*

LT-PEM Fuel Cells diagnosis based on EIS, clustering, and automatic parameter selection

Damien Chanal, Nadia Yousfi Steiner, Didier Chamagne, Marie-Cécile Pera

***Abstract*— In the field of fuel cells, early detection of faulty conditions can significantly improve the lifetime. Then, signal analysis techniques such as electrochemical impedance spectroscopy combined with machine learning algorithms can generate a representation of the system state of health space using data in known conditions. Although the onboard measurement of EIS can be done by controlling the harmonic content of the power converter at the output of the fuel cell, the implementation of in-vehicle diagnostic algorithms is still limited by the absence of a large database listing the evolution of performance throughout the life cycle. This paper presents a fast-diagnostic method able to consider the occurrence of new data to adapt the dimensional space representing the health state and compensate for the lack of data. Available measurements come from two low-temperature proton exchange membrane fuel cell technologies characterized by two laboratories. The results presented in the paper show that the automatic parameter selection provides performances as good as the ones obtained by an expert. The feasibility of the approach has also been demonstrated on a low-cost embedded platform.**

***Index Terms*— Clustering methods, Fault diagnosis, Feature Selection, Fuel cells, Standardization**

I. INTRODUCTION

One of the promising technologies in the context of low greenhouse gas emission technologies is the proton exchange fuel cell (PEMFC). PEMFCs are energy converters that transform hydrogen and oxygen into electricity, heat, and water. Their electric efficiency is generally about 50% at the beginning of life. Currently, PEMFCs are of particular interest in the fields of transportation and stationary applications and the low-temperature proton exchange membrane fuel cell is one of the most developed technologies.

The development of fuel cells is hampered by their limited lifetime as well as their susceptibility to defects. The U.S. Department of Energy's ultimate goal is to increase the lifetime of PEMFCs to 8,000 and 80,000 hours for transportation and stationary applications [1].

To achieve and improve these lifetime goals, monitoring, and diagnostic tools suitable for fuel cell systems should be used to detect early and allow correction of any abnormal condition that may occur.

One of the most widely used fuel cell characterization techniques is Electrochemical Impedance Spectroscopy (EIS). According to [2], EISs provide information on many fuel cell conditions, such as membrane degradation, catalyst activity decrease, reactant poisoning, humidification, and aging. The principle of an EIS is to inject a small AC disturbance and analyze the voltage response of the fuel cell to extract its impedance. The operation is repeated for different frequencies of disturbances. The obtained impedances are then analyzed in Nyquist and Bode diagrams to determine some physical parameters such as membrane resistance, gas diffusivity, or polarization resistance. However, the utilization of EIS in automotive applications is complex. According to [3], a device capable of injecting an AC signal is too expensive for automotive applications. One of the solutions is to use the DC/DC converter connected to the fuel cell terminals to generate the disturbance. It allows the collection of the EIS online, at a low cost without additional equipment [4]. The design and realization of EIS through the converter have been studied in European projects, the Health Code project [5], and the RUBY project [6] suit.

Once the impedances are obtained, the diagnosis algorithms can be used to determine the State of Health (SoH) of the fuel cell.

Diagnosis methods can be classified as model-based and non-model-based (data-driven) approaches. A review of these methods is proposed in [7], [8]. However, it is worth noting that in the two approaches, artificial intelligence can be used to establish a relationship between inputs and outputs without using any physical knowledge. Indeed, PEMFCs are considered complex systems due to the interaction of several phenomena (thermal, electrical, fluidic ...), which makes accurate modeling difficult. For this reason, the model-based approach is questionable for the real-time diagnosis of this system. Generally, a database is needed to train with known data (off-line part) before being able to analyze unknown data (online part) and return the SoH of the system. For batteries, large databases are already available in open source as referenced in [9]. Currently, the number of open-access fuel cell databases is low, which is a bottleneck for data-driven algorithm development. To take this limitation into account, the diagnosis

This project has received funding from the Fuel Cells and Hydrogen 2 Joint Undertaking (JU) under grant agreement No 875047 Website: https://www.rubyproject.eu/. This work has been supported by the EIPHI Graduate School (contract ANR-17- EURE-0002) and the Region Bourgogne Franche-Comté. We acknowledge the European project HEALTH CODE which provides the data used in this paper. Website: http://pemfc. health-code.eu/. The corresponding author of this paper is Damien CHANAL.

All authors are with Université de Franche-Comté, CNRS, institut FEMTO-ST, FCLAB, F-90000 Belfort, France. Their respective e-mails are: damien.chanal@femto-st.fr, nadia.steiner@univ-fcomte.fr, didier.chamagne@univ-fcomte.fr and marie-cecile.pera@univ-fcomte.fr

algorithms must be able to re-train quickly on new data acquired during operation.

In several scenarios, model-based and non-model-based approaches have been applied to diagnose fuel cell systems. In [10], a computational efficiency approach based on fuzzy logic combined with clustering is proposed to detect several levels of flooding and drying in a fuel cell stack. Authors in [11] proposed to use a probabilistic Bayesian neural network to detect faulty conditions in a PEMFC system. Four faults were tested, 3 related to the auxiliaries (fan, cooling system, and hydrogen supply line) and one to the increase of the fuel crossing inside the fuel cell. In [12], an online implemented support vector machine used to monitor individual cells in a fuel cell is presented. The approach shows good performance to detect pressure anomalies, drying, and air starvation conditions. Another approach based on fuzzy and pattern recognition named Visual Block-Fuzzy Inductive Reasoning is presented in [13]. The authors compare their method with a model-based methodology to detect 5 faulty conditions linked to the stack voltage, the oxygen management, and the compressor. Also, results obtained show favorable performances for the Visual Block-Fuzzy Inductive Reasoning approach.

The common feature of the data-driven approach is the impossibility of extrapolating to unknown conditions without a re-training step which is not always possible. For that purpose, an interesting approach to the diagnosis using Fuzzy C-Means clustering is presented in [14]. The approach shows good performances for two different databases. According to the authors, the choice of clustering allows improving both the training time because of the simplicity of use and the capacity to process large databases that are generally the penalizing points of the other diagnostic methods. However, a limitation of this method is the need for empirical tests for the feature selection step but also the user expertise to determine the number of clusters in the classification step. To improve these points which have a huge impact on the results and limit the usability of the method, this paper introduces the use of a specific robust criterium to have an autonomous algorithm that can quickly retrain itself when measuring new data and a low need for user expertise.

Section 1 is dedicated to the presentation of the so-called classical approach presented in Fig. 1. The different steps leading to the classification of the EIS spectra will be detailed followed by a presentation of the databases used.

Section 2 is dedicated to the description of different methods of data standardization. These methods can impact the results by reducing the importance of outliers and the computation time.

Section 3 presents a way to improve the feature selection and some popular clustering validation indices. The classic feature selection is based on the use of ranked features and empirical tests to determine how many have to be used while the improved selection uses a simple threshold. The clustering indices are used to automate the clustering part of the diagnostic approach allowing the addition of new data.

The results are presented in section 4. The parameters that allow the user's expertise to be reduced as much as possible and the generalization of the algorithm are highlighted.

II. Presentation of the Diagnosis Approach

A. The approach

The method developed in the Health Code project is based on the use of a Fuzzy C-means classifier to detect the SoH of a fuel cell from EIS measurements performed online through a relevant control of the fuel cell output converter. A global presentation of the diagnosis approach is given in Fig. 1 and detailed in this section, however, more information about this method and data are available in [14]. The offline processing is composed of the following steps: First, features from the EIS are extracted. These features are standardized which is the step this paper is focused on. Then, a selection of the ones containing the best information to discriminate the SoH of the fuel cell is done. Finally, data are classified using Fuzzy C-means clustering.

In the developed algorithm the extracted features are: the minimum and maximum magnitudes of impedance respectively named (*mm*) and (*Mm*); the difference between the maximum and minimum magnitude (*ΔMag*); the polarization resistance (*R_pola*); the minimum and maximum phase respectively (*mp*) and (*Mp*); the phase at a frequency of 0.1 Hz (*P1*); the difference between P1 and Mp (*ΔPha).* Also, an analysis of the phase during a linear part of the Bode diagram is done ([0.1 -1] Hz). Equation (1) describes phase as the first-order equation of frequency (f):

$$Phase = A \times f + B \tag{1}$$

Coefficients A and B are extracted as features.

The standardization method used was based on quantile information to make data follow a uniform distribution. This method was selected because of its ability to handle outliers and noisy data. The feature selection approach used the Pearson Correlation Coefficient (PCC) to filter data with a high linear correlation and then an ANOVA F-Test to rank features. Once the generation of features is done, the diagnosis algorithm consists of using a Fuzzy C-means clustering to create clusters that will be used to detect the SoH of training data. During this step, the experience of the user is required. As a matter of fact,

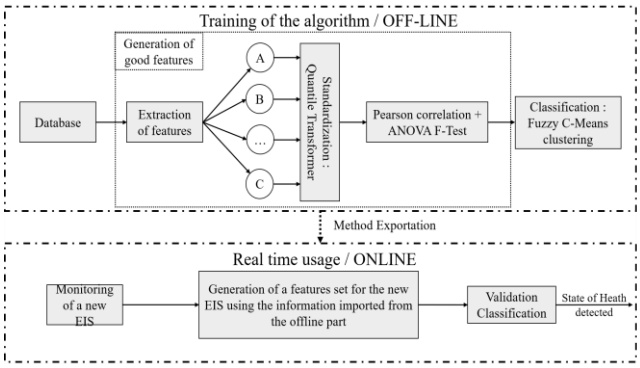

Fig. 1. Global principle of diagnosis tool developed in [14]

to optimize the creation of clusters, for each fault, the user will give only data associated with the faults and enter the desired number of clusters (in the presented diagnosis it was the number of faults' level tested). It permits optimizing the localization of clusters for each fault even if it modifies the non-supervised character of Fuzzy C-means. Concerning one of the studied faults, fuel poisoning, a specific data clustering is made to identify CO poisoning in the first place. It is easy to detect as it exhibits positive values of the imaginary part of the impedance.

The online step consists in using the information obtained during the off-line step. For that purpose, the best fest features to extract, bounds of standardization, and cluster centers coordinates are transferred to the system implemented online. In order to associate a known condition with new measured data, the algorithm proceeds in two steps. First, it extracts the best features and standardizes them according to the information received from the offline part. Second, it computes the Euclidean distance between the transformed features and the previously computed cluster centers. The associated SoH corresponds to the closest cluster.

B. Datasets' presentation

Two datasets are tested in this paper. They came from two fuel cell stacks tested during the European project Health Code. For each stack, the number of spectra retained for each condition is presented in Fig. 2.

The first one is a short hydrogen-oxygen stack that is supposed to replicate the operations of a real backup system coupled with an electrolyzer. A total of 5 conditions have been tested: nominal, flooding, drying, hydrogen starvation, and oxygen starvation. For the nominal and flooding conditions, acquired spectra are considered as one level according to the analysis by an expert. For other conditions (i.e drying and starvations), spectra are separated into 3 levels (low, medium, and high) depending on the experimental conditions and time exposure to the degradation.

The second stack is a hydrogen-air technology intended to be implemented in a micro-CHP system fed by natural gas. Seven conditions have been tested: nominal, flooding, drying, anode starvation, cathode starvation, and poisoning (carbon monoxide & sulfur). Only one level has been associated with the nominal conditions. Regarding the water management conditions (i.e flooding and drying) two levels have been determined which correspond to faults occurring at the anode and cathode sides. Two levels have also been determined for the starvation conditions. They are associated with low and moderate/high degradation levels. Concerning the poisoning faults, several rates of poisoning have been studied. For carbon monoxide contamination, 4, 8, 12, and over 80ppm were tested. Spectra are grouped into 4 levels however because the recovery of CO poisoning is complicated (platinum reduction), the exposure time has a high impact on spectra, so this separation is not representative. The same study has been done for sulfur poisoning using 4 rates: 4, 6, 8, and 10ppm, and because the recovery of sulfur poisoning is impossible (platinum dissolution), it is not possible to recover original performances between two consecutive experimental tests. For this reason, like with carbon monoxide, the choice of 4 levels is only informative and not representative of reality.

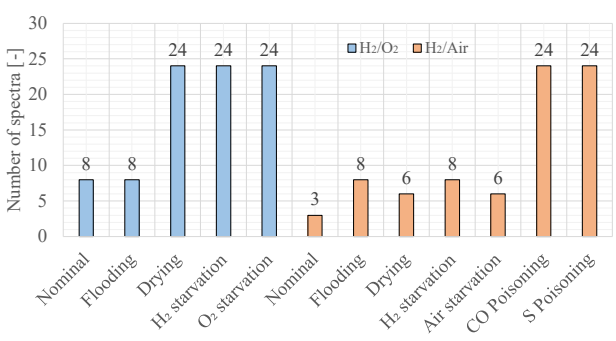

Fig. 2. Composition of the two datasets used of healthy and faulty conditions

Table I shows the input variables used for EIS measure.

TABLE I : INPUT VARIABLES FOR EIS MEASUREMENT

Input	Value for laboratories test
Frequency	10 kHz – 10 mHz (log scale)
Current value	5 - 10% of DC H_2/O_2: 210 A H_2/Air: 40 A
Number of periods	1 – 20 (depending on frequency)
Sampling frequency	At least 100 times the injected frequency

III. STANDARDIZATION METHODOLOGIES

One of the key points in the development of machine learning algorithms is the generation of good-quality features. Indeed, a good feature generation decreases the predominance of possible outliers and noises, reduces the computation time but also improves the accuracy and the robustness of the results. In the case of classification algorithms that rely on distance calculations, the choice of a relevant standardization method is crucial. It consists in adjusting data values when they are not in the same range to eliminate distortions of the SoH space and make them comparable. The magnitude of features affects algorithms' performances, especially when some features have much larger values than others. There are three main families of methods to standardize data: Normalization, Linear scaling, and Nonlinear transformation. A short presentation of the main standardization of each family is presented below. Each algorithm presented is implemented in Scikit-learn [15] and the interested reader can refer to [16]–[18].

A. Normalization

In general, it is the features of the dataset which are standardized, however, it is also possible to standardize each sample so that its norm equals 1. This method of standardization is named normalization. It is interesting to normalize samples when the objective is to quantify the similarity of any pair of samples.

Mathematically a norm is the total size or length of all vectors

in a vector space of matrices. The norm of a vector *x* can be calculated at several levels *(p)* by using the equation below:

$$\|x\|_p = \sqrt[p]{\sum_i |x_i|^p} \quad (2)$$

where $p \in R$ is the level of the norm and *x* is the vector to be normalized. In machine learning, the normalization uses generally 3 levels of the norm which are: L1 norm is the sum of absolute values of vector *x* (*p*=1). L2 norm corresponds to the second level of the norm (*p*=2) which is the sum of squared values of *x*. The infinite norm corresponds to the level when $p \to \infty$. Once the norm is calculated, each member of the vector x is divided by the norm to obtain a unit vector. The formula is presented in (3):

$$x_{\text{normalized}} = \frac{x}{\|x\|_p} \quad (3)$$

Normalization is a powerful process, which can be used for tasks where it is possible to observe variability between the different conditions to classify. It is well adapted for clustering and text classification, however, in the case of noisy data, normalizers are sensitive to outliers which can impact the norm calculation.

B. Linear scaling

Linear standardization methods are the most widely used methods to scale features. They are quite simple to implement and work well for most databases. In addition, linear scalers are very useful to accelerate algorithms that use descent gradients. Indeed, in the case where one feature is higher than the other, it is more difficult to converge to the optimal value of the function. Different linear scaling methods use several indicators to standardize.

The first scaling method consists in scaling data in the range [0-1], it is also called “Min-Max feature scaling”. It consists of using minimal and maximal data as boundaries and rescaling data. Min-Max scaling is represented by (4):

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (4)$$

One of the advantages of the Min-Max scaler is that it allows putting in the same interval features that can be very different while keeping all information since the distance ratios are kept. In the case of algorithms based on the distance between points, it allows comparison between items with small and large values.

The second method of scaling data is called "Max Absolute Scaling". It uses the maximum absolute value of a vector x to scale the features in the range [0, 1] or [-1, 1] depending on whether they are negative values. This method consists in dividing the vector *x* by its maximal absolute value as shown in (5):

$$x_{\text{scaled}} = \frac{x}{\max(|x|)} \quad (5)$$

Max Absolute scaler is very similar to Min-Max scaler, nevertheless, it should be used for data that are already centered on zero.

The third method of linear scaling is called "Standard scaler". The objective of this method is to transform the features so that they have a mean of zero and a standard deviation of one as shown in (6):

$$x_{\text{scaled}} = \frac{x - \mu_x}{\sigma_x} \quad (6)$$

With μ the mean and σ the standard deviation.

Standard scaler allows for data centering and makes easier the use of statistical machine learning algorithms such as Principal Components Analysis (PCA). The main disadvantage of the three linear scalers presented above is that they are very sensitive to outliers in the dataset.

This is why standardization algorithms using statistics were developed. It is the case of a robust scaler that uses the median and interquartile range (IQR) of data to reduce the importance of outliers. The formula to standardize data is:

$$x_{\text{scaled}} = \frac{x - \text{median}}{\text{IQR}} \quad (7)$$

Equation 7 looks similar to (6), however median and IQR are more robust to outliers than mean and standard deviation because they use the position of the data rather than the values.

C. Non-linear transformation

Even if the "robust scaler" permits the reduction of the importance of extreme values, it can be better to use non-linear transformations. These non-linear transformations allow transforming the data so that they change their distribution. Two types of standardization allow doing this: power transformations and quantile transformations.

Power transformations are parametric and monotonic transformations. They are useful to stabilize the variance of features that are heteroscedastic and map data to make them more Gaussian-like. There are 2 main power transformations: Box-Cox and Yeo-Johnson transformations. Box-Cox transformer [19] is defined by (8):

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases} \quad (8)$$

With *x* vector to transform, and λ the power parameter of transformation which is determined through maximum likelihood estimation.

Box-Cox transformer allows transforming a dataset into a Gaussian-like distribution. However, it is limited in that it allows only strictly positive values. Because data from EIS are positive and negative, it is not possible to use this transformer. This is not the case with the Yeo-Johnson transformer [20]

which has no restrictions. It is defined in (9):

$$x_i^{(\lambda)} = \begin{cases} \frac{[(x_i+1)^{\lambda}-1]}{\lambda} & \text{if } \lambda \neq 0, x_i \geq 0 \\ \ln(x_i+1) & \text{if } \lambda = 0, x_i \geq 0 \\ \frac{-[(-x_i+1)^{2-\lambda}-1]}{(2-\lambda)} & \text{if } \lambda \neq 2, x_i < 0 \\ -\ln(-x_i+1) & \text{if } \lambda = 2, x_i < 0 \end{cases} \quad (9)$$

The Box-Cox and Yeo-Johnson methods have the same objectives; however, they are slightly different. Indeed, in the case where the values are strictly positive, the Yeo-Johnson transformation is identical to the Box-Cox power transformation of (x+1). However, these two methods are regularly used in many domains such as machine learning. In [21], properties of Box-Cox transformation for pattern classification are presented. In [22], the effect of standardization is studied on speech emotion recognition; the Yeo-Johnson transformer is compared to linear scaling and normalizer. While in [23], the authors study the effect of linear scalers and non-linear transformers with K-nearest-neighbor and support vector machine algorithms.

In addition to the power transformer which makes data Gaussian-like, the quantile transformer uses information contained in the quantile to make data follow a uniform or normal distribution. The quantile transformer formula is presented in (10):

$$G^{-1}\big(F(x)\big) \quad (10)$$

With F the cumulative distribution function of x and G^{-1} the quantile function of output distribution G.

Quantile transformers are very useful to reduce the importance of outliers. The negative point of this function is that it distorts correlations and distances within and across features because it smooths the original distribution. Nevertheless, the characteristics measured at different scales are more easily comparable. In addition, it is worth noting that when a new sample is transformed with a quantile transformer, it is not possible to extrapolate it, unlike other standardization methods. Indeed, if the new data are larger or smaller than those used to determine the transformation boundaries, the standardized value is limited to the minimum or maximum fitted value. For example, in the case of a uniform distribution, the possible range is [0, 1], so if a new outlier appears, the standardized value will be 0 or 1.

IV. AUTOMATION OF PARAMETER SELECTION

A second key point to consider when developing a Machine Learning algorithm is its ease of use. Indeed, from a computational cost point of view, a complex algorithm that takes a long time to train or retrain (for example neural networks) will require much more effort to set up than a simple algorithm. Moreover, another need for diagnosis algorithms to be used in a system throughout its lifetime is the need for a database containing a large number of different conditions at different times. Indeed, a nominal state at the beginning of life and a nominal state at the end of life might not be identical and can be easily classified as a faulty state if the fuel cell degradation is not listed in the database or updated during the lifetime (using a re-training of the algorithm).

To reduce the need for empirical testing as well as to facilitate lifelong learning, two modifications have been made to the diagnosis approach presented in section II.A: the first one facilitates the automatic selection of the number of features to use and the second one automates the choice of the number of clusters characterizing a defect.

A. Automatic feature selection

As explained in section II.A, features are extracted from EIS spectra using physical knowledge. These are then standardized, and the best ones are selected using a filtering step with the Pearson Correlation Coefficient and then ranked with ANOVA F-Test. The combination of filtering and ranking steps reduces the number of features to use and obtains better results of classification as shown in [14]. However, it needs an empirical study to determine the best number of features to select. Indeed, using features containing little information can both increase the computation time unnecessarily but also distort the state of health space, and reduce classification performance.

For that purpose, the proposed improvement consists in keeping the filtering and ranking steps, however, instead of empirically testing the features, the obtained scores are represented in percentages. Then, the algorithm selects all features which are below a threshold defined by the user. Three thresholds have been studied and their performances are compared in section V. Fig. 3 shows the synoptic of the automatic feature selection process as explained above and in section II.A. The use of a limit has the advantage of selecting only features containing sufficient information, thus reducing complexity and computation time.

B. Automatic cluster number selection

In the first developed approach, the number of clusters used to characterize a condition is defined by the user and is equivalent to the number of degradation levels tested in the database. However, this methodology implies a precise knowledge of the database and cannot be applied when the user wants to re-train the algorithm using the newly classified EIS. To overcome these difficulties, and to determine the optimum number of clusters (i.e., to solve a cluster validity problem), it is possible to use validation clustering indices. This section is devoted to the presentation of the fuzzy clustering used in this study as well as the presentation of the various clustering indices that have been retained.

1. Fuzzy C-Means clustering algorithm

The fuzzy C-Means clustering algorithm is one of the most used fuzzy clustering algorithms [24]. The fundamental aspect of fuzzy clustering is to determine the similarity measure in which the distances between pairs of data points are calculated. In fuzzy C-Means clustering, the models are treated as vectors

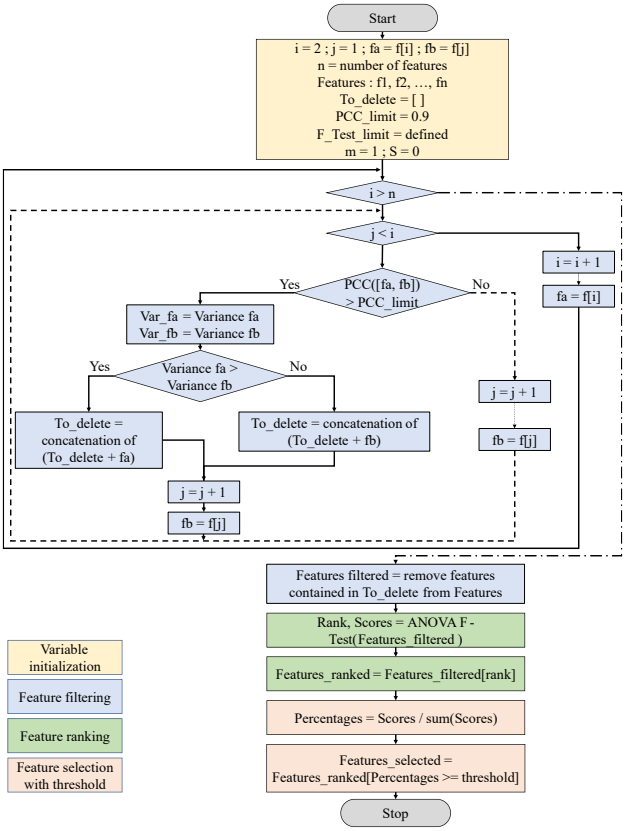

Fig. 3. Flow chart detailing the full process of feature selection designed

in Euclidean space.

For a collection of *n* data in a dataset $X = \{x_1, x_2, \ldots, x_n\}$ to be separated into *c* clusters, the objective function J_m to minimize is defined as shown in (11):

$$J_m(U,V) = \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^m (d_{ij})^2 \quad (11)$$

$$d_{ij} = \|x_j - c_i\| \quad (12)$$

where u_{ij} is the membership of the data *j* in the cluster *i* and

$m \in [1, \infty]$ a fuzzifier that controls the fuzziness of membership of data. The membership can be calculated using (13):

$$u_{ij} = \frac{\left(1/d_{ij}^2\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{c}\left(1/d_{ij}^2\right)^{\frac{1}{m-1}}} \quad (13)$$

And the cluster coordinates can be calculated using (14):

$$c_i = \frac{\sum_{j=1}^{n} (u_{ij})^m \cdot x_j}{\sum_{j=1}^{n} (u_{ij})^m} \quad (14)$$

2. Cluster validity indexes

As shown above, the C-Means clustering algorithm needs the user's expertise to inform the number of clusters to use. As this is not always possible, it is necessary to use validation criteria for clustering. These criteria are designed to analyze the structure of the data and compare the results obtained for several numbers of clusters to determine which one is optimal. Among the validation criteria reported in the literature, this study focuses on the criteria below. Other cluster validity indexes can be found in the literature such as [25] and more recently [26], however, these indices introduce one or several thresholds used to exclude noisy data. Because we aim to propose a method that refers as less as possible to expert knowledge, these indices are not retained.

In 1974, Bezdek proposed the first indices named Partition Coefficient (PC). PC computes the relative mean V_{PC} of the fuzzy intersection between pairs of fuzzy subsets by their algebraic product. It is defined in (15):

$$V_{PC} = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^2 \quad (15)$$

The best number of clusters is obtained by maximizing V_{PC}

A modification of V_{PC} has been proposed by Dave in [29] to correct the monotonic tendency by applying a linear transformation. The Modified Partition Coefficient (V_{MPC}) is defined in (16):

$$V_{MPC} = 1 - \frac{c}{c-1}(1 - V_{PC}) \quad (16)$$

In addition to the partition coefficient, Bezdek defined another validation clustering index based on the Shannon entropy function [30]. This index is named Partition Entropy (PE) and its objective is to describe the fuzzy uncertainty contained in each data. To calculate this fuzzy uncertainty in a subset, it calculates the average of the fuzzy entropies V_{PE} as shown in (17):

$$V_{PE} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij} \log_\alpha u_{ij} \quad (17)$$

where $\alpha \in (1, \infty)$, in this study we retained only $\alpha = 1$ because it is the most common value associated to $\log_\alpha$. The best number of clusters is obtained by minimizing V_{PE}.

To compensate for the monotonic tendency of PE to decrease with the augmentation of clusters, a first modification has been proposed in [30], [31] with the Scaled Partition Entropy (SPE). The idea of V_{SPE} is to refine the lower limit of PE and is defined in (18):

$$V_{SPE} = \frac{V_{PE}}{\log_\alpha c} \quad (18)$$

Another adaptation of PE is presented in [31] with the Normalized Partition Entropy (NPE). NPE is Dunn's normalized version of PE and such as for SPE, its objective is to counter the tendency of PE to monocratically decrease. V_{NPE} is defined as shown in (19):

$$V_{NPE} = \frac{V_{PE}}{(1-\frac{c}{n})} \quad (19)$$

Other validity indices which use other metrics than those based on PC or PE can be found in the literature. Some indices such as Fukuyama-Sugeno [32], fuzzy hypervolume [33], Xie and Beni [34], Kwon [35], PBM [36], and PCAES [37] can be cited.

Fukuyama-Sugeno (FS) validity index is based on the difference between compactness and separation metrics. Compactness is calculated by the intra-cluster distance while separation is calculated by the inter-cluster distance. FS validity index V_{FS} is defined by (20):

$$V_{FS} = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \left(\left\|x_j - c_i\right\|^2 - \|c_i - \bar{c}\|^2 \right) \quad (20)$$

where $\bar{c}$ is the average of data represented by (21):

$$\bar{c} = \frac{1}{n}\sum_{j=1}^{n} x_j \quad (21)$$

The optimal number of clusters is obtained when V_{FS} reaches the minimum value.

Gath and Geva proposed in 1989 the fuzzy hypervolume (FHV) validity index which uses the fuzzy covariance matrix and is developed in (22, 23):

$$V_{FHV} = \sum_{i=1}^{c}[\det(F_i)]^{\frac{1}{2}} \quad (22)$$

$$F_i = \frac{\sum_{j=1}^{c}(u_{ij})^m (x_j - c_i)(x_j - c_i)^T}{\sum_{j=1}^{n}(u_{ij})^m} \quad (23)$$

The optimal number of clusters is obtained when V_{FHV} reaches the minimum value.

In 1991, Xie and Beni proposed an index for clustering [34] using *m=2*. In 1995, this index has been modified by Pal and Bezdek [38] to accept different values of *m* as shown in (24):

$$V_{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^m \left\|x_j - c_i\right\|^2}{n \min_{i \neq j}\left(\left\|c_i - c_j\right\|^2\right)} \quad (24)$$

In (24), the numerator represents the compactness of the fuzzy partition, and the denominator the grade of the separation between clusters. The optimal number of clusters is obtained by minimizing V_{XB}. However, Xie and Beni stated that the validity index decreases monotonically when the number of clusters is close to *n*.

In 1998, Kwon proposed a validity index to eliminate the monotonically decreasing tendency when the number of clusters becomes very large. The equation is presented in (25):

$$V_K = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^2 \left\|x_j - c_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|c_i - \bar{c}\|^2}{\min_{i \neq k}(\|c_i - c_k\|^2)} \quad (25)$$

The optimal number of clusters is obtained when V_K reaches the minimum value.

Pakhira proposed in 2003 the PBM validity index which is used for crisp clustering and propose a modified version that incorporates fuzzy distances called the PBMF validity index [36]. PBMF equation is shown in (26):

$$V_{\mathrm{PBMF}} = \frac{1}{c} \times \frac{E_1}{J_m} \times D_c \quad (26)$$

with

$$E_{\mathrm{i}} = \sum_{j=1}^{n} u_{ij}\left\|x_j - c_i\right\| \quad (27)$$

$$D_{\mathrm{c}} = \max_{i,j}\left\|c_i - c_j\right\| \quad (28)$$

$$J_{\mathrm{m}} = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m\left\|x_j - c_i\right\| \quad (29)$$

The optimal number of clusters is obtained when V_{PBMF} reaches the minimum value.

In 2005, Wu and Yang proposed the partition coefficient and exponential separation index (PCAES) [37] which pays special attention to outliers and noisy data while validating the partitioning results. PCAES combines a measure of compactness and separation criteria of partitioning. In [37] PCAES is calculated as shown in (30):

$$V_{PCAES} = \sum_{i=1}^{c}\sum_{j=1}^{n}\frac{u_{ij}{}^2}{u_M} - \sum_{i=1}^{c}\exp\left(\frac{-\min_{i \neq k}\{\|c_i - c_k\|^2\}}{\beta_T}\right) \quad (30)$$

$$*u_M = \min_{1 \leq i \leq c}\{\sum_{j=1}^{n} u_{ij}{}^2\} \quad (31)$$

$$\beta_T = \frac{\sum_{i=1}^{c}\|c_i - \bar{c}\|^2}{c} \quad (32)$$

In (31), $*u_M$ is calculated using minimal compactness, however, in [37], the authors state that $*u_M$ is bounded between]0, 1] and calculate the most compact cluster partitioning coefficient. However, by calculating the minimal value it is the less compact cluster that the partitioning coefficient is calculating. We assume that this is an error and use the equation of $u_M$is proposed in [40] and detailed in (33):

$$u_M = \max_{1 \leq i \leq c}\{\sum_{j=1}^{n} u_{ij}{}^2\} \quad (33)$$

V. Results

To define if an algorithm is powerful or not, it is necessary to define metrics able to measure the correct classification of data. In addition to relevant metrics, it is better to evaluate the classification of data with different training and testing sets to have a fairer view of performances. A good method to measure the generalization ability without increasing the need of data is to use a cross-validation process. It is a statistical method that consists in dividing the database into several parts (*k* parts) to train it with k-1 parts and test it on the last part. It exists several ways to divide the dataset into *k* parts but the retained one is the "Leave One Out" which consists of a training dataset with all data except one and proceeds by iteration to be able to test all data.

A. Evaluation of algorithms

One of the most useful ways to measure the effectiveness of a machine learning algorithm is to define multiple metrics. The interest in using several indices (5 in our paper) is to observe the most common types of errors to have a better understanding

of the algorithm and perhaps to add extra steps when detecting certain conditions to limit the risk of errors. In this study, widely used indices are computed to evaluate the performances and analyze the type of mistakes if any. The first index is the confusion matrix which allows observing the 4 cases of classification for a specific condition "f" as shown in Table II:

- "*Tp*" is the number of samples correctly assigned to "f"
- "*Fn*" is the number of samples wrongly assigned to "f"
- "*Fp*" is the number of samples wrongly not assigned as "f"
- "*Tn*" is the number of samples correctly not assigned as "f"

TABLE II : REPRESENTATION OF CONFUSION MATRIX

	Actual condition	
Detected condition	***True***	***False***
True	Tp	Fp
False	Fn	Tn

The second index is the accuracy score which provides a representation of the number of correct classifications under all samples. Equation (34) shows the formula to determine the accuracy score:

$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{34}$$

The third index is the precision score which is useful to observe the ratio of correct positive classification to all positively detected classifications. The formula for the precision score is presented in (35):

$$Precision = \frac{Tp}{Tp + Fp} \tag{35}$$

Fourthly, the recall score, also called sensitivity, is defined as the ratio of correct positive classification to all occurrences of actual true conditions as shown in (36):

$$Recall = \frac{Tp}{Tp + Fn} \tag{36}$$

Finally, the F1 score is one of the useful indexes to evaluate an algorithm. It permits the measurement of the weighted average of precision and recall scores. The F1 score formula is presented in (37):

$$F1\ score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{37}$$

B. Impact of standardization

The same standardization methods are applied to both datasets. The objective of this comparison is to visualize the impact of standardization on data with different characteristics, at the same scale as well as on data with outliers. Fig. 4 and Fig. 5 show the results obtained using the H_2/O_2 and H_2/Air databases respectively but also the number of features needed to obtain the best results. Data obtained with the H_2/O_2 stack have all the same order of magnitude, which is not the case with data from the H_2/Air stack. This is due to the fuel poisoning faults which, at high concentrations, lead to impedance values much larger than in other operating conditions. Results were obtained using the "Leave One Out Cross-Validation" (LOO CV) methodology. This allows getting as close as possible to utilization in real life where the EIS would be tested 1 by 1, but also, to use a maximum of spectra for training since the number of available data is low. The study of standardization impact is based on the work done in [41].

As shown in Fig. 4 and Fig. 5, the standardized data allow for improving the efficiency of the diagnostic algorithm. Indeed, the choice of a correct standardization methodology allows for improving the F1 score by about 12% and 30% for H_2/O_2 and H_2/Air stacks respectively. The results in table form are presented in appendixes A and B.

In the case of the H_2/O_2 stack, the best results are provided by the main linear scaling methods and nonlinear transformations. However, it is interesting to note that the three normalizers generate more confusion in the algorithms (a 7 to 10% decrease in the F1 score compared to the case with raw data). This loss of performance means that samples are not different enough from each other to obtain good-quality features. Max Absolute scaler doesn't improve classification results compared to other scalers which provide a F1 score better than 90%. Nevertheless, only three methods obtain more than 95% of correct classification: Robust scaler, Yeo-Johnson, and Uniform Quantile Transformer. The specificity of these three methods is that they consider outliers that can be present in data even if they are all of the order of magnitude.

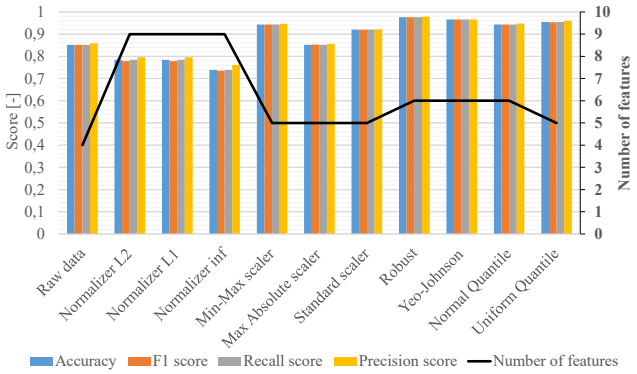

Fig. 4. Validation results obtained for H_2/O_2 dataset (LOO CV)

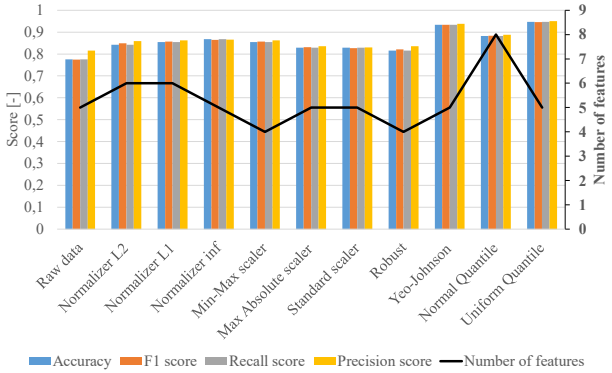

Fig. 5. Validation results obtained for H_2/Air dataset (LOO CV)

Regarding the H_2/Air results, it can be observed that compared to the first database, normalizers improve classification results by 5-10% due to the presence of samples at different scales. However, in comparison to the first database, almost all standardization methods give results below 90%. In this configuration, poisoning fault highly impacts the standardization of data to have a correct standardization even if methods such as Robust scaler and Normal Quantile transformer are dedicated to reducing the outlier importance. The best methods are Yeo-Johnson and Uniform Quantile transformers which allow obtaining better than 90% of correct classification.

The results obtained for both datasets confirm the weakness of normalizers and linear scalers in handling outliers. Normalizers need sufficiently different data to work, which makes them more efficient in dealing with these outliers, but the results obtained with them are insufficient compared to other standardization methods. Only the uniform quantile and Yeo-Johnson transformers perform well (>90%) for both datasets, making them good candidates for generic use.

In the following, only the uniform quantile transformer will be retained. The Yeo-Johnson could have been used as well since both methods give similar results.

C. Impact of automatic feature selection

Once the standardization method is fixed for both databases, it is interesting to investigate how the automatization of the feature selection impacts the results. For this, three threshold values are tested to determine the minimum percentages of information to be retained in each feature. The thresholds tested are 10%, 5%, and 1%. In addition to the performances, the features selected will be analyzed too. For this, LOO CV will be run twice, the first one to detect the features selected by the algorithm most often and the second one to measure performance with fixed features to simulate an online evaluation. Fig. 6 shows the results obtained with the 2 datasets according to the threshold used to detect features containing too little information while Fig. 7 shows the percentage of feature number retained depending on the threshold used. Looking at the results, it is possible to observe that the threshold used has a moderate impact on the results. Indeed, compared to the results obtained in Fig 4 and Fig 5, the F1 score decreases by a maximum of only 5.5% and 2.6% respectively for H_2/O_2 and H_2/Air datasets.

It is possible to note that for both databases the maximum performance is reached using a limit of 5%. The algorithm succeeds in obtaining the same results as in Fig. 4 and Fig. 5. The limits of 10% and 1% lead to performance losses of about 5-3% for the H_2/O_2 database and 1.3 - 2.6% for the H_2/Air database. Even if the lost performances are quite low, this shows the importance of selecting the features correctly. Too many variables containing little information lead to an increase in the computation time as well as distortions within the health state space. On the contrary, a too-small space containing not enough information will not give good results.

The threshold of 5% allows obtaining the same performance (i.e. keeping only the most important information). In the framework of this study, a limit of 5% seems to fit well, it allows keeping only the variables containing the main information. In addition, it is worth noting that in the case of the 10% and 5% limits, the first 5 features are most often selected as opposed to the 1% limit which tends to add 2 other features. This shows that in general the most useful variables contain more than 10% of information but keeping the features containing between 5% and 10% of information allows having certain flexibility during the training which improves the final results.

In addition to the number of features, it is interesting to study which features are selected for both datasets. In the case of the H_2/O_2 dataset, the ones retained are *mp, Mp, Coefficient B, Coefficient A,* and *mm* where *Mm, R_Pola* are added when the 1% threshold is used. For the H_2/Air dataset, it is: *mp, Mp, Coefficient B, Coefficient A, ΔPha,* with *R_Pola* and *P1* if the 1% threshold is used.

It is interesting to note that several selected features are common to both databases: *mp, Mp, Coefficient B, and Coefficient A* (+ *R_Pola*). This shows that these features are relevant and allow good separation of the information. They provide respective information on the charge transfer of hydrogen oxidation reaction (*mp*), electrolyte membrane-related degradation (*Mp*), and the charge transfer of oxygen reduction reaction (*Coeff A & B*). According to Fig. 3.3 in [42], these four features give information at frequency situated in the starvation and water management conditions which are the

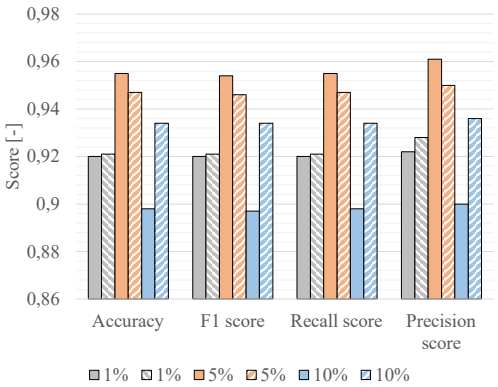

Fig. 6. Results obtained with automatic feature selection considering 3 information thresholds (1, 5 and 10%) using LOO CV

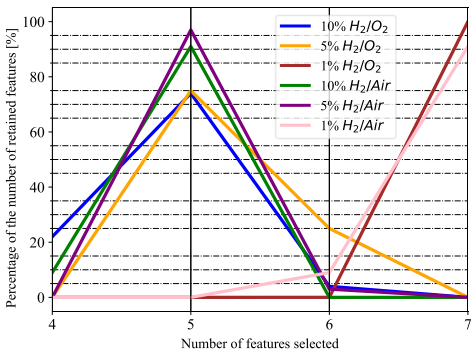

Fig. 7. Percentage of feature selected during the LOO CV depending of the threshold used

common faults between the two datasets.

Other features retained (i.e mm, *ΔPha*, *P1*, and Mm) give respectively information about the total ohmic resistance of the FC stack; the height of the phase spectra; diffusion phenomena; and are used with incomplete spectra when the imaginary axis is not crossed by EIS spectra. With both datasets, the two features added when the 1% threshold is used are redundant which justifies the loss of performance. Indeed, Mm and R_pola give similar information because all spectra don't cross the imaginary axis. Also, P1 in the H_2/Air dataset doesn't provide new information because it is calculated with Mp and *ΔPha* which are already selected.

D. Impact of clustering validity indices

Once the feature selection step is improved, it is interesting to focus on the clustering step. Indeed, the number of clusters defined by a user is limited by its knowledge of the database while the use of scores can allow detecting nuances that are invisible to the user. In this section, the cross-validation is run twice. For the first run, the algorithms are run with the automatization of feature selection and clustering steps to simulate the offline step. Then, the features and number of clusters are fixed, and the LOO CV process is run for a second time to simulate an online step. For each condition tested, the minimum number of clusters is fixed at $c_{min} = 2$ and the maximum number of clusters $c_{max} \cong \sqrt{n}$ which is considered a rule of thumb according to [38].

Fig. 8 and Fig. 9 present the most often number of clusters according to the cluster validity indices as well as the results obtained with the online step. Results show that several cluster validity indices provide good results close to the ones obtained when the true number of fault levels is used.

In the case of the H_2/O_2 dataset, V_{PE}, $V_{NPE,}$ and V_{PBMF} indices do not properly capture the separation between the data. They concatenate data in only two clusters for all conditions. In addition, they give the lowest performances in classification. The best performances are given by the V_{MPC} and V_{FS} indexes with a F1 score of about 0.95. Both methods detect more clusters than needed for nominal and flooding conditions. This can be explained by the fact that only 1 fault level is tested while

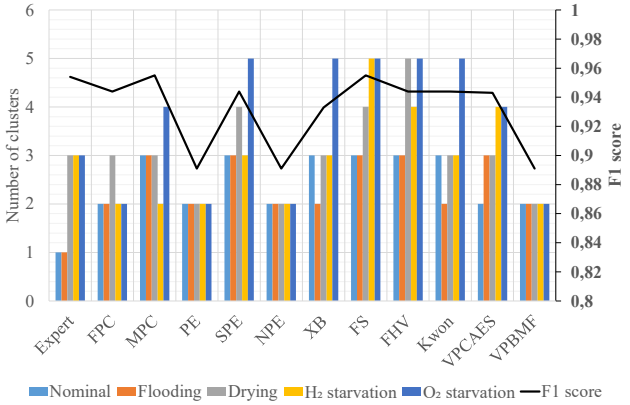

Fig. 8. Number of clusters selected and classifications performances according to the different clustering indices for the H_2/O_2 dataset

the minimum number of possible clusters is 2. However, for all

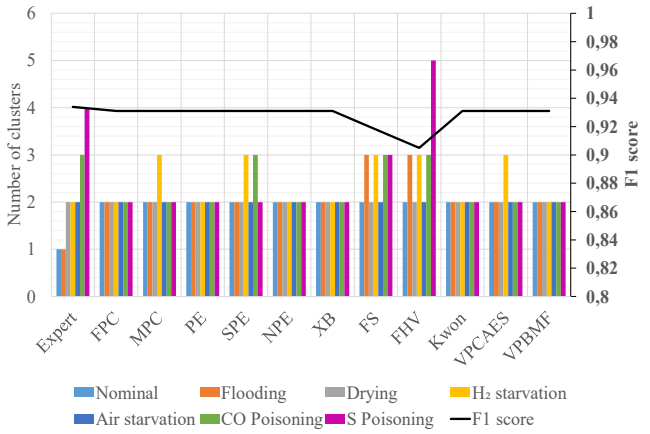

Fig. 9. Number of clusters selected and classifications performances according to the different clustering indices for the H_2/Air dataset

other conditions tested the V_{MPC} index correctly approximates the correct number of clusters (within ∓ 1 cluster). The V_{FS} index detects 2 and 1 too many clusters respectively for the starvations (H_2 and O_2) and drying conditions.

Regarding the H_2/Air dataset, the worst performances are given by V_{FS} and V_{FHV} with a decrease in performance of 1.5 and 2.8%. As with the H_2/O_2 database, they detect more clusters than necessary which shows a certain monotonic tendency that can be explained by a low amount of data. The V_{PE}, $V_{NPE,}$ and V_{PBMF} indices again detect 2 clusters for each condition as well as the V_{FPC}, V_{XB}, $V_{kwon,}$ and V_{PBMF}. However, they provide the same results as the ones given by V_{MPC}, $V_{SPE,}$ and V_{PCAES} (F1 score $\cong$ 0.93).

As the results above show, cluster validity can impact the performances of clustering algorithms. Too many clusters can generate more confusion between two conditions, while too few clusters can lead to not detecting a fault level which can generate confusion between the different conditions. In both databases, the V_{MPC} index provides performances similar to the ones obtained with the correct number of fault levels. V_{MPC} index is retained in the following of this paper. However, it is worth noting that the size of the two databases is relatively small and a similar study should be conducted with a larger sample size. Indeed, even if the best results are currently given by the V_{MPC} index, a more robust and complex index (e.g. V_{PCAES} or V_{PBMF}) can provide better performances when the database size is larger because they are robust against the monotonic tendency.

E. Analyze of misclassifications

To better measure the impact of the automation steps on classification performance, it is interesting to look at the classification errors. Table III highlights the confusion obtained using the expert-obtained parameters (*expert*) and the results obtained with the automated steps (*auto*). Results show that generally, the same confusions appear between the expert approach and the automatic one.

In the H_2/Air dataset, the confusions are mainly between the two poisoning faults which can be explained by the low severity

TABLE III: CONFUSIONS OBTAINED FOR THE TWO DATABASES

	TRUE CONDITION	*DETECTED CONDITION*	*NUMBER OF CONFUSIONS*
EXPERT H_2/O_2	O_2 starvation	H_2 starvation	4
AUTO H_2/O_2	O_2 starvation	H_2 starvation	1
	H_2 starvation	O_2 starvation	1
	Drying	Nominal	2
EXPERT H_2/AIR	Nominal	Drying	1
	Flooding	Drying	1
	CO Poisoning	S Poisoning	1
	S Poisoning	CO Poisoning	2
AUTO H_2/AIR	Nominal	Drying	2
	Flooding	Drying	1
	S Poisoning	CO Poisoning	2

of the fault condition. Both conditions have similar mechanisms at low intensity, so the features are similar.

There is also the presence of false positives linked to the drying condition. Indeed, 3 conditions are detected as drying while they were labeled as nominal and flooding. In this case, the confusion can be explained by the low severity of conditions combined with the small number of data which impacts the cluster centers calculation highly during the LOO process (3 – 8 – 6 for respectively nominal, flooding, and drying). Regarding the H_2/O_2 dataset, the same conditions have been confused (i.e. O_2 starvation and H_2 starvation) however, the automatic procedure generates confusion between drying and nominal conditions. Starvation conditions are easily confused due to the noise generated on spectra and their likeness. Drying confusions, as for the H_2/Air dataset, can be explained by a low fault level combined with a small number of data (i.e. 8 for nominal and 8 for weakly drying conditions). In both cases, the automatic parameter selection does not generate aberrant confusion and remains very close to the optimal results obtained with an expert study. To analyze and understand the confusion generated by the automatic selection of parameters, it is interesting to plot spectra in Nyquist diagrams. Fig. 10 shows the misclassified spectra for the H_2/O_2 and Fig. 11 for the H_2/Air dataset. Only low levels of poisoning are shown to improve the visibility of the graph. Moreover, no errors were detected for high levels of poisoning. It is possible to observe that all confusions are located at the intersections between 1 or more conditions. This confirms the difficulty of properly isolating the weak conditions because they are all located in the same area of SoH space.

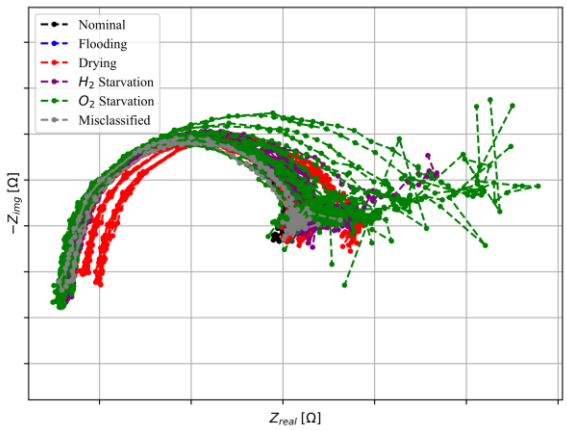

Fig. 10. Nyquist plots highlighting the misclassified EIS spectra using the automatic selection of parameters with the H_2/O_2 dataset (axes are hidden for confidentiality reasons)

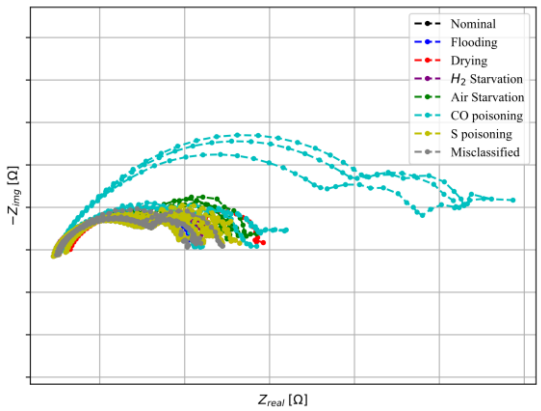

Fig. 11. Nyquist plots highlighting the misclassified EIS spectra using the automatic selection of parameters with the H_2/Air dataset (axes are hidden for confidentiality reasons)

F. Computation time measuring

Because this method needs to be easily implemented for practical use, it is necessary to test computation time on a low-cost embedded system. For this, a Raspberry Pi (RPi) Model B rev 2, with a 1 core 700 MHz BCM2835 CPU and 512 MB of RAM has been used. A comparison has been done with a computer equipped with an Intel(R) Core (TM) i7-8650U CPU @ 1.90GHz 2.11 GHz and 16Go of RAM to show the possible computation times both with a cheap system and a more powerful system. The algorithms have been run in LOO cross-validation 5 times. Training and prediction times have been measured for each loop of LOO CV and are shown in Fig. 12. It is possible to observe that considering the tested technologies, the execution times remain relatively low. Indeed, with a computer, the average training times are about 0.24 and 0.18 seconds for each database. Using RPi, these times increase to about 14 and 10 seconds. Given the specificities of the RPi system, these run times are normal although they are significantly longer than those of a computer. The average prediction time for a computer is 0.016 - 0.018 seconds and for the RPi is 0.04 - 0.14 seconds. Except for the execution time of the H_2/Air database on the RPi, the times are approximately the same between the two tested technologies. The increase in prediction time for the H_2/Air database can be explained by the additional CO classification step that is not done with the oxygen database used. In comparison with the training times given by more powerful diagnosis methods such as neural networks that take several minutes on a recent computer, this approach has the advantage to be efficient and easier to re-train. This shows the possibility of using this approach to regularly retrain the diagnostic algorithm with fresh data acquired online.

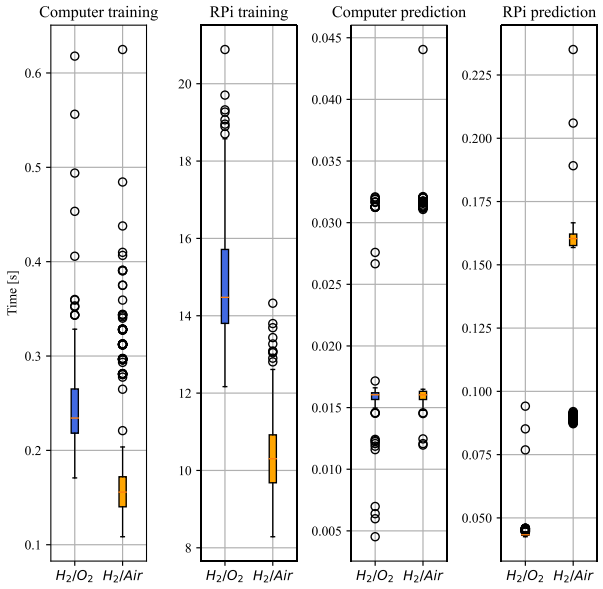

Fig. 12. Execution times of the algorithms implemented on a RPi system and a computer

VI. Conclusion

This paper presents an improved diagnosis approach based on EIS-extracted features and an automatic selection of the best parameters applied for PEMFC. Several standardizations have been studied to determine which ones can be considered to best generalize the method. The non-linear Yeo-Johnson and uniform quantile transformers produce excellent results (>93%) for classifying the spectra of two different databases at any point (noise, fault tested, experimentation materials …). To reduce the need for user expertise, several thresholds have been investigated to distinguish variables containing a significant amount of information and disturbing variables. It appeared that deleting the variables containing less than 5% of information allowed keeping the main information while deleting the disturbing variables. Also, the extracted variables are equal to those determined by the empirical method. The last investigation carried out in this article was to measure the impact of various clustering validation indices on classification rates. Good knowledge of a database is not an obvious thing to do this is why using metrics to automatically detect the right number of clusters is one of the studied axes. They allow the algorithm to adapt if data is added during the operation and to reduce the user's expertise. As a result, several indicators allow reaching the same performances as those obtained after the analysis of the databases, while granting a fast calculation time.

To reduce the measurement time of EIS spectra, future research will focus on reducing the number of frequencies to be used in the EIS measurement combined with the reconstruction of spectra by an equivalent circuit. This will reduce the experimental time without reducing the quality of the information that can be recovered to further improve the online implementation of the method.

Appendixes

A. Validation results for different standardization: H_2/O_2 database

TABLE IV: LOO CV VALIDATION RESULTS OBTAINED USING RAW DATA AND NORMALIZERS - H_2/O_2 DATASET

	Raw data	*Norm L2*	*Norm L1*	*Norm inf*
Accuracy [%]	85,2	78,4	78,4	73,9
F1 score [%]	85,2	77,9	77,9	73,5
Recall score [%]	85,2	78,4	78,4	73,9
Precision score [%]	85,9	79,6	79,6	76,2
Number of features	4	9	9	9

TABLE V: LOO CV VALIDATION RESULTS OBTAINED USING LINEAR SCALERS - H_2/O_2 DATASET

	Min-Max scaler	*Max absolute scaler*	*Standard scaler*	*Robust scaler*
Accuracy [%]	94,3	85,2	92,0	97,7
F1 score [%]	94,3	85,3	92,0	97,7
Recall score [%]	94,3	85,2	92,0	97,7
Precision score [%]	94,7	85,6	92,2	97,9
Number of features	5	5	5	6

TABLE VI: LOO CV VALIDATION RESULTS OBTAINED USING NON LINEAR TRANSFORMERS - H_2/O_2 DATASET

	Yeo-Johnson	*Normal Quantile*	*Uniform Quantile*
Accuracy [%]	96,6	94,3	95,5
F1 score [%]	96,6	94,3	95,4
Recall score [%]	96,6	94,3	95,5
Precision score [%]	96,6	94,8	96,1
Number of features	6	6	5

B. Validation results for different standardization: H_2/AIR database

TABLE VII: LOO CV VALIDATION RESULTS OBTAINED USING RAW DATA AND NORMALIZERS - H_2/AIR DATASET

	Raw data	*Norm L2*	*Norm L1*	*Norm inf*
Accuracy [%]	77,6	84,2	85,5	86,8
F1 score [%]	77,4	84,9	85,7	86,4
Recall score [%]	77,6	84,2	85,5	86,8
Precision score [%]	81,6	85,9	86,2	86,6
Number of features	5	6	6	5

TABLE VIII: LOO CV VALIDATION RESULTS OBTAINED USING LINEAR SCALERS - H_2/AIR DATASET

	Min-Max scaler	*Max Absolute scaler*	*Standard scaler*	*Robust scaler*
Accuracy [%]	85,5	82,9	82,9	81,6
F1 score [%]	85,7	03,1	82,7	82,1
Recall score [%]	85,5	82,9	82,9	81,6
Precision score [%]	86,2	83,6	83,0	83,6
Number of features	4	5	5	4

TABLE IX: LOO CV VALIDATION RESULTS OBTAINED USING NON LINEAR TRANSFORMERS - H_2/AIR DATASET

	Yeo-Johnson	*Normal Quantile*	*Uniform Quantile*
Accuracy [%]	93,4	88,2	94,7
F1 score [%]	93,4	88,3	94,6
Recall score [%]	93,4	88,2	94,7
Precision score [%]	93,8	88,8	95,0
Number of features	5	8	5

ACKNOWLEDGMENT

This project has received funding from the Fuel Cells and Hydrogen 2 Joint Undertaking (JU) under grant agreement No 875047 Website: https://www.rubyproject.eu/.

This work has been supported by the EIPHI Graduate School (contract ANR-17- EURE-0002) and the Region Bourgogne Franche-Comté.

We acknowledge the European project HEALTH CODE which provides the data used in this paper. Website: http://pemfc. health-code.eu/.

REFERENCES

[1] "Fuel Cells | Department of Energy." https://www.energy.gov/eere/fuelcells/fuel-cells (accessed Mar. 11, 2022).

[2] R. I. Salim, H. Noura, and A. Fardoun, "A review on fault diagnosis tools of the proton exchange Membrane Fuel Cell," in *2013 Conference on Control and Fault-Tolerant Systems (SysTol)*, Oct. 2013, pp. 686–693. doi: 10.1109/SysTol.2013.6693877.

[3] J. Aubry, N. Y. Steiner, S. Morando, N. Zerhouni, and D. Hissel, "Fuel cell diagnosis methods for embedded automotive applications," *Energy Reports*, vol. 8, pp. 6687–6706, Nov. 2022, doi: 10.1016/j.egyr.2022.05.036.

[4] A. Narjiss, D. Depernet, D. Candusso, F. Gustin, and D. Hissel, "On-line diagnosis of a PEM fuel cell through the PWM converter," Dec. 2008.

[5] "Real operation pem fuel cells HEALTH-state monitoring and diagnosis based on dc-dc COnverter embeddeD Eis;H2020, European project, Horizon 2020; Health-Code." http://health-code.eu/ (accessed May 04, 2021).

[6] "RUBY – EU project." https://www.rubyproject.eu/ (accessed Feb. 02, 2021).

[7] R. Petrone *et al.*, "A review on model-based diagnosis methodologies for PEMFCs," *International Journal of Hydrogen Energy*, vol. 38, no. 17, pp. 7077–7091, Jun. 2013, doi: 10.1016/j.ijhydene.2013.03.106.

[8] Z. Zheng *et al.*, "A review on non-model based diagnosis methodologies for PEM fuel cell stacks and systems," *International Journal of Hydrogen Energy*, vol. 38, no. 21, pp. 8914–8926, Jul. 2013, doi: 10.1016/j.ijhydene.2013.04.007.

[9] "BatteryArchive.org." https://www.batteryarchive.org/index.html (accessed Jul. 01, 2022).

[10] Z. Zheng, M.-C. Péra, D. Hissel, M. Becherif, K.-S. Agbli, and Y. Li, "A double-fuzzy diagnostic methodology dedicated to online fault diagnosis of proton exchange membrane fuel cell stacks," *Journal of Power Sources*, vol. 271, pp. 570–581, Dec. 2014, doi: 10.1016/j.jpowsour.2014.07.157.

[11] L. A. M. Riascos, M. G. Simoes, and P. E. Miyagi, "A Bayesian network fault diagnostic system for proton exchange membrane fuel cells," *Journal of Power Sources*, vol. 165, no. 1, pp. 267–278, Feb. 2007, doi: 10.1016/j.jpowsour.2006.12.003.

[12] Z. Li *et al.*, "Online implementation of SVM based fault diagnosis strategy for PEMFC systems," *Applied Energy*, vol. 164, pp. 284–293, Feb. 2016, doi: 10.1016/j.apenergy.2015.11.060.

[13] A. Escobet, À. Nebot, and F. Mugica, "PEM fuel cell fault diagnosis via a hybrid methodology based on fuzzy and pattern recognition techniques," *Engineering Applications of Artificial Intelligence*, vol. 36, pp. 40–53, Nov. 2014, doi: 10.1016/j.engappai.2014.07.008.

[14] D. Chanal, N. Yousfi-Steiner, R. Petrone, D. Chamagne, and M.-C. Péra, "Online Diagnosis of PEM Fuel Cell by Fuzzy C-means clustering," *Encyclopedia of Energy Storage*, p. 41.

[15] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*, p. 6.

[16] A. Zheng and A. Casari, *Feature Engineering for Machine Learning*. O'Reilly Media, Inc, USA, 2018.

[17] C. M. Bishop, "Neural Networks for Pattern Recognition," p. 498.

[18] J. Brownlee, *Data Preparation for Machine Learning Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery, 2020.

[19] G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.

[20] I.-K. Yeo and R. A. Johnson, "A New Family of Power Transformations to Improve Normality or Symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000.

[21] M. Bicego and S. Baldo, "Properties of the Box–Cox transformation for pattern classification," *Neurocomputing*, vol. 218, pp. 390–400, Dec. 2016, doi: 10.1016/j.neucom.2016.08.081.

[22] T. J. Sefara, "The Effects of Normalisation Methods on Speech Emotion Recognition," in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, Nov. 2019, pp. 1–8. doi: 10.1109/IMITEC45504.2019.9015895.

[23] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Aug. 2020, pp. 729–735. doi: 10.1109/ICSSIT48917.2020.9214160.

[24] T. J. Ross, "Fuzzy Logic with Engineering Applications, 3rd Edition | Wiley," *Wiley.com*. https://www.wiley.com/en-us/Fuzzy+Logic+with+Engineering+Applications%2C+3rd+Edition-p-9780470743768 (accessed Jun. 24, 2022).

[25] D.-W. Kim, K. H. Lee, and D. Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," *Pattern Recognition*, vol. 37, no. 10, pp. 2009–2025, Oct. 2004, doi: 10.1016/j.patcog.2004.04.007.

[26] Y. Hu, C. Zuo, Y. Yang, and F. Qu, "A robust cluster validity index for fuzzy c-means clustering," in *Proceedings 2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, Dec. 2011, pp. 448–451. doi: 10.1109/TMEE.2011.6199238.

[27] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *J. Math. Biology*, vol. 1, no. 1, pp. 57–71, May 1974, doi: 10.1007/BF02339490.

[28] J. C. Bezdek†, "Cluster Validity with Fuzzy Sets," *Journal of Cybernetics*, vol. 3, no. 3, pp. 58–73, Jan. 1973, doi: 10.1080/01969727308546047.

[29] R. N. Dave, "Validating fuzzy partitions obtained through c-shells clustering," *Pattern Recognition Letters*, vol. 17, no. 6, pp. 613–623, May 1996, doi: 10.1016/0167-8655(96)00026-8.

[30] J. Bezdek, "Mathematical Models for Systematics and Taxonomy," Jan. 1975.

[31] J. C. Bezdek, M. P. Windham, and R. Ehrlich, "Statistical parameters

of cluster validity functionals," *International Journal of Computer and Information Sciences*, vol. 9, no. 4, pp. 323–336, Aug. 1980, doi: 10.1007/BF00978164.

[32] Y. Fukuyama, "A new method of choosing the number of clusters for the fuzzy c-mean method," *undefined*, 1989, Accessed: Jun. 25, 2022. [Online]. Available: https://www.semanticscholar.org/paper/A-new-method-of-choosing-the-number-of-clusters-for-Fukuyama/a79b0c54ffada673179d17fad783872309620771

[33] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 773–780, Jul. 1989, doi: 10.1109/34.192473.

[34] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, Aug. 1991, doi: 10.1109/34.85677.

[35] S. H. Kwon, "Cluster validity index for fuzzy clustering," *Electronics Letters*, vol. 34, no. 22, pp. 2176–2177, Oct. 1998, doi: 10.1049/el:19981523.

[36] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, vol. 37, no. 3, pp. 487–501, Mar. 2004, doi: 10.1016/j.patcog.2003.06.005.

[37] K.-L. Wu and M.-S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1275–1291, Jul. 2005, doi: 10.1016/j.patrec.2004.11.022.

[38] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, Aug. 1995, doi: 10.1109/91.413225.

[39] pal, "On cluster validity for the fuzzy c-means model | IEEE Journals & Magazine | IEEE Xplore." https://ieeexplore.ieee.org/document/413225 (accessed Jun. 26, 2022).

[40] M. H. F. Zarandi, M. R. Faraji, and M. Karbasian, "An Exponential Cluster Validity Index for Fuzzy Clustering with Crisp and Fuzzy Data," p. 16.

[41] D. Chanal, N. Y. Steiner, D. Chamagne, and M.-C. Pera, "Impact of standardization applied to the diagnosis of LT-PEMFC by Fuzzy C-Means clustering," in *2021 IEEE Vehicle Power and Propulsion Conference (VPPC)*, Oct. 2021, pp. 1–6. doi: 10.1109/VPPC53923.2021.9699234.

[42] H. Wang, X.-Z. Yuan, and H. Li, Eds., *PEM Fuel Cell Diagnostic Tools*, 1st ed. CRC Press, 2017.

AUTHORS INFORMATION

Damien CHANAL is a Ph.D. student at Femto-ST Institute and the UAR FCLAB. His Ph.D. thesis is part of a European collaboration in which he is in charge of diagnostic and prognostic tasks to improve the performance of PEM fuel cells. He has a master's degree in thermal and energy engineering from the University of Franche-Comté in Belfort. He has a specialization in the field of hydrogen and energy efficiency with a Master of Engineering: CMI Hydrogen and Energy Efficiency course. In terms of publication, Damien is the first author of a chapter named "Online Diagnosis of PEM Fuel Cell by Fuzzy C-Means Clustering" which is published in the "Encyclopedia of Energy Storage". In addition, he is the first author of 2 international conference papers named "Impact of standardization applied to the diagnosis of LT-PEMFC by Fuzzy C-Means clustering" (IEE VPPC 2021) and "Voltage prognosis of PEMFC estimated using Multi-Reservoir Bidirectional Echo State Network" (IEE ICSC 2022).

Nadia YOUSFI STEINER received a master's degree in mathematics and a master's degree in Fluidics and Energetics in 2006. She obtained a Ph.D. in Engineering Science in collaboration between the University of Franche-Comté and the European Institute for Energy Research in Karlsruhe, Germany in 2009. From 2009 to 2014, she worked as R&D Project Manager in charge of collaborative projects on Hydrogen and Fuel Cells in Germany.

Her research deals with Fuel Cell systems characterization diagnostics, prognostics and Fault Tolerant Control. She is currently Full Professor at the University of Franche-Comté and held a 6-year Research Chair of excellence within the Energy Department in Belfort, France.

Didier CHAMAGNE obtained a Ph.D. from the University of Franche-Comte, in 1991. These activities of research are made in the FEMTO-ST Institute / Energy Department (UMR CNRS 6174). It concerns the design - modeling and optimization of systems of energy conversion as part of a multiphysical approach. The research works concern three themes: optimization of energy systems - design and optimization of innovative machines for the hybrid vehicles and for the energy conversion - diagnostic and prognostic of PEM fuel cells. Since 2008, he is a Full Professor at the University of Franche-Comte.

Marie-Cécile Péra received a Ph.D. in electrical engineering in 1993. From 1994 to 1999, she was an Associate Professor at the University of Reims Champagne Ardennes, where she studied non-linear dynamics of electrical systems, based on chaos theory. Since 1999, she has joined the University of Franche-Comte (UFC) where she launched the activities on Fuel Cell Systems and became a full Professor From 2012 to 2019, she was the Deputy Director of the FEMTO-ST Institute (750 members). Since 2020, she is the Director of FCLAB, Center for Service and Research (140 members). She is the chair of the board of the CoNRS, National Council of CNRS, for photonics and electrical engineering. She works on energy management of hybrid electric power generation systems (fuel cells, PEMFC and SOFC, supercapacities, batteries), their control, diagnosis and prognostics. She has contributed to more than 275 publications in peer-reviewed international journals and international conferences.