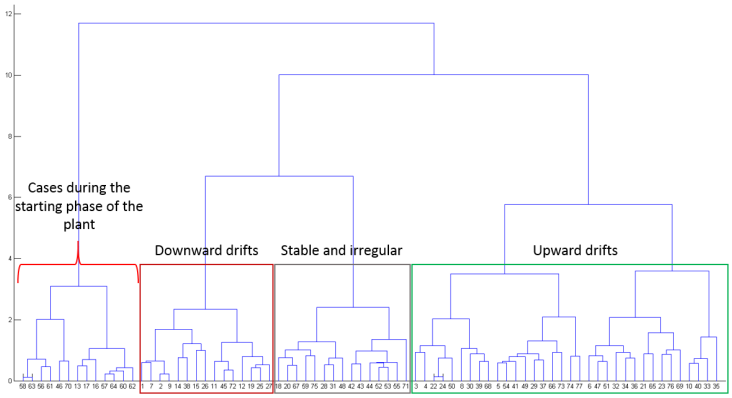


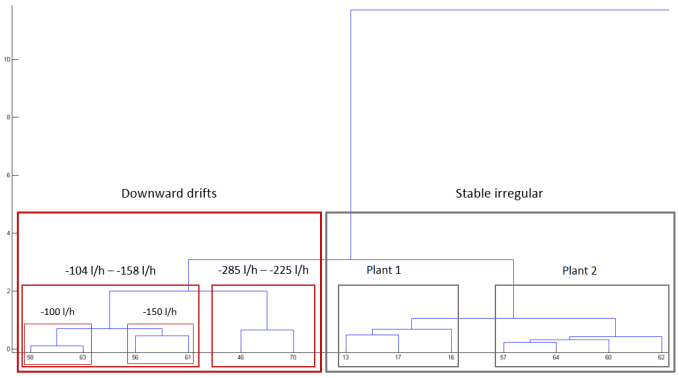
Figure 9. Clustering tree: adjusting the weights with experts' support

Note 9: y-axis represents dissimilarity measure and x-axis shows case number rearranged.

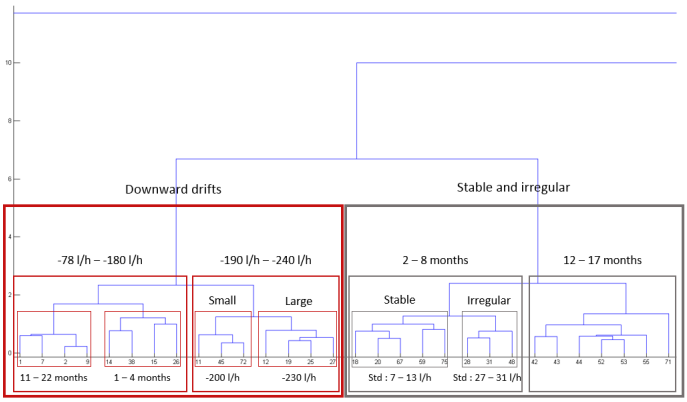
Looking further, finer categories can be pointed out.

- Regarding the transient cases (Figure 10(a)), on the right branch stand the downward drifts and on the other branch the stable irregular behaviors. Further differentiations are made on the drift amount and on the plant;
- Regarding the downward drift and stable and irregular (Figure 10(b)), distinctions are made on the drift amount and duration and on the stability;
- Focusing on the upward drift (Figure 10(c)), cases are clustered regarding the drift amount and its duration.

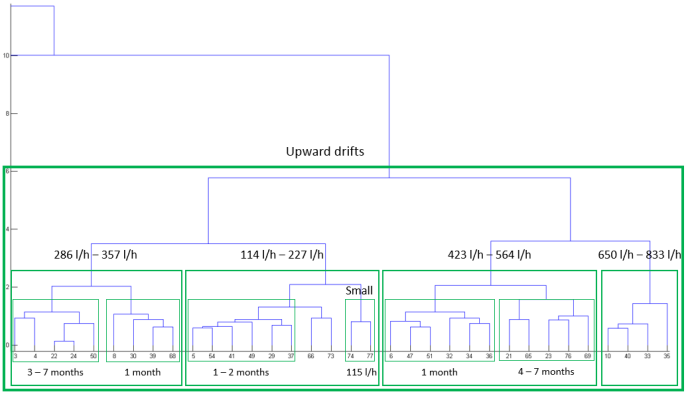
These finer clusters could be related to diagnosed faults. As previously suggested, a way to improve this evaluation is to extend its scope by judging directly from the observed degradations. The prerequisite of this is to clean up the database by filling all missing expertises and observed degradations.



(a) Transient Cases



(b) Downward Drift and Stable and Irregular cases



(c) Upward Drift

Figure 10. Classification tree details

Note 10: y-axis represents dissimilarity measure and x-axis shows case number rearranged.

This evaluation is really positive in the sense that both the four generic leak flow categories and the finer details are well clustered when using the similarity measure defined with the support of experts.

5.2. Refinement of the measure using artificial data

Artificial data helps to extend the case base and avoid over-fitting in statistical methods. In this study, the two main advantages when dealing with artificial data are:

- The amount of data is flexible;
- The degree of similarity between cases is known in advance, by construction.

These advantages are possible only because artificial data correctly replicate the equipment behavior.

5.2.1. Artificial data

Generated artificial data need to mimic the global interactions within the equipment. Such a construction requires first a good knowledge of each sub-equipment and materials of the monitored equipment, and secondly a good understanding of all their interactions.

Unfortunately, these interactions are very complex and hard to model – in particular the relations between a given degradation and the related measurements. Therefore, in this study, the generated artificial data are limited to the leak flow. The objective is then to adjust the parameters of the global similarity measure to find their relative influence on the case comparison.

In this context, the remaining descriptors included in the artificial case structure are the leak flow preprocessed time series and the *Identified symptoms* category (refer to Table 1: evolution, mean level, step, noise, drift amount, speed).

Several parameters are introduced to generate the artificial signals. They ensure these signals are as close as possible to real leak flow measurements:

- Amount of drift: from -200L/h to 600L/h;
- Duration: from 1 month to 1 year;
- Mean level: low (300L/h), normal (450L/h) or high (700L/h);
- Steps: with or without steps;
- Noise: 20 to 40L/h of amplitude.

5.2.2. Problem statement

The problem is here framed as a classification problem. Categories are identified by construction and the objective is to find the most adapted set of parameters involved in the similarity measure that replicates these predefined clusters.

When defining a global similarity measure between cases, several parameters have been defined: a set of weights and additional variables introduced in the definition of local similarities (refer to 4.2). The resulting problem is a constrained non-linear optimization problem. The aim is to minimize the misclassification by refining the global similarity measure.

For this purpose, during the generation of artificial cases, classes have been established. Two parameters are considered to have a lower influence on the comparison: the noise amount and the presence of steps. Generated data includes pairs of identical signals except for their noise; these paired cases are considered closest to each other. Three levels of step and two level of noise lead to six similar signals. As a result, by construction, each signal is classified in a two-element category – two noise amounts – which is itself part of a much larger six-element category – including different level of steps. A sample of these artificial data is illustrated Figure 11.

As a result, this construction eases the identification of the closest or the five closest cases to a given situation. However, its comparison with other classes is not well-established. This lead to restrict the objective functions to the followings:

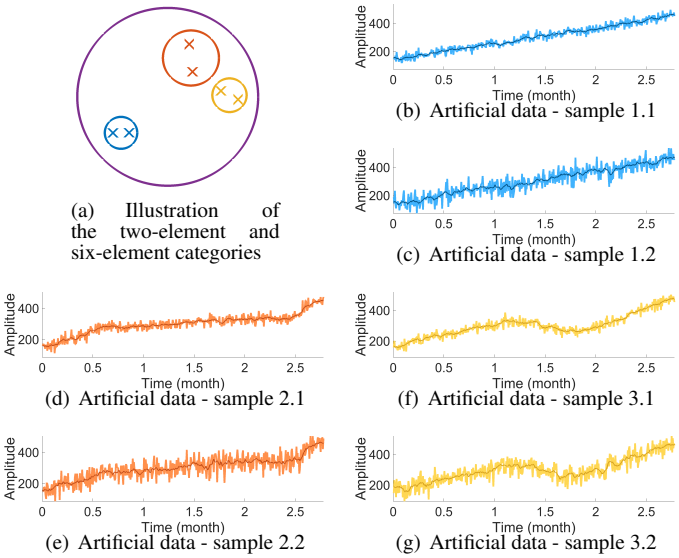


Figure 11. Sample of artificial leak flow time series

Note 11: These figures show a sample of the generated artificial data. This example illustrates the six-element class associated to a given set of identical parameters except noise and steps.

Function 1: Signals are associated with their closest.

Function 2: The five closest cases are correct.

These two functions will be used in the optimization.

5.2.3. Optimization using the genetic algorithm

A Genetic Algorithm (GA) is a search heuristic that is used to solve optimization problems according to (Mitchell, 1996). Usually starting from a population of randomly generated individuals, it iteratively adjusts the parameters to approach the optimum. In this study, this algorithm terminates when a maximum number of iterations has been reached.

The considered population is the set of weights and additional parameters introduced in the definition of local similarity measures. Due to these additional parameters, the resulting optimization problem is non-linear and amount to 18 parameters.

Several approaches have been tested:

1. *Applying the GA on all parameters*:
The first idea was to apply the GA directly on the 18 parameters. The results are not satisfactory: the error rate is up to 30% on some sets of artificial data. Some potential reasons have been raised:

- The 18-parameters optimization problem may be over-fitted;
- Some of these parameters are linked which affects the optimization results. Different combinations of these parameters lead to the same optimal value.

To improve the optimization, the GA has been applied to a subset of parameters: Identified symptoms, DTW and Euclidean distances with raw and residual signals. Then, fixing these local weights, the GA will be applied on the global system to optimize the remaining weights.

2. *Applying the GA on the Symptoms parameters*:
Multiple sets of artificial data have been used with both objective functions alternatively – one or five closest neighbors. The average resulting set of parameters is the following (Table 2).

Table 2. Symptoms parameters

Weights

Step	1.5	Drift	125	Noise	1.1
Level	88	Stability	34	Duration	69

Max

Drift (L/h)	466	Noise (L/h)	74	Duration (Months)	3.6
Level (L/h)	255	Stability (L/h)	53		

Note Table 2: The weights amount for the importance of each factor in the global similarity measure (No units). The max are defined in the local similarity measures for the saturation effect.

The error rate remains high, up to 25% on some artificial data sets. The next section will show that the global weights will provide a finer model and consequently lead to better results.
The maximum values mostly depend on the range of the factors compared. The weights explain the relative impact of the symptoms on the case comparison. Judging from these results, these factors are ordered as follows: drift, level, duration, stability and then, at a distance, step and noise. These are not inconsistent with the earlier expert adjustment.

3. *Applying the GA on the Time series parameters*:
In order to adjust the parameters involved in the time series comparisons, the GA has first been applied on each time series similarity measure independently: raw data or residual data, with time and/or amplitude normalization. This optimization helps to find the two parameters related to local similarities, namely cost and limit parameters. The results are close for each similarity: around 400 for the cost and 8000 for the limit. These parameters will be fixed in the following optimizations.

The second step of the process is to apply the GA to adjust the different weights, local parameters – cost and limit – being fixed. The results are the following (Table 3).

Table 3. Time Series weights

Euclidean

Raw	92.4	Residual	61.6	Norm Time	88.5
Norm Amplitude	91.3	Norm Time & Amplitude	13.1		

DTW

Raw	90.5	Residual	79.8	Norm Time	83.0
Norm Amplitude	82.7	Norm Time & Amplitude	3.48		

Note Table 3: The weights amount for the importance of each factor in the global similarity measure (No units).

The simultaneous normalizations in time and amplitude are disregarded for both euclidean and DTW distance measures. In fact, signals whose evolution is different may be found similar after this transformation. The other weights remains in the same range, between 61.6 and 92.4 and none of them is really prevalent on the others.

The third step is to remove the disregarded factors and optimize

again the set of parameters. The results are the following (Table 4).

Table 4. Time Series weights

Euclidean

Raw	99.9	Residual	98.0	Norm Time	98.6
Norm Amplitude	0.02				

DTW

Raw	93.4	Residual	81.9	Norm Time	56.4
Norm Amplitude	30.1				

Note Table 4: The weights amount for the importance of each factor in the global similarity measure (No units).

Regarding the Euclidean distance, the amplitude normalization is disregarded. In fact, as suggested earlier, this normalization lead to find similar signals potentially very different before the amplitude transformation. The three other comparisons remain equivalent even the normalization in time. Increasing the set of durations involved in the artificial data generation may affect this factor by reducing its relative importance.
Regarding the DTW distance, the elasticity of the measure reduces the preprocessing impact by adapting itself. It results in high weights for raw data or residual comparison and not negligible weights for the normalized signals comparisons.
As a result, normalizations have less impact on the comparison than raw data or residual data. However, these results should be put into perspective since in this study, the objective functions only take into account the closest or the five closest cases. In any event, they are not inconsistent with the observations previously stated.

4. *Applying the GA on the global model*:
This last optimization has been made by fixing all previously established parameters. The results are the following (Table 5).

Table 5. Global parameters

Symptoms	37	Time Series	30

Note Table 5: The weights amount for the importance of each factor in the global similarity measure (No units).

Using this set of parameters, the error rate remains up to 15% with some artificial data samples.

5.2.4. Comments on the results

Parameters adjustments obtained using artificial data are consistent with the one performed with the experts' support. They helped to adjust the weights relative to the leak flow factors. As a result some symptoms and some time-series preprocessing preponderate on others in the comparison. In particular, normalizations, noise and steps are more or less disregarded in comparison with other factors.

These results should be put in perspective. Indeed, some difficulties have been encountered and partly tackled in this artificial data adjustment:

- Artificial data should perfectly mimic the reality. Otherwise, including these results in the global model may alter its representativity;
- Even whether these signals have been generated artificially, the establishment of a predefined similarity is not that simple. The solution employed was to restrict the objective function to the closest or the five closest cases. This restriction may lead to incoherent results since, as an example, two identical signals except for their mean level may be considered as identical by the experts but they are not in this model;
- When dealing with many parameters, some of them may be correlated which potentially alter the representativity of the results. Indeed, multiple parameters configuration lead to the same optimum. This is partly solved by adjusting subsets of parameters;
- Optimizing subsets of parameters may give a local optima which differ from the global optima of the optimization problem.

Some prospects can be laid out. Extending the artificial data scope to other data than the leak flow may increase its representativity. Ideally, it would not only take into account the four generic categories but directly the associated fault. Moreover this optimization is not specific to the application case and the overall approach could be used for other application cases.

6. PROGNOSTIC PROSPECTS

Given a situation, the implemented tool is able to find the closest situation in the past event case-base. This closest case is related to additional documents regarding the diagnostic or the corrective actions which could inform on the expected remaining useful life.

In this study, the prognostic approach is based on the time series comparison. By synchronizing the time series, the closest case can be used to assess the expected remaining life before an alarm or an outage. The figure 12 illustrates this prognostic.

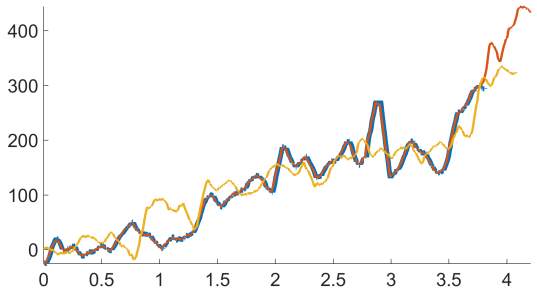


Figure 12. Prognostic illustration

Note 12: This figure illustrate the two closest signals to a given one (represented in blue). These two signals can help to predict the evolution of the current signal.

Judging from these closest cases, the leak flow may increase up to 138L/h in 11.8 days or 29L/h in 8.2 days. These cases give a prognosis based on the real experienced abnormalities. This information could, in addition to already stored data regarding the closest cases, support the experts in their decisions.

7. CONCLUSION

The establishment of the *Condition Based Maintenance* policy requires a perfect mastery of the state of functioning of monitored equipment in real-time. For this purpose, early detection tools warn the operators when a potential anomaly is about to occur. Once a situation is detected, it is analyzed and diagnosed to identify the adapted corrective actions that should be performed.

This paper presented the application of the Case-Based Reasoning approach for the monitoring and diagnostic of nuclear power plants equipment. The proposed work encompasses different methods from the literature to create a case structure, define a similarity measure, and then evaluate and refine the model. The resulting tool supports, by case comparison, the experts in their decisions. Many challenges have been tackled such as the establishment of a similarity measure for heterogeneous data, the multidimensional signal comparison and the optimization of the similarity measure.

The main difficulty remained in the establishment of a standardized case structure. Each case need to be rigorously and completely defined, in particular concerning the diagnostic and the corrective actions. These were missing or incomplete for the seal #1 case study, which lead to adapt the classification method.

Using this approach, a smart choice of the similarity measure led to great classification results. The proposed approach can be extended to other case studies. This study opens up new prospects for diagnostic and prognostic approaches.

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-based reasoning; foundational issues, methodological variations, and system approaches. *AI COMMUNICATIONS*, *7*(1), 39–59.

Condition monitoring and diagnostics of machines – vocabulary (Standard). (2012, September). International Organization for Standardization (ISO).

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical clustering. In *Cluster analysis* (pp. 71–110). doi: 10.1002/9780470977811.ch4

Jardine, A. K., Lin, D., & Banjevic, D. (2006, oct). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510. doi: 10.1016/j.ymssp.2005.09.012

Lesot, M.-J., Rifqi, M., & Benhadda, H. (2008, December). Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *1*(1), 63-84. doi: 10.1504/IJKESDP.2009.021985

Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA, USA: MIT Press.

Senin, P. (2008, December). *Dynamic Time Warping Algorithm Review* (Tech. Rep. No. CSDL-08-04). Department of Information and Computer Sciences, University of Hawaii, Honolulu, Hawaii 96822.

Song, Q. (2001). Ecm - a novel on-line, evolving clustering method and its applications. In *In m. i. posner (ed.), foundations of cognitive science* (pp. 631–682). The MIT Press.

Stanfill, C., & Waltz, D. (1986, December). Toward memory-based reasoning. *Commun. ACM*, *29*(12), 1213–1228. doi: 10.1145/7902.7906

Todorov, Y., Feller, S., & Chevalier, R. (2015, July). Making the investigation of huge data archives possible in an industrial context an intuitive way of finding non-typical patterns in a time series haystack. In *Informatics in control, automation and robotics (icinco), 2015 12th international conference on* (Vol. 01, p. 569-581).

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2012, feb). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275–309. doi: 10.1007/s10618-012-0250-5

BIOGRAPHY

MAI Carole was born in Vernon, FRANCE, in 1992. She received an Engineering degree from SUPELEC, France in 2014 and a Master degree of Mathematical Modeling with distinctions from UCL, United-Kingdom in 2014 (Frank T. Smith Prize). In 2014, she joined the R&D department of EDF - Éléctricité de France - as a research engineer. Her current research interests include machine learning and data analysis - both numerical and textual - applied to condition based maintenance optimization.