IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR ELECTRIC POWER SYSTEMS

Received December 6, 2019, accepted December 24, 2019, date of publication December 27, 2019, date of current version January 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962775

Deep Convolutional Variational Autoencoder as a 2D-Visualization Tool for Partial Discharge Source Classification in Hydrogenerators

RYAD ZEMOURI[1,2,3], MÉLANIE LÉVESQUE[2], NORMAND AMYOT[2], CLAUDE HUDON[2], OLIVIER KOKOKO[2], AND ANTOINE TAHAN[3]

[1]CEDRIC Laboratory of the Conservatoire National des Arts et Métiers (CNAM), HESAM Université, 750141 Paris Cedex 03, France
[2]Institut de Recherche d'Hydro-Québec (IREQ), Varennes, QC J3X 1S1, Canada
[3]École de Technologie Supérieure, Montréal, QC H3C 1K3, Canada

Corresponding author: Ryad Zemouri (ryad.zemouri@cnam.fr)

This work was supported by the Institut de Recherche d'Hydro-Québec (IREQ).

ABSTRACT Hydrogenerators are strategic assets for power utilities. Their reliability and availability can lead to significant benefits. For decades, monitoring and diagnosis of hydrogenerators have been at the core of maintenance strategies. A significant part of generator diagnosis relies on Partial Discharge (PD) measurements, because the main cause of hydrogenerator breakdown comes from failure of its high voltage stator, which is a major component of hydrogenerators. A study of all stator failure mechanisms reveals that more than 85 % of them involve the presence of PD activity. PD signal can be detected from the lead of the hydrogenerator while it is running, thus allowing for on-line diagnosis. Hydro-Québec has been collecting more than 33 000 unlabeled PD measurement files over the last decades. Up to now, this diagnostic technique has been quantified based on global PD amplitudes and integrated PD energy irrespective of the source of the PD signal. Several PD sources exist and they all have different relative risk, but in order to recognize the nature of the PD, or its source, the judgement of experts is required. In this paper, we propose a new method based on visual data analysis to build a PD source classifier with a minimum of labeled data. A convolutional variational autoencoder has been used to help experts to visually select the best training data set in order to improve the performances of the PD source classifier.

INDEX TERMS Hydrogenerators, diagnosis, partial discharges, deep neural networks, convolutional variational autoencoder, data visualization, feature extraction, model interpretation, generative model.

I. INTRODUCTION

One of the main problems that all industries face is the massive high dimensionality unlabeled data. Artificial Neural Networks (ANNs) and Deep Learning (DL) are actually the leading machine-learning tools for intelligent condition monitoring and diagnosis used for mechanical systems. However, a major assumption accepted by default, is that the training and testing data are taking from same feature distribution [1]. Like many other utilities, Hydro-Québec which has an electric generating capacity of 36 GW from its 62 hydroelectric power plants, has collected a large number of measurement files. Each hydrogenerator is worth several million to tens of millions of dollars and is subject to preventive maintenance comprising both systematic and conditional maintenance.

The associate editor coordinating the review of this manuscript and approving it for publication was Canbing Li.

Availability, reliability and durability of hydrogenerators are key features that have driven electrical utilities to implement monitoring and diagnostic methods in order to evolve toward Condition Based Maintenance (CBM). More than ten years ago, Hydro-Quebec has implemented a web-based application, MIDA (Methodology for Integrated Diagnostic of hydrogenerators [2]), which was developed at its research institute. MIDA is an integrated diagnostic system for hydrogenerators and is based on the aggregation of individual health indexes from seven diagnostic tools [2].

MIDA gives a ranking of more than 300 hydrogenerators and thus helps maintenance engineers to better plan maintenance actions. The MIDA centralized database contains all the data from every diagnostic measurement performed on each hydrogenerator since the 1990's. The diagnostic data from MIDA can then be used to identify symptoms of physical degradation states involved in failure mechanism and

is used as input in a Failure Mode and Symptom Analysis (FMSA) approach applied to hydrogenerator prognostics [3].

Data from MIDA used in conjunction with PHM techniques can feed a prognostic model that will provide useful equipment information and lead to the implementation of predictive maintenance. However, diagnostics need to improve in order to generate a more detailed information input for the prognostic model. One of these diagnostic tools on which we will focus in this paper is PD measurements. The prognostic model is built on more than 100 failure mechanisms for the stator, the main component of hydrogenerators. These mechanisms are consigned in the form of causal trees or graphs [4]. A large number of these failure mechanisms involve the presence of PD [3]. At Hydro-Quebec, PD measurements on hydrogenerators have been carried out for the past 30 years and this extensive PD database is integrated in MIDA. As of December 2019, MIDA has more than 33 000 unlabeled PD measurement files. Each of these files must be analyzed and classified by the experts of Hydro-Quebec for PD source recognition which is very time-consuming and can only be done for PD files easily recognizable by the expert. A structured analysis of this huge amount of data is of paramount importance to understand the behavior and evolution of the PD activity. The implementation of an automatic recognition based on an intelligent classification of all PD sources would significantly improve diagnostic and prognostic models for hydrogenerators.

However, to perform such an intelligent classification, experts must first label a number of PD measurement files for the training process. To do this, experts faced two challenging problems: how to select the most significant PD data for labeling, and what is the minimum data size required to complete the training process and which would be sufficient to define each class? One solution to help the expert choosing new training data is to use active learning [5], [6]. Active learning is an efficient machine learning technique that can simultaneously improve the quality of the classification model and decrease the complexity of training samples. It is frequently deployed in scenarios where large scale data sets are easily collected, but labeling them is expensive and/or time-consuming [7]. By using active learning, a classification model can iteratively interact with experts to only select the most significant instances for labeling and to further promote its performance as quickly as possible [7]. However, several previous studies have indicated that the performance of active learning can be easily disrupted if the data set has an imbalanced distribution [7]. Many research works have been recently developed to improve the performance of learning from imbalanced data [8]–[12]. However, none of these methods relies on a visual analysis of the learning data.

One of the first research dealing with feature selection based on 2D-data visualization has been developed by Gill *et al.* [13]. The objective of this study was to enhance the health-monitoring system for helicopters by visual data analyzing of the structural vibrations, in order to recognize different flight conditions directly from sensor information. As introduced by [13], we consider that the visualization of the learning data to understand their nature and to extract more information from the 2D-space distribution is an important step during the conception of the diagnosis model.

Otherwise, it is usually considered that classification improves with the number of labeled data, but this is not always true. In fact, it depends on the spatial distribution of data used in training process. Figure 1 gives a basic illustration of two different training data sets in a 2D-dimensional space. The first one has more data points than the second data set, but the second data set is better dispersed than the first one. The grey zone represents the conflict area, which is the boundary between classes, where most of the false positive predictions are produced by the classifier. Better is the spatial distribution of the training data set, better is the accuracy of the prediction. To reduce this ''dead zone'', the expert must choose new data points to label them as belonging to or as located nearby this conflict area.

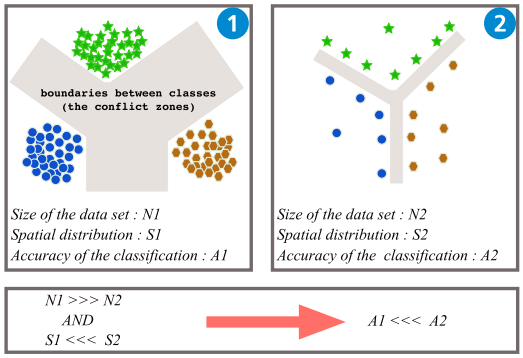


FIGURE 1. Basic illustration of an arbitrary 2D-representation of two different training data sets with two different spatial distributions. The data set #1 has more samples than the data set #2 but the accuracy of the classification is better for the data set #2.

Another reason that motivates our work is the lack of interpretability obtained by the ANNs. It is very hard to understand what happens in the hidden layers and why a trained NN gives a positive diagnosis for a given input sample. This ''black-box'' aspect is very restrictive in many application fields, where the interpretation of a decision can lead to serious legal consequences especially in safety-critical applications [14] (e.g., medical diagnosis [15], [16], autonomous driving, electric power generation. . . etc.) When the Convolutional Neural Networks (CNNs) are used for image processing, several methods have been developed to visualize what happens in the intermediate layers. Some of these algorithms are for example visual explanations from CNNs via gradient-based localization [17]; a visualization technique of the input stimuli that excite individual feature maps at any layer in a CNN model [18] or; a deep Taylor decomposition method for interpreting generic multilayer NNs by decomposing the network classification decision into contributions of its input elements [19]. When the input data are not images, as for industrial data, the interpretability of the hidden layer's activities is less obvious.

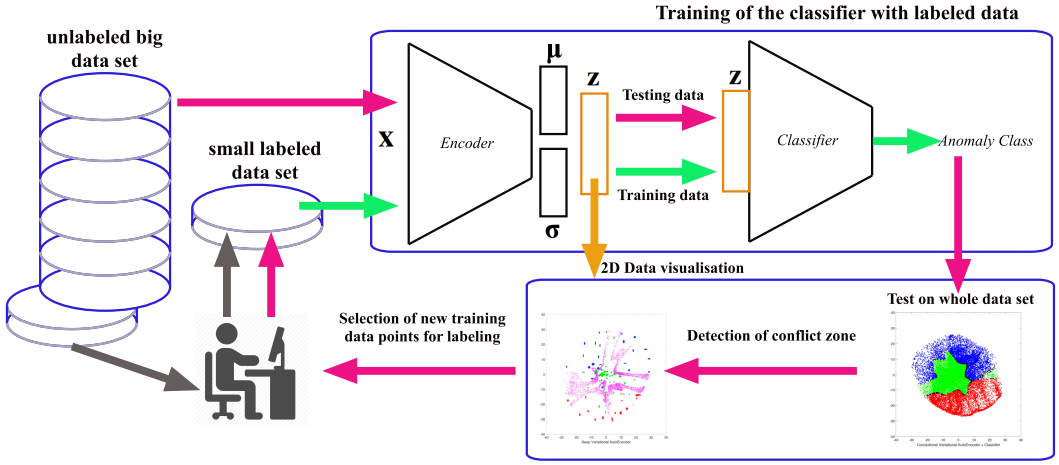


FIGURE 2. The framework of the proposed PD source classification methodology using an encoder function of a CVAE as a data projection for a 2D-visualization. The input vectors are encoded into a 2D-latent space. If the conflict zone, identified on the 2D-visualization, is too large, new data nearby these conflict zones are selected and labeled by experts. These new labeled data are then added to the initial training data set for a new training iteration.

Some visualization techniques, as the t-distributed stochastic neighbor embedding projection (t-SNE) [20], converts a high-dimensional data set into a 2D-matrix of pairwise similarities. Feature maps of the model are then obtained. The t-SNE method has been used and reported in many research publications such as [8], [14], [21]–[25]. The main constraint of this method is the lack of repeatability due to the minimization of the Kullback-Leibler divergence between the input space distribution and the embedding space distribution [22].

In this paper, a new method based on a visual data analysis to build a PD source classifier with a minimum of labeled data files is proposed. The framework of this methodology for PD source classification is given in Fig. 2. The encoder, part of a Convolutional Variational Autoencoder (CVAE), is used for data projection in a 2D-visualization latent space. The input vectors are encoded and displayed into this 2D-space, which helps the expert to visually analyze the spatial distribution of the training data set. At the beginning, few data files from the huge unlabeled database are selected from the 2D-latent space. These data files are labeled by experts and used to train a neural network classifier. The obtained classifier is then tested over all unlabeled data set. To identify conflict areas, i.e., the gray zones illustrated in Fig. 1, several classifiers are trained on the same labeled data set and tested using the entire MIDA unlabeled data set. The resulting conflict areas are identified on the 2D-space by analyzing conflicting results between classifiers. As it can be seen in Fig. 1, if the conflict zone is too large, the class boundary defined by the classifier, is more uncertain, which leads to an erroneous classification near this zone. New data files from these conflict zones are then selected and labeled by experts in order to reduce the conflict area. These new labeled data files are then added to the initial training data set for a new training iteration. Therefore, several iterations between the PD experts and DL experts are necessary to refine the input vector, reduce the boundary of conflict areas and improve classification of PD source. This focuses the work on a limited number of data files out of the overwhelming existing data.

The paper will focus in section 2 on the description of how an Autoencoder (AE) and a Variational Autoencoder (VAE) can be used to resolve real industrial challenges by classifying PD sources obtained from hydrogenerators. Thus, minimal background on partial discharge will be given in section 3, in order to understand how PD source classification would be a major breakthrough in this field. A description of PD phenomena, their long term deleterious effect leading to premature failure and the definition of the feature vector will be discussed. In section 4, the convolutional variational autoencoder used jointly with a neural network classifier is described. The way that conflict areas are identified and treated will be developed. Then, all the visual analysis obtained by the encoder are illustrated in section 5. Some confrontation results between the architecture used (i.e., the encoder part of the CVAE used jointly with a classifier) and a traditional classifier without the convolutional encoder are also presented. Finally, a discussion and a conclusion are given in sections 6 and 7.

II. VARIATIONAL AUTOENCODERS

An AE is an unsupervised NN trained to recreate or reproduce the input vector x [23], [26]–[28]. The AE is composed by two main structures: an encoder and a decoder (Fig. 3) which are multilayered NNs parameterized by ϕ and θ, respectively. The first one encodes the input data x into a latent representation z by the encoder function $z = f_\phi(x)$, whereas the second one decodes this latent representation onto $\hat{x} = h_\theta(z)$ which is an approximation or reconstruction of the original data. In an AE, an equal number of units are used in the input/output layers while less units are used in the latent space (Fig. 3). The AEs are usually used for data compression (i.e., feature extraction/reduction), noise removal and pre-trained

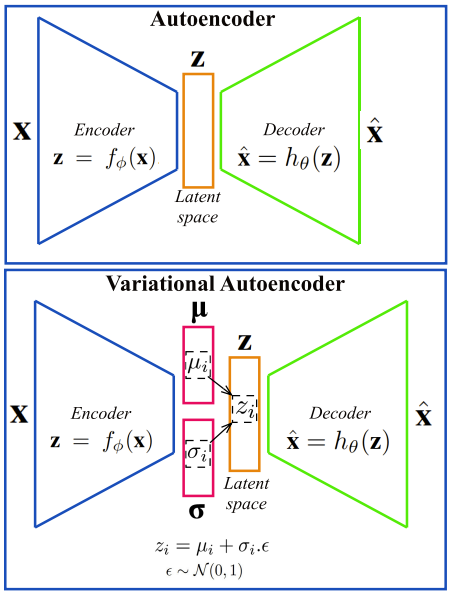


FIGURE 3. Schematic architecture of a standard deep autoencoder and a variational deep autoencoder. Both architectures have two parts: an encoder and a decoder.

parameters for a complex network.

A VAE has the same functions as the AE in the sense that it is composed by an encoder and a decoder (Fig. 3). VAE becomes a popular generative model by combining Bayesian inference and the efficiency of the NNs to obtain a nonlinear low-dimensional latent space [29]–[32]. The Bayesian inference is obtained by an additional layer used for sampling the latent vector z with a prior specified distribution $p(z)$, usually assumed to be a standard Gaussian $\mathcal{N}(0, I)$, where I is the identity matrix. Each element z_i of the latent layer is obtained as follow:

$$z_i = \mu_i + \sigma_i \cdot \epsilon \tag{1}$$

where μ_i and σ_i are the i^{th} components of the mean and standard deviation vectors, ϵ is a random variable following a standard Normal distribution ($\epsilon \sim \mathcal{N}(0, 1)$). Unlike the AE which generates the latent vector z, the VAE generates vector of means μ_i and standard deviations σ_i. This allows to have more continuity in the latent space than the original AE. The VAE loss function given by the equation 2 has two terms. The first term $\mathcal{L}_{rec}$ is the reconstruction loss function (equ. 3). Usually the negative expected log-likelihood (e.g., the cross-entropy function) is used ([30], [31], [33]–[35]) but the mean squared error [32] can also be used. The second term $\mathcal{L}_{KL}$ (equ. 4) corresponds to the Kullback-Liebler (KL) divergence loss term that forces the generation of a latent vector with the specified Normal distribution [36], [37]. The KL divergence is a theoretical measure of proximity between two densities $q(x)$ and $p(x)$. It is asymmetric ($KL(q \parallel p) \neq KL(p \parallel q)$) and nonnegative. It is minimized when $q(x) = p(x)$ [38]. Thus, the KL divergence term measures how close is the conditional distribution density $q_\phi(z \mid x)$ of the encoded latent vectors from the desired Normal distribution $p(z)$. The value of KL is zero when two probability distributions are the same, which forces the encoder of VAE $q_\phi(z \mid x)$ to learn the latent variables that follow a multivariate normal distribution over a k-dimensional latent space.

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{KL} \tag{2}$$

$$\mathcal{L}_{rec} = -E_{q_\phi(z|x)}(log(p_\theta(x \mid z))) \tag{3}$$

$$\mathcal{L}_{KL} = KL(q_\phi(z \mid x) \parallel p(z)) \tag{4}$$

When the VAE is trained, each function (i.e., the encoder and the decoder) can be used separately, either to reduce the space dimension by encoding the input data, or to generate synthetic samples by decoding new variables from the latent space (Fig. 4). Most of the generative applications deal with image processing as in [33] where a VAE was trained to generate face images with much clearer and more natural noses, eyes, teeth, hair textures as well as reasonable backgrounds. In [30], a generative model is constructed to create new random realizations of faces that are indistinguishable from samples.

In nonlinear processes monitoring, VAE have been recently used for high-dimensional process fault diagnosis. The most relevant characteristics of the process are extracted by the latent variable space by projecting the high-dimensional process data into a lower-dimensional space [8], [29], [31], [32], [34], [39]–[42].

III. PARTIAL DISCHARGE ANALYSIS

PD are minute sparks that occur within voids inside high voltage insulation or in the air around the insulating system. Each PD event does not cause immediate failure, but it will slowly erode the insulation system and will lead to breakdown in years to decades [43], [44]. The impulses can be detected on-line from sensors connected to hydrogenerators, which is a major advantage, because diagnosis of upcoming problems is possible while the machine is running and generating power. Over the past 30 years, Hydro-Québec has gathered an extensive PD database using two types of commercial measurement instruments. One of the instruments used is a 2D Partial Discharge Analyzer (PDA), which displays the rate of discharge pulses as a function of their amplitude. Up to now, over 33 000 measurement files have been recorded yearly by plant personnel. Fig. 5 gives a representation of the PD activity measured using the 2D PDA technique. Here the graph shows the discharge pulse rate (PD/s) as a function of the amplitude (in mV) and where positive pulses are in red and the negative discharges are in yellow. The second type of instrument used is a 3D system with much greater PD recognition capabilities.

In hydrogenerator diagnostic, it is important to recognize normal internal PD giving symmetrical distribution of positive and negative PD from other type of discharge sources [44]. Most measurement systems only rely on the maximum PD amplitude and it's trending for quantification, regardless of the type of PD source. Even though it is well

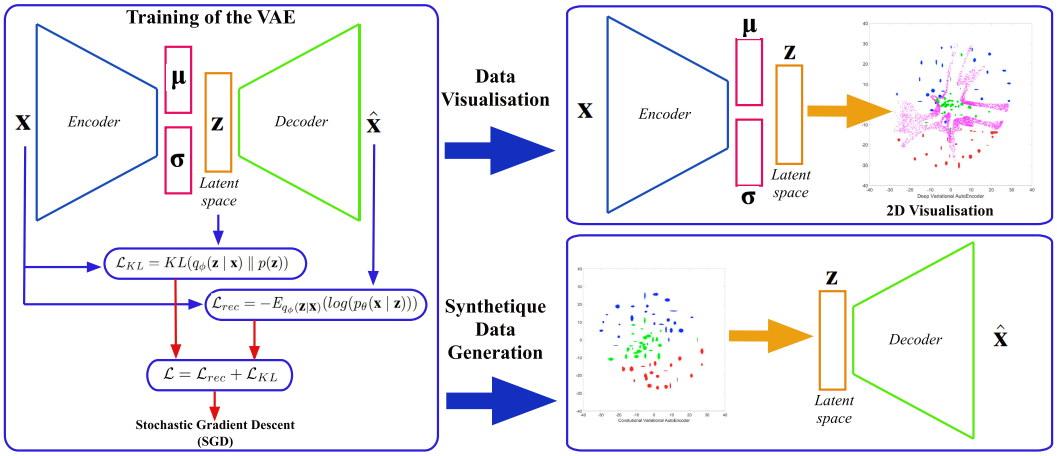


FIGURE 4. The VAE loss function. The first term $\mathcal{L}_{rec}$ is the reconstruction loss function. The second term $\mathcal{L}_{KL}$ corresponds to the Kullback-Liebler divergence loss term that forces the generation of a latent vector with the specified Normal distribution. When the VAE is trained, the two functions encoder/decoder can be used separately even to reduce the space dimension by encoding the input data or to generate synthetic samples by decoding new variables from the latent space.

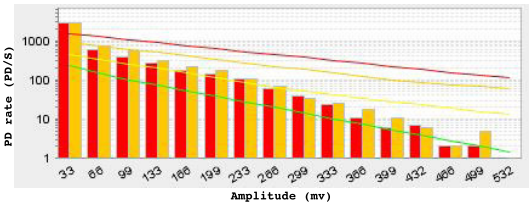


FIGURE 5. A simple 2D representation of the PD showing the discharge rate (PD/s) as a function of 16 channels of amplitudes (mV) for positive PD are (red) and negative PD (yellow).

recognized in the industry that different types of PD sources represent different risks for the equipment. Some PD sources may lead to failure after only 10 years of operation, for a machine that would typically last 45 years or more. One of the reasons that quantification is not done for each type of PD independently, is that there is no easy way to automatically recognize all PD signatures. To do this, it would require having classification rules that work for all files. Expert rules used up to now, consider asymmetries such as the one of slot PD giving more activity for positive PD or in the case of decohesion of insulation at the high voltage conductor giving the opposite asymmetry with more negative PD [44].

A ratio of 2 between the two PD polarities is a clear indication of asymmetry, but as this ratio reduces, experts should analyze each PD signature, channel by channel, and exception are more common than PD signature respecting the expert rules.

Another type of PD source called gap type discharge takes place in the end winding outside the magnetic core of the stator. Because of their location and signal propagation mode, these PDs experience very little attenuation. Thus, they reach the detection point with apparent amplitudes much greater than those for other types of PD. However, they do not represent a higher risk as would suggest their high amplitude. Since most systems tend to quantify the severity of PDs by using their amplitude, these gap discharges should not be rated on the same scale as other PD sources.

The logical solution would be to have different evaluation criteria for different discharge sources, but because of the overwhelming quantity of data, it is currently impossible to sort all PD signatures. However, the current expert rules can be used as a starting point to identify easily recognizable PD signatures with distinctive asymmetry or other features. In order to do this, the PD distribution in Fig. 5 can be described by the following vector:

$$PD = \begin{pmatrix} pd_1^x, \ldots, pd_i^x, \ldots, pd_{16}^x \\ pd_1^y, \ldots, pd_i^y, \ldots, pd_{16}^y \end{pmatrix} \quad (5)$$

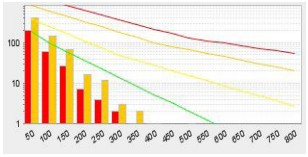
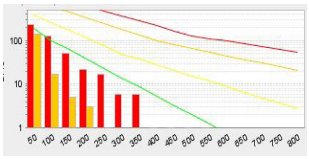
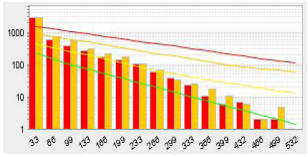
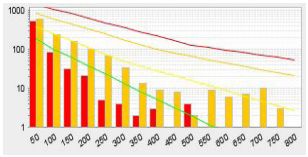
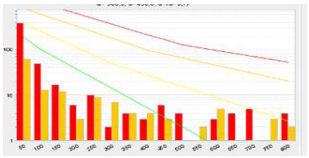
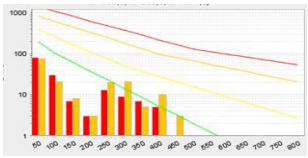
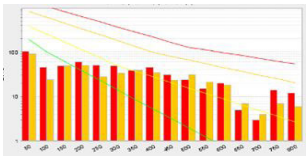
where pd_i^x and pd_i^y are respectively the positive and negative discharge rate in PD/s in each amplitude channel i of the horizontal axis in Fig. 5 (where i=1 to 16).

Thus, the rules provided by experts can be used to define the seven following classes. Table 1 gives a representative illustration of each PD source:

- PD source 1 (F_1): Negative Asymmetry,
- PD source 2 (F_2): Positive Asymmetry,
- PD source 3 (F_3): Symmetry,
- PD source 4 (F_4): Negative Asymmetry with Gap,
- PD source 5 (F_5): Positive Asymmetry with Gap,
- PD source 6 (F_6): Symmetry with Gap,
- PD source 7 (F_7): Gap.

As can be seen, some of these classes are combination of single classes, such as the PD source 5 (i.e., positive asymmetry with gap), which is a combination of PD source 2 and PD source 7. This feature is important to capture because hydrogenerators often have more than one PD active source. Some typical examples of each class have been found in the overall PD database and have been used for training.

TABLE 1. An illustration of the seven PD sources F_i.

PD source 1 (F_1)	PD source 2 (F_2)	PD source 3 (F_3)	
			
PD source 4 (F_4)	**PD source 5 (F_5)**	**PD source 6 (F_6)**	**PD source 7 (F_7)**
			

A. FEATURE VECTOR DEFINITION

The starting point of the above method is to extract features which are efficient in characterizing the 2D-representation of the PD activity.

In this study, we use traditional handcrafted features extraction by adding some of the expert rules and knowledge in the input vector. The feature vector $\mathbb{S}$ is composed by four instances S_i (equ. 6). Each instance S_i is extracted from the expert rules associated to at least one of the previous PD sources. Table 2 highlights the relationship between all features S_i and each PD source F_j.

$$\mathbb{S} = \left(S_1, S_2, S_3, S_4\right) \tag{6}$$

The first set of features S_1 is composed by all the positive and negative discharges as follows:

$$S_1 = (pd_1^x, -pd_1^y \ldots, pd_i^x, -pd_i^y \ldots, pd_{16}^x, -pd_{16}^y) \tag{7}$$

The second feature instance S_2 is given by the equation 8. At each element i, a cumulative difference w_i between the positive and negative discharges is calculated (as shown by the equation 9.)

$$S_2 = \left(w_1, \ldots, w_i \ldots, w_{16}\right) \tag{8}$$

$$w_i = \left((\sum_{j=1}^{i} pd_j^x) - (\sum_{j=1}^{i} pd_j^y)\right) \cdot \kappa_i \tag{9}$$

where κ_i is a parameter that depends on the amplitude channel. The cumulative difference w_i have a significant impact for the identification of the asymmetry or symmetry between the positive and negative discharges.

Finally, to characterize the gap type discharges described on the previous section, two features S_3 and S_4 have been used:

$$S_3 = \left(a_1, a_2\right) \tag{10}$$

$$S_4 = \left(b_1, b_2\right) \tag{11}$$

These two features have been defined by formalizing the expert's knowledge through PD analysis when gap PD occurs. The algorithms 1 and 2 give the computation procedure for each of these two features. The variables a_1, a_2 of the feature S_3, respectively b_1, b_2 of the feature S_4, increase each time the discharge rate (PD/s) increase between two consecutive channels. Note that parameters ϵ, ε and η are defined by the user.

TABLE 2. The relationship between all the used features S_i and each PD source F_i.

PD source	Features			
	S_1	S_2	S_3	S_4
F_1	✓	✓		
F_2	✓	✓		
F_3	✓	✓		
F_4	✓	✓	✓	
F_5	✓	✓	✓	
F_6	✓	✓	✓	
F_7				✓

IV. DEEP NEURAL ARCHITECTURES

A. THE CONVOLUTIONAL VARIATIONAL AUTOENCODER

The CVAE architecture is illustrated in Fig. 6. This architecture includes two parts, an encoder and a decoder, which are two symmetrical and reversed structures. Each one is composed by two convolutional layers and two fully connected layers. For the encoder, we use convolutional layers with 4×1 kernels and the same padding. The stride was 1×1 for the

Algorithm 1 The computation procedure for the feature S_3

Initialization:

$a_1 = 0, a_2 = 0$

$\epsilon > 1, \varepsilon < 1$

for *i=2 to 16* **do**

  if $pd_i^x > pd_{i-1}^x.\epsilon$ **then**

    $a_1 = a_1 + \varepsilon$

  end

  if $pd_i^y > pd_{i-1}^y.\epsilon$ **then**

    $a_2 = a_2 + \varepsilon$

  end

end

Algorithm 2 The computation procedure for the feature S_4

Initialization:

$b_1 = 0, b_2 = 0$

for *i=2 to k (k<16)* **do**

  if $pd_i^x > pd_{i-1}^x$ **then**

    $b_1 = b_1 + \frac{pd_i^x}{pd_{i-1}^x}.\eta$

  end

  if $pd_i^y > pd_{i-1}^y$ **then**

    $b_2 = b_2 + \frac{pd_i^y}{pd_{i-1}^y}.\eta$

  end

end

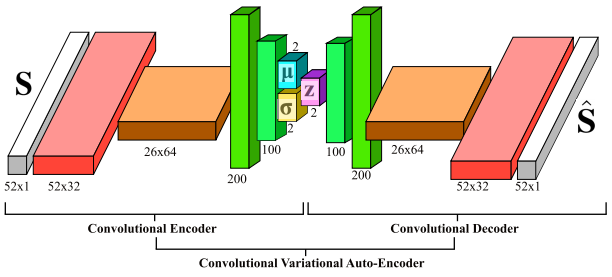


FIGURE 6. The CVAE architecture.

first convolutional layer and 2×1 for the second. The latent two-dimensional space is represented by two layers for the encoder: the mean and the standard deviation layers (i.e., μ and σ), and one sampling layer ($\mathbf{Z}$) for the decoder.

The first step was to train the whole CVAE architecture for the reconstruction of the feature vector ($\hat{\mathbb{S}} = \mathcal{F}(\mathbb{S})$). Training the CVAE does not need the label information of the input data. However, for an efficient data encoding, an indirect labeling is used, since all training samples belong to the PD sources described above. When the training process of the CVAE is successfully done, the encoder part is then used jointly with a neural classifier, as presented in Fig. 7. The mean layer μ of the convolutional encoder is considered as the input 2D-vector of the classifier. The second step is then to train the classifier for PD source recognition. The encoder parameters obtained by the previous step are frozen during

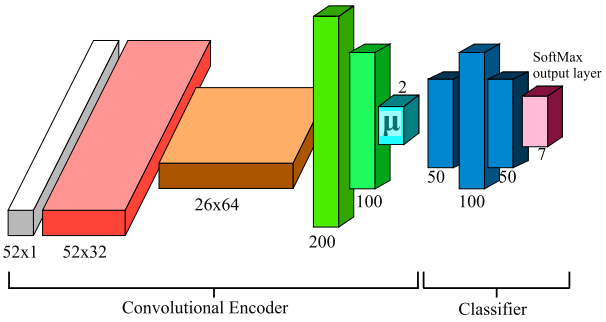


FIGURE 7. The encoder used jointly with a neural classifier. The mean layer of the convolutional encoder is considered as the input 2D-vector of the classifier.

the classifier training step. All the details of the CVAE and the classifier, are presented in Table 3.

TABLE 3. Proposed DNN architectures.

Layer	Type	Neurons	Kernels
Encoder			
0	Input feature vector	52×1	-
1	Convolution	52×32	4×1
2	Convolution	26×64	4×1
3	Fully connected	200	-
4	Fully connected	100	-
5	Mean layer	2	-
5	Standard deviation layer	2	-
Decoder			
0	Sampling layer	2	-
1	Fully connected	100	-
2	Fully connected	200	-
3	Deconvolution	26×64	4×1
4	Deconvolution	52×32	4×1
5	Output reconstructed vector	52×1	-
Classifier			
0	Input (mean layer)	2	-
1	Fully connected	50	-
2	Fully connected	100	-
3	Fully connected	50	-
4	SoftMax output layer	7	-

B. VISUALIZATION OF LATENT VECTORS AND PERFORMANCE COMPARISON

Considering that the latent vectors are the encoding representation of the input feature vectors, it is interesting to visualize the 2D-representation of the original vectors and to evaluate the similarities between each PD source F_i according to each feature set S_j. In the first test (Fig. 8), we have compared the visualization performance of the CVAE with two state-of-the-art dimension reduction methods including principal components analysis (PCA) [45] and t-SNE [20]. All the

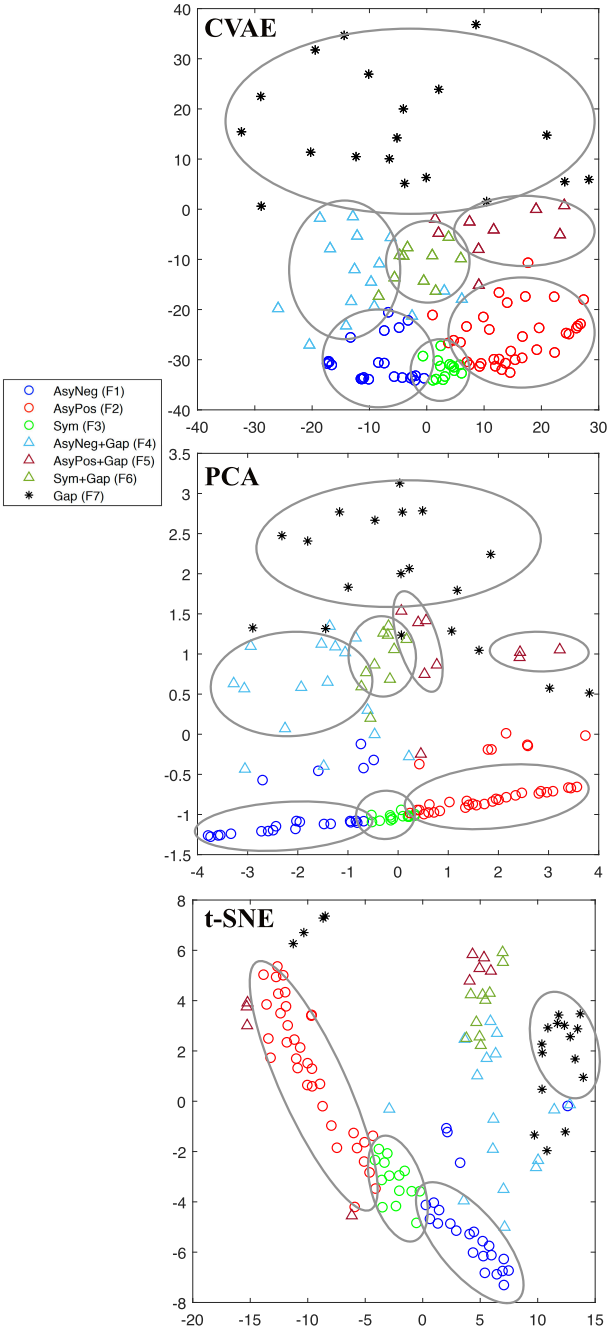


FIGURE 8. The 2D-visualization using three dimension reduction methods: the CVAE, the principal components analysis (PCA) and t-SNE.

training samples belong to one of the PD source F_i. The color of each sample corresponds to the labelling knowledge provided by experts. To quantitatively assessing the performance of these methods in dimension reduction and visualization, we compared the PD clusters obtained on the original 52-dimensionnal space (OS) with the 2D reduced subspaces, obtained by the three reduction methods (i.e. CVAE, PCA and t-SNE). We have first calculated the mean vector of each PD source clusters in each dimension space (i.e. the original space and in the three reduced spaces). Then we used the k-nearest neighbors algorithm (*k-NN*) [46] to classify each vector. By comparing the true label and the label obtained by the *k-NN*, we have calculated the accuracy given by the following expression:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where TP, TN, FP, FN are respectively the true positive, true negative, false positive and false negative. Table 4 gives the summary of the obtained accuracy. For each PD source F_i, we highlight the highest and lowest value with the given color. The mean accuracy is calculated for the original space and for each of the reduction method.

TABLE 4. The accuracy obtained by the 52-dimensionnal original space (OS) and the three reduction methods with the k-nearest neighbors algorithm (*k-NN*).

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	Av.
CVAE	0.94	0.93	0.91	0.93	0.94	0.94	0.96	0.94
PCA	0.92	0.91	0.86	0.91	0.94	0.96	0.89	0.91
tSNE	0.94	0.88	0.92	0.91	0.91	0.93	0.94	0.92
OS	0.92	0.91	0.84	0.93	0.98	0.97	0.97	0.93
	High						Low	

By these comparative tests, the CVAE showed better separation of the PD sources than the two other dimension reduction methods. Indeed, the best average accuracy has been obtained by the CVAE (0.94%), as shown by the table 4. This result is visually perceived by analyzing the clusters obtained by each reduction method of the figure 8.

In the second test (Fig. 9), we have visually evaluated the performance of the CVAE according to different combinations of features S_1, S_2, S_3, S_4 described before. We can see on the figure 9 that the PD source separation showed poor performances when we consider only the feature S_1. Clustering of the different PD source is less efficient. All PD sources are mixed on one dense and localized cluster (colors overlap). It is obvious that, in this 2D-representation, the performance of the PD source classification will be very poor. On the contrary, data files from PD sources showing similar features of asymmetry (positive or negative) or of symmetric PD tend to cluster together when the feature S_2 is used jointly with S_1. Thus, three clusters are formed and spread over a larger area of the latent representation: cluster 1 is formed by the latent points of the PD source F_1 and F_4 while cluster 2 is formed by F_2 and F_5. Finally cluster 3 groups points of source F_3 and F_6. Moreover, the dissociation of the Gap is possible only when the feature S_3 is used. The PD sources with gap (F_4, F_5, F_6) are then separated from the other sources. Finally, due to the feature S_4, the PD source F_7 is separated from the other PD sources and can form an isolated cluster. This sensitivity study confirms that the latent space simplifies data visualization. By improving the input vector, isolated clusters of all PD sources are more easily defined.

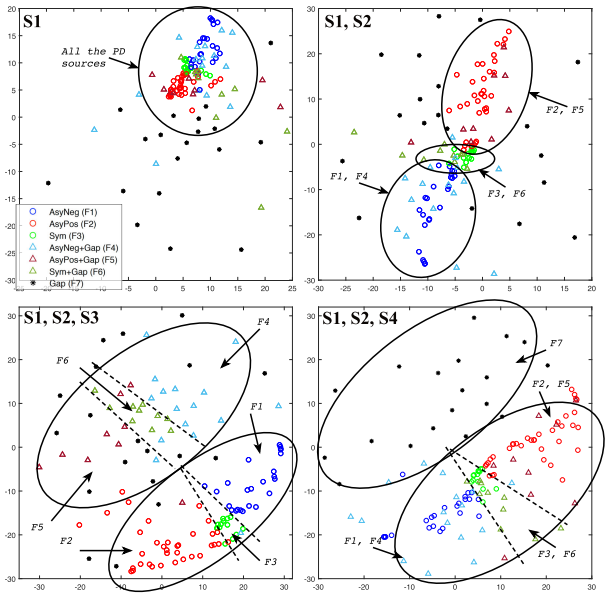


FIGURE 9. 2D-latent representation obtained by the convolutional encoder with different combinations of features S_1, S_2, S_3, S_4.

C. IDENTIFICATION AND REDUCTION OF THE CONFLICT ZONE

As described in section I (see Fig. 1), the conflict zone is the area between clusters where most of the false predictions occur. This area is due to the poor distribution of data used in the training process. Because the human brain can perceive only two or three-dimensional space, it is impossible for humans to process the 52-features dimensional space used to define the input vector of a PD data file. It is even worse when trying to compare files between them. Therefore, projecting data in 2D-latent space helps experts to easily identify clusters of similar PD sources and locate areas or conflict zones where it is necessary to add new data to label. When the conflict zone is too large, the boundary between classes, fixed by the classifier, leaves a large number of files with uncertain classification. This means that two different classifiers C_i and C_j with $i \neq j$ will definitely have two opposite responses for the same input data k:

$$\Psi_i(k) \neq \Psi_j(k) \tag{13}$$

where $\Psi_i(k)$ and $\Psi_j(k)$ are respectively the output class obtained by the classifiers C_i and C_j for the input sample k. To identify these conflict zones, several classifiers C_i were trained with the same training data set. Subsequently, all the trained classifiers have been tested over the entire MIDA database. For each input sample k of the MIDA database, if two classifiers have two opposite responses, the input sample k is then considered as a conflict sample. To reduce the proportion of false predictions, the size of the conflict zone must be reduced. For this, experts will choose new learning data files in these conflict areas in order to adjust the learning of the classifier (Fig. 2). These new samples are then labeled by experts and added to the previous training data set even if they do not strictly respect the initial expert rules. Algorithm 3 shows the steps of the entire model training process and new data selection for reduction of the conflict zone. The entire procedure is then repeated until the conflict zone is considered as acceptable by experts. This is done without having to consider all data files from the database, but just a few additional files located in the conflict zone.

Algorithm 3 The training algorithm and data selection

Data

Ω_{train} : is the training labeled dataset
Ω_{mida} : is the whole MIDA unlabeled dataset
Ω^{2D}_{train}: is the 2D-representation of Ω_{train}
Ω^{2D}_{mida} : is the 2D-representation of Ω_{mida}
$Round = 1$: is incrementing each time the steps 2,3,4 are computed
*

Step1: Train the CVAE

→ Train the CVAE on Ω_{train}
→ Use the encoder to convert the data from the feature space to the 2D-latent space:
$\Omega_{train} \overset{\text{Encoder}}{\Longrightarrow} \Omega^{2D}_{train}$
$\Omega_{mida} \overset{\text{Encoder}}{\Longrightarrow} \Omega^{2D}_{mida}$
*

Step2: Train the classifier

→ Train 10 classifiers C_i on Ω^{2D}_{train} as follow :
for *i=1 to 10* **do**
  Train the classifier C_i on Ω^{2D}_{train}
  Save the parameters of C_i
end
*

Step3: Identification of the conflict zone

→ Test the classifier C_i on Ω^{2D}_{mida}
→ $\Psi_i(k)$: is the output class obtained by the classifier C_i for the input sample k
for *each element k of the whole dataset* Ω^{2D}_{mida} **do**
  for *each two classifiers* C_i *and* C_j *with* $i \neq j$ **do**
    if $\Psi_i(k) \neq \Psi_j(k)$ **then**
      The element k is considered to be part of the conflict zone
    end
  end
end
*

Step4: Evaluation of the diagnosis model

if *Performance criterion is obtained* **then**
  End of the training process
else
  → Choose new unlabeled samples from the whole dataset Ω^{2D}_{mida} in order to reduce the conflict area
  → Label these new samples
  → Add these new samples to the training dataset Ω^{2D}_{train}
  → Repeat the steps 2, 3 and 4
  → $Round = Round + 1$
end

V. RESULTS

This section presents the experimental results and is split into four parts. The first part details the way the training data set has been built. By applying the procedure of the algorithm 3, three training rounds were sufficient to adjust the conflict zone between classes. Then, in the second part, the effect of the selected features on the classification performances was visually evaluated. Several combinations of features S_i have been tested and applied on the overall MIDA database. In the third part, the effect of the dimension reduction obtained by the convolutional encoder was further studied and evaluated. For this, we have compared the model used in this work, called "convolutional encoder classifier" against a classical classifier, called "traditional classifier" (i.e., a classifier without dimension reduction). Finally, in the last part, the validation results are presented. These results were obtained from a reference data set that have been labeled by experts of Hydro-Québec and not used during the learning process.

A. BUILDING OF THE TRAINING DATA SET AND REDUCTION OF THE CONFLICT ZONE

In this section the way that the training data set have been built with the support of the 2D-representation is presented. Fig. 10 shows the evolution of the conflict zone obtained by the learning process described by the algorithm 3. This conflict area is represented by the magenta points. Starting with an initial training data set (Fig. 10.A), three rounds of the algorithm 3 were sufficient to obtain an acceptable conflict zone. At each round R_i, new data files were chosen from the 2D-latent space, labeled by experts and then added to the training data set. These new samples are pointed by an arrow in the Fig. 10 B and C for rounds R_2 and R_3 respectively. Table 5 gives the size of the training data set used for each PD source F_i and for each of the three rounds R_1, R_2, R_3. The reduction of the conflict zone when new data files were added to the training set is shown for each round in Fig. 10 (D, E and F). It is important to note that the conflict zone is the boundary between classes, obtained by the classifier during the training process. This boundary is less confrontational when new learning data files are wisely chosen. A total of 127 data files, which represent 0.38% of the whole MIDA database, were used for the training process during the third round R_3.

B. EFFECT OF THE SELECTED FEATURES

Fig. 11 shows boundaries (i.e., the conflict zones) obtained on the overall MIDA database by the convolutional encoder classifier according to the feature combinations presented earlier in Fig. 9. The smallest boundaries and the best defined clusters for each PD source were obtained when all features S_1, S_2, S_3, S_4 were used. All other combinations are worse and as it can be seen in Fig. 11.B, when only the feature S_1 is considered, the largest boundary size with the least defined clusters is obtained. In fact, the size of the boundary zone can be used as an indicator of the quality of classification.

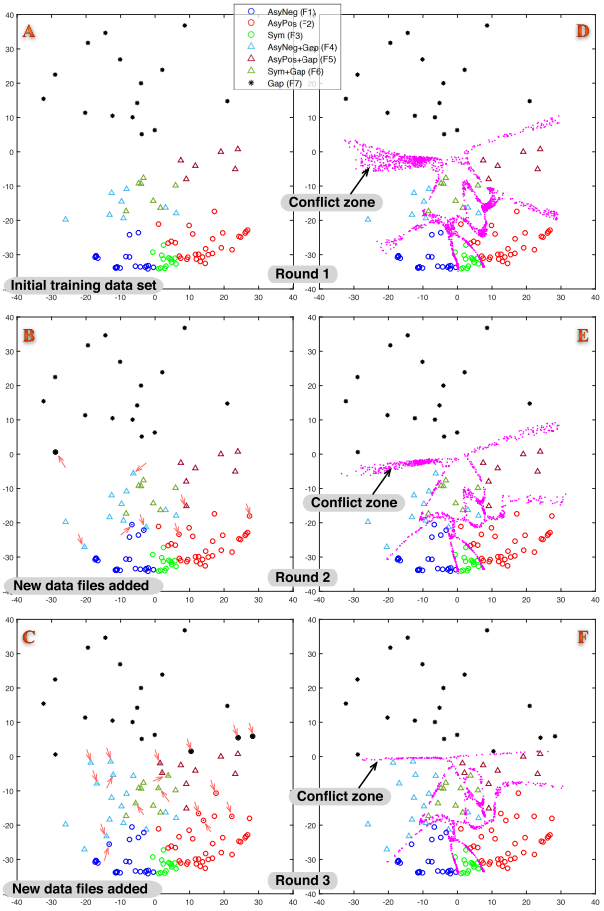


FIGURE 10. The training data set and the conflict area obtained at each round of the algorithm 3: (A,D) for the 1st round, (B,E) for the 2nd round and (C,F) for the 3rd round.

For instance, during our three rounds of training process the boundary zone decreased from 6 % to 3 % of the latent space.

C. EFFECT OF THE DIMENSION REDUCTION

It can be argued that the use of an additional processing function, such as the encoder function of the CVAE, cause loss of information and deterioration of the classification performances. To assess the effect of the additional use of this encoding function, the output classification results obtained by the architecture used (i.e., the encoder part of the CVAE used jointly with a classifier, which is shown in Fig. 7) was confronted with the output classification results obtained without the use of the convolutional encoder.

Fig. 12 shows the confusion matrices for these confrontation tests obtained over the entire MIDA database. Several combinations of features S_1, S_2, S_3, S_4 have been tested. At each test, the classifier used without VAE was trained on the same data set obtained after round R_3 of the learning process (see table 5). The output classes #1 to #7 represent the PD sources F_1 to F_7. The class #8 is the ambiguity class, it means that none of the output classes #1 to #7 is higher than the decision threshold (for these tests, a decision

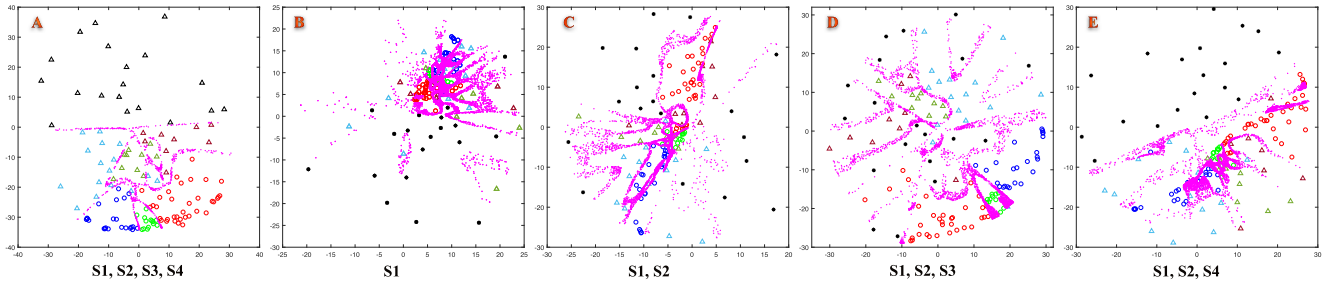


FIGURE 11. The conflict zones obtained by the convolutional encoder classifier according to different combinations of the features S_1, S_2, S_3, S_4. The more confused result is shown by the Fig. B, when only the feature S_1 is considered, and the best result is obtained by the Fig. A.

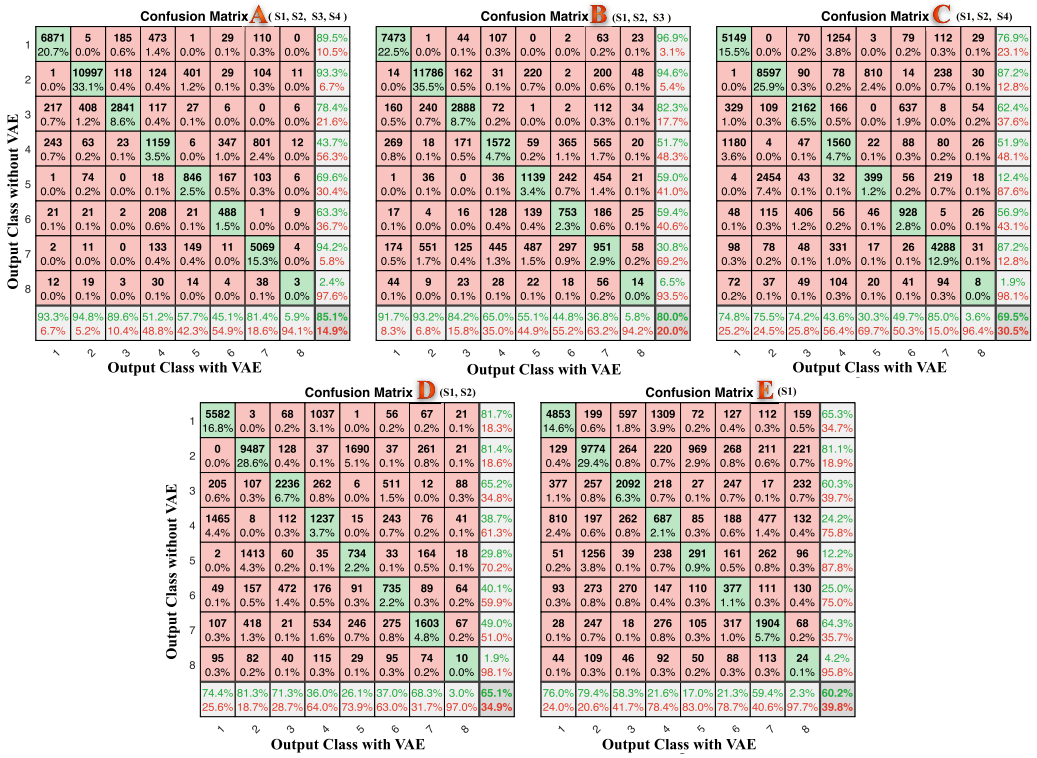


FIGURE 12. The confusion matrices for the confrontation tests between the convolutional encoder classifier and a traditional classifier. Several feature combinations were used for these confrontation tests.

TABLE 5. The evolution of the training data set at each round R_i of the algorithm 3. The bold values are the percentage relative to the whole MIDA database (33 223 measurement files). The Total column gives the number of training data files used at each round of the algorithm.

Round		PD sources							Total
		F_1	F_2	F_3	F_4	F_5	F_6	F_7	
R_1		21	29	14	8	6	7	15	100
	%	**0.063**	**0.087**	**0.042**	**0.024**	**0.018**	**0.021**	**0.045**	**0.3**
R_2		+2=23	+2=31	+0=14	+3=11	+1=7	+0=7	+1=16	+9=109
	%	**0.069**	**0.093**	**0.042**	**0.033**	**0.021**	**0.021**	**0.048**	**0.33**
R_3		+1=24	+4=35	+0=14	+5=16	+2=9	+3=10	+3=19	+18=127
	%	**0.072**	**0.1**	**0.042**	**0.048**	**0.027**	**0.03**	**0.057**	**0.38**

threshold of 0.5 was used.) It can be seen that the best result of the concordance rate (85.1%) is obtained when all features were used jointly (Fig. 12, matrix A). The worst score of 60.2% is obtained when only the feature S_1 is used (Fig. 12, matrix E). It is interesting to see how features S_i influence the concordance rates between the two models. For example,

TABLE 6. The distribution of the validation data set according to each PD source F_i.

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	**Total**
#(%)	7(8.4%)	16(19.3%)	12(14.5%)	17(20.5%)	8(9.6%)	11(13.3%)	12(14.5%)	83(100%)
CVAE + Classifier (Classifier with dimension reduction)								
TP	6	15	4	8	6	6	9	54(65.06%)
FP	5	7	3	4	2	4	4	29(34.94%)
PPV	54.55%	68.18%	57.14%	66.67%	75.00%	60.00%	69.23%	65.06%
Classifier without CVAE (Without dimension reduction)								
TP	6	14	5	10	4	4	6	49(59.04%)
FP	2	8	6	6	4	3	5	34(40.96%)
PPV	75.00%	63.64%	45.45%	62.50%	50.00%	57.14%	54.54%	59.04%
Expert Rules								
TP	1	7	8	5	1	5	3	30(36.14%)
FP	1	3	3	5	10	6	8	36(43.37%)
PPV	50%	70%	72.73%	50%	09.09%	45.45%	27.27%	45.45%

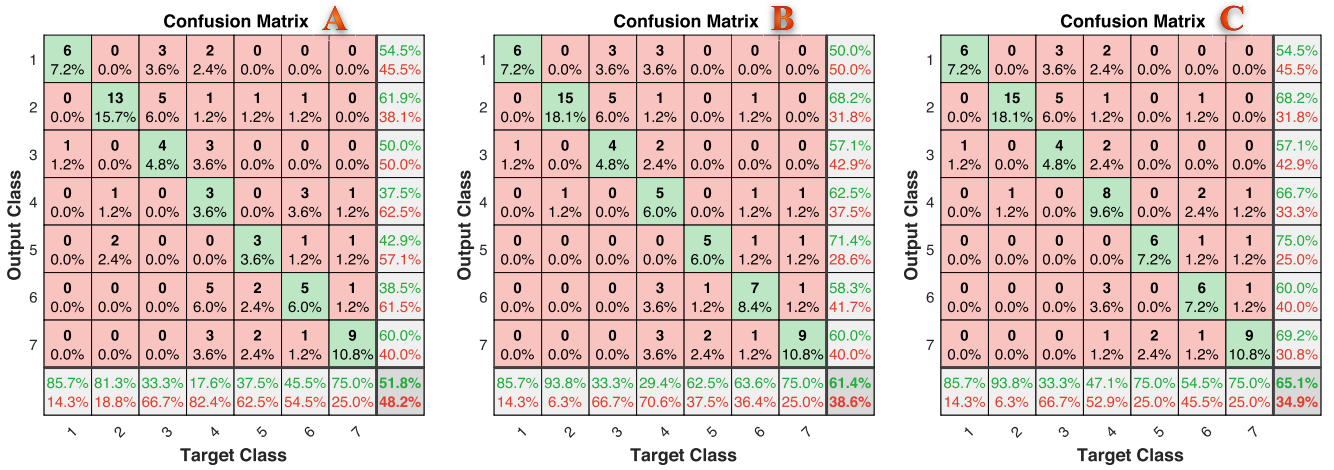


FIGURE 13. The confusion matrices of the convolutional encoder classifier for each learning round of the algorithm 3 (matrices A, B and C for respectively the round 1,2 and 3).

the concordance rate of the PD source F_7 is higher when the feature S_4 is used (Fig. 12, matrix C).

D. TEST ON A VALIDATION DATASET

This section presents results obtained from the validation data set on which the performances of the convolutional encoder classifier have been evaluated. Table 6 provides the distribution of this validation data set according to each PD source F_i. A total of 83 PD data files have been labelled by the experts of Hydro-Québec. As shown by the table 6, the proposed PD recognition method (i.e. classifier with dimension reduction) has been confronted to the PD recognition obtained by the classical classifier on the one hand, and by the rules used by the experts of Hydro-Québec on the other hand. For this comparison test, we provide the true positive (TP), the false positive (FP) and the Positive Predictive Value ($PPV = TP/(TP + FP)$) obtained for each PD source. These results show that the convolutional encoder classifier outperforms the two other methods. For example, the total PPV of 65.06% was obtained by the proposed method while 59.04% was obtained by the classical classifier and 45.45% by the expert rules. It should be noted that 17 PD files were not classified by the expert rules.

Fig. 13 shows the confusion matrices of the convolutional encoder classifier for each learning round of the algorithm 3. The total classification accuracy is improved at each round of the learning algorithm (51.8% at round R_1, 61.4% at round R_2 and 65.1% at round R_3). The best improvements were obtained for both classes F_4 and F_5 with 17.6% to 47.1% for F_4 and 37.5% to 75% for F_5. As seen in the Fig. 13.C, the accuracy obtained for each class F_i is extremely disparate (e.g., 93.8% for F_2, and 33.3% for F_3).

It is interesting to analyze these results visually on the 2D-latent space representation of the classified data files (Fig. 14). The color of the square represents the label given by expert while the color of the point in the center of each square is the output class obtained by the convolutional encoder

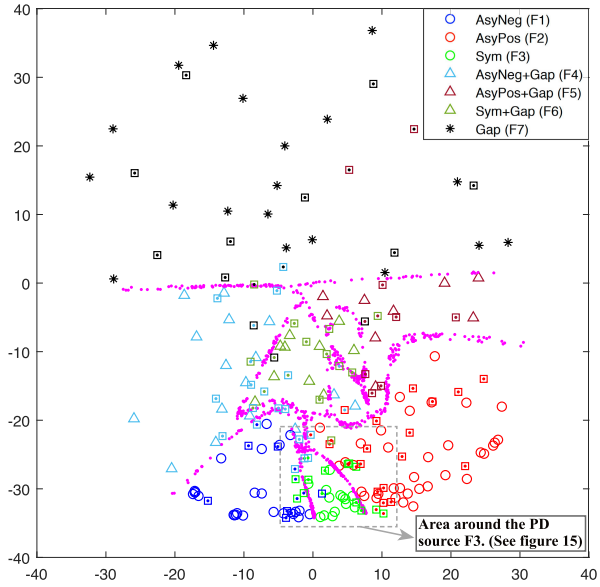


FIGURE 14. The 2D-latent space representation of the validation test shown jointly with the training dataset. The validation data files are presented by square. The color of the square represents the label given by experts while the color of the point in the center of each square is the output class obtained by the convolutional encoder classifier.

classifier. An enlargement of a zone of interest is presented in Fig. 15.

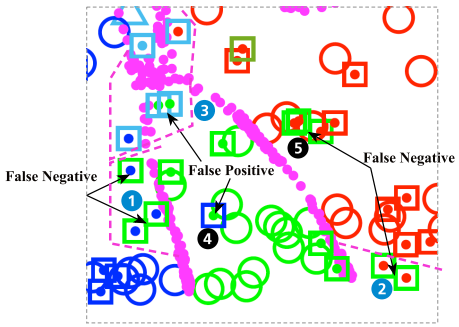


FIGURE 15. A better view of the area around the PD source F_3 obtained from the Fig. 14.

For the worst result obtained in the case of PD source F_3 (considered here as the positive class), which is composed of 12 data files as presented in table 6, the confusion matrix C in Fig. 13 leads to the following observations:

- 4 data files of the PD source F_3 have been correctly classified as F_3 by the convolutional encoder classifier (True Positive = 4).
- 8 data files of the PD source F_3 have been incorrectly classified (False Negative = 8) as follows: 3 data files incorrectly classified as F_1 and 5 data files incorrectly classified as F_2.
- 3 data files have been incorrectly classified as F_3 (False Positive = 3) as follows: 1 data file from F_1 and 2 data files from F_4.

All these results can be visualized in Fig. 15. This 2D-latent space representation shows that most of the false predictions are located near the conflict area. For example, the two false positives on the top, are identified by the blue circle number 3. The three false negatives of the left side are identified by the blue circle 1, and the two false negatives on the right side are identified by circle 2. They all are located near the boundaries. These false predictions are due to a lack of knowledge on this part of the 2D-latent space representation during the learning process. It can be supposed that conflict areas between classes obtained during the learning process are not yet optimized and could be improved if additional data files close to the conflict area were added to the training data set. Thus, the new conflict area between classes could be redefined as shown by the pink dashed lines in Fig. 15. The data files here are part of the validation data set. Thus, other files close by should be identified, labeled and injected in the training data set in additional training rounds.

A less coherent result concerns the isolated false positive, identified by the black circle 4 and the three false negatives, identified by the black circle 5 in Fig. 15. These data files are very close to other data files used in the learning process and yet belong to opposite classes as identified by experts. These false predictions cannot be corrected by boundary refinement. These data files represent some of the particular cases that should be thoroughly analyzed by experts to capture new knowledge currently not included in the input vector. One of the main reasons of this misclassification could be due to the expert's parameters rules (κ_i, ϵ, ε and η) used to build the input feature vector. There could be an additional parameter to introduce in the input vector. Indeed, it is essential to find the best setting of these feature parameters to improve the classification performances. The adjustment of these parameters is part of future developments.

Finally, the last criterion evaluated in this section is the Area Under the Curve (AUC) which is a performance measurement for classification problem at various decision thresholds settings. The closer the AUC value is to 1, the better will be the classifier. The tables 7 and 8 detail the AUC for each PD source class F_i. The bold values represent the maximum result for each column. The results state that, except for the PD source F_4, in most of the cases, the convolution encoder classifier outperforms the traditional classifier.

TABLE 7. The AUC values obtained at each round and for each PD source F_i.

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
R_1	0.97	0.92	0.86	0.77	0.87	0.80	0.91
R_2	0.96	0.95	0.88	0.82	0.83	0.87	**0.92**
R_3	**0.98**	**0.97**	**0.90**	**0.83**	**0.94**	**0.90**	**0.92**

VI. DISCUSSION

In this paper, we have described an innovative method based on a CVAE to build a PD source classifier. Two major observations follow:

- The first is the use of the expert rules to build the input feature vector.

TABLE 8. The AUC values obtained by the convolution encoder classifier and the traditional classifier according to several combinations of the features F_i.

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
CVAE + Classifier							
S_1,S_2,S_3,S_4	**0.98**	**0.97**	**0.90**	**0.83**	**0.94**	**0.90**	**0.92**
S_1	0.86	0.82	0.83	0.77	0.71	0.68	0.79
S_1,S_2	0.84	0.77	0.72	0.72	0.84	0.87	0.84
S_1,S_2,S_3	0.97	0.94	0.80	0.78	0.85	0.84	0.61
S_1,S_2,S_4	0.88	0.84	0.70	0.61	0.61	0.78	0.66
Classifier							
S_1,S_2,S_3,S_4	0.93	0.93	**0.83**	0.84	**0.94**	0.83	**0.90**
S_1	**0.94**	**0.95**	0.78	0.87	**0.94**	0.82	0.89
S_1,S_2	0.91	0.92	0.81	0.84	0.92	**0.85**	0.86
S_1,S_2,S_3	**0.94**	0.92	**0.83**	**0.92**	0.92	**0.85**	0.89
S_1,S_2,S_4	0.92	0.93	0.81	0.81	0.89	0.81	0.85

- The second is the use of a CVAE to display a 2D-representation of the feature space, used for visual analysis by experts.

A. EXPERT KNOWLEDGE FOR TRANSFER LEARNING

The PD sources are often difficult to characterize because of the ambiguity caused by the overlapping sources (i.e., different classes). The class boundary is therefore not easy to define and often, for complicated cases, disagreement as well as conflicting diagnoses between experts are not excluded. Nevertheless, this only represent a small fraction of all files. The diversity of PD sources to be treated makes the diagnosis even more difficult when dealing with real online measurement. As reported in [43], [47], [48], most of PD source recognition techniques developed in the literature were mainly trained on artificial defect models in laboratory, which may not function well when put into practice in PD online measurement scenarios. In our case, we tested our method on more than 33 000 unlabeled PD measurement files collected by Hydro-Québec during the last 30 years.

Another difficulty for characterizing PD sources is that one measurement can contain different sources in a single file. In the case of this study, we restricted to a maximum of two sources present at the same time, but it is not excluded that one measurement can contain three sources.

The expert in his analysis, takes into account analytical data based on certain rules defined over time, such as the ratio between the positive and negative discharges but also less formal information such as:

- the specific form of the histogram that characterizes each PD source (as shown in Table 1) to see for example if the gap PD activity is present (which can lead to special characteristic of the PD distribution);
- the date of the measurement, to see if it corresponds to a dry period as during winter because it impacts on a number of PD sources;
- results from other diagnostic tools and visual inspections.

All the challenge is then to extract the expert knowledge and to find the right formalization of the expert rules to build the input characteristic vector processed by the NN classifier. Many studies in the literature claim in recent years that DNNs, particularly the CNNs, are able to extract characteristic features themselves from raw data provided from a big dataset. However, the results presented in this article show quite the opposite. Indeed, we have seen that the more we put knowledge in the input vector, the more the data space becomes discriminant, thus making the classification easier. We have seen that the separation between classes is transformed from a non-linear separability with high overlapping clusters in the space feature S_1 (Fig. 11.B), into a pseudo-linear one in the space feature S_1, S_2, S_3, S_4 (Fig. 11.A) with less overlapping clusters. This reduces the ambiguities and conflict areas between classes. The training process of the NN, which is based on the rules extracted from the expert knowledge, would be even more efficient than the rules defined as: *If* $x \geq threshold$ *Then* $a = y$. The main difficulty with these kind of rules is to treat the exceptions cases. Changing a rule to fit an exception may disrupt other properly processed cases. Adding new rules can become very complicated. Often for these exceptions, the expert transgresses the rules, which makes these methods not very suitable for online diagnosis. The advantage of the NNs is their ability to fit the boundaries between classes in a non-linear way, supervised by experts. For the exception's rules, the NN redefined the boundary around this exception in a non-linear way, without modifying the knowledge previously acquired.

It is therefore essential to inject expert knowledge into the characteristic vector to improve the classification. Results demonstrate that the use of the expert knowledge, as hand-crafted input features, is essential for an appropriate clustering of data belonging to the same PD source. The paradigm that the NN performs better when trained on big-data is not always true, the quality of the data being another parameter to be considered for the training. Therefore, *right-data* is as relevant or even more so as *big-data*.

B. VISUAL DATA ANALYSIS FOR MODEL CONFIDENCE

The different steps to follow when using ANNs (shallow or deep) are: (1) model choosing; (2) model building; (3) model learning; and (4) model checking. In the first step, one neural architecture must be chosen. In the second step, the size of the NN must be defined: how many layers? how many units per layer? how many convolution filters and what is their size? In the third step, the NN will be trained by supervised techniques. In the last step, the confidence of the NN must be validated.

It is usually very hard to build the architecture of a NN. For example, when using a CNN, it is not obvious for a non-expert to define the best size of the convolution filters and to choose the best number of convolution, pooling or fully-connected layers. A good alternative to this drawback is to consider the neural architecture as a hyper-parameter evolving during the learning process. The NN is built step-by-step during the

learning process, until a convergence criterion is reached. Recently, several promising studies about the constructive and pruning algorithms were published (see [49], [50] for a complete survey).

A more fundamental problem is the model confidence evaluation and the interpretability of the obtained results. ANNs learn to associate an output according to a given input, but they do not learn to give any reason or interpretation associated to this response. However, the outstanding performance of the classification models does not lead to easy model interpretations or model confidence, although DNNs reveal superior performance and have been extensively used in multiple disciplines. Moreover, in most cases, they are considered as black-boxes and the interpretation of their internal working mechanism is usually challenging [14].

In the proposed method, the CVAE is used as a support for data visualization and PD source classification. However, after training of the CVAE, the encoder part is used as a data projection from input features mapping to a latent 2D-representation. The CVAE is trained without using class labels to learn properties in the data. The power of CVAE resides in capturing the complicated data features from the multi-dimensional input space and compressing them into a smaller 2D-latent space, making it easier for visualization by a human expert. The CVAE reveals a promising tool to produce a 2-dimensional embedding of high dimensional data with the goal of simplifying the identification of clusters when used jointly with a classifier.

As articulated in the results section, it is quite easy to understand the diagnosis given by the neural classifier by visually analyzing the spatial distribution of the classified samples. The knowledge area of the NN and the boundaries between the classes are easily perceived by experts. The conflict regions which are poorly covered by the training data file are then easily identified. The NN is then less perceived as a ''black-box'' in the sense that its knowledge area is visible, the interpretation of the false predictions and their understanding become achievable.

C. METRICS DEFINITION FOR MODEL PERFORMANCES

The last condition of the proposed algorithm 3 (i.e., step 4) is the model evaluation through the testing of a performance criterion. As stated before, for the tests presented in this paper, we do not address this condition of the algorithm. The only criterion used is the reduction of the conflict area, which is visually perceived by the expert, at each round of the algorithm. The performances obtained for the PD source F_3, discussed in the previous section, show that this condition is not sufficient (Fig. 15). Indeed, the classification accuracy of F_3 is the worst in spite of the reduced conflict zone (i.e., a thin boundary delimiting this class). More training rounds could be tried. However, it is clear from our study that the expert assisted CNN only has to process a fraction of the available data in order to decrease the conflict area. In this example, only 0.38 % of all available data was used for training; so even with more rounds, it should stay below 1 %.

To improve the performances of the proposed method, other metrics than the reduction of the conflict area must be defined and used to evaluate the performance of such model during the building process. These metrics should give some answers to these non-exhaustive questions:

- How to evaluate the quality of the spatial distribution of an input training dataset?
- How to evaluate the features (expert knowledge) which are closely related to the quality of the input vector? We have observed that overlapping of opposite classes decreases when more expert rules are injected to the input vector.
- How to evaluate the performances of the CVAE, especially the encoder function?
- A more fundamental question concerns the minimum size of the conflict area tolerated by the expert. This is a real dilemma for safety-critical applications. For some borderline cases, an ambiguous response from the NN can be more acceptable than a false prediction. It is then preferable that the NN, for these critical cases, lets the expert decide rather than propose a catastrophic false prediction.

VII. CONCLUSION AND FURTHER WORK

In this paper, the use of a CVAE as a visual support for the interpretation of PD sources has been investigated. The starting point of the proposed method was to extract features which are efficient in characterizing the PD distribution profile. Finding the best features that identify PD sources in order to pinpoint which failure mechanisms are possibly active is fundamental for an accurate diagnosis.

Despite their many advantages, results reveal that the DNNs are not always able to extract the best features. In the case of the PD sources classification, the most suitable way is to integrate the expert knowledges into the input vector based on handcrafted feature extraction. The boundaries between the different classes become obvious when the expert knowledge is used as a building block of the input vector. Results demonstrate that the use of the encoder function of the CVAE does not deteriorate the classification performances. The method was tested on more than 33 000 unlabeled PD measurement files collected by Hydro-Québec.

Although promising results were obtained by the proposed method, the accuracy on the reference data set still should be improved. A global classification accuracy up to 65% has been achieved but, for certain PD sources, this accuracy does not exceed 35%. Future developments will be focused on the improvement of the extraction of the expert's rules to define the best discriminant feature vector characterizing all the PD with one or more sources. In addition, optimization of the conflict zones could be improved.

REFERENCES

[1] T. Han, C. Liu, W. Yang, and D. Jiang, ''Deep transfer network with joint distribution adaptation: A new intelligent fault diagnosis framework for industry application,'' *ISA Trans.*, Aug. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0019057819303489

[2] C. Hudon, M. Belec, and D. N. Nguyen, "Innovative Web system for condition-based maintenance of generators," in *Proc. IEEE Electr. Insul. Conf.*, May 2009, pp. 234–245.
[3] M. Lévesque, N. Amyot, C. Hudon, M. Bélec, and O. Blancke, "Improvement of a hydrogenerator prognostic model by using partial discharge measurement analysis," in *Proc. Annu. Conf. Prognostics Health Manage. Soc.*, vol. 8, 2017, p. 7.
[4] O. Blancke, A. Tahan, D. Komljenovic, N. Amyot, M. Lévesque, and C. Hudon, "A holistic multi-failure mode prognosis approach for complex equipment," *Rel. Eng. Syst. Saf.*, vol. 180, pp. 136–151, Dec. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0951832017312632
[5] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, CS Tech. Rep., 2009.
[6] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation and active learning," in *Proc. 7th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Cambridge, MA, USA: MIT Press, 1994, pp. 231–238. [Online]. Available: http://dl.acm.org/citation.cfm?id=2998687.2998716
[7] H. Yu, X. Yang, S. Zheng, and C. Sun, "Active learning from imbalanced data: A solution of online weighted extreme learning machine," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1088–1103, Apr. 2019.
[8] W. Mao, Y. Liu, L. Ding, and Y. Li, "Imbalanced fault diagnosis of rolling bearing based on generative adversarial network: A comparative study," *IEEE Access*, vol. 7, pp. 9515–9530, 2019.
[9] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.
[10] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[11] Fanny and T. W. Cenggoro, "Deep learning for imbalance data classification using class expert generative adversarial network," *Procedia Comput. Sci.*, vol. 135, pp. 60–67, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918314364
[12] J. Błaszczyński and J. Stefanowski, "Improving bagging ensembles for class imbalanced data by active learning," in *Advances in Feature Selection for Data and Pattern Recognition* (Intelligent Systems Reference Library), vol. 138, U. Stańczyk, B. Zielosko, and L. Jain, Eds. Cham, Switzerland: Springer, 2018.
[13] W. S. Gill, I. T. Nabney, and D. Wells, "Helicopter vibration sensor selection using data visualisation," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2012. pp. 1–6.
[14] J. Wang, L. Gou, W. Zhang, H. Yang, and H.-W. Shen, "DeepVID: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation," *IEEE Trans. Visual. Comput. Graph.*, vol. 25, no. 6, pp. 2168–2180, Jun. 2019.
[15] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841517301135
[16] R. Zemouri, N. Zerhouni, and D. Racoceanu, "Deep learning in the biomedical applications: Recent and future status," *Appl. Sci.*, vol. 9, no. 8, p. 1526, Apr. 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/8/1526
[17] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2016, *arXiv:1610.02391*. [Online]. Available: https://arxiv.org/abs/1610.02391
[18] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 818–833.
[19] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316303582
[20] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.
[21] M. Usama, J. Qadir, A. Raza, H. Arif, K.-L.-A. Yau, Y. Elkhatib, A. Hussain, and A. Al-Fuqaha, "Unsupervised machine learning for networking: Techniques, applications and research challenges," *IEEE Access*, vol. 7, pp. 65579–65615, 2019.
[22] G. C. Linderman and S. Steinerberger, "Clustering with t-SNE, provably," *SIAM J. Math. Data Sci.*, vol. 1, no. 2, pp. 313–332, Jan. 2019, doi: 10.1137/18M1216134.
[23] S. Yu and J. C. Príncipe, "Understanding autoencoders with information theoretic concepts," *Neural Netw.*, vol. 117, pp. 104–123, Sep. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608019301352
[24] X. Wang, D. Peng, P. Hu, and Y. Sang, "Adversarial correlated autoencoder for unsupervised multi-view representation learning," *Knowl.-Based Syst.*, vol. 168, pp. 109–120, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705119300176
[25] A. Mishra, M. S. K. Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," Sep. 2017, *arXiv:1709.00663*. [Online]. Available: https://arxiv.org/abs/1709.00663
[26] Y. Bengio, "How auto-encoders could provide credit assignment in deep networks via target propagation," 2014, *arXiv:1407.7906*. [Online]. Available: http://arxiv.org/abs/1407.7906
[27] D. J. Im, S. Ahn, R. Memisevic, and Y. Bengio, "Denoising criterion for variational auto-encoding framework," 2015, *arXiv:1511.06406*. [Online]. Available: http://arxiv.org/abs/1511.06406
[28] Y. J. Fan, "Autoencoder node saliency: Selecting relevant latent representations," *Pattern Recognit.*, vol. 88, pp. 643–653, Apr. 2019.
[29] G. S. Martin, E. L. Droguett, V. Meruane, and M. Das Chagas Moura, "Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis," *Struct. Health Monit.*, vol. 18, no. 4, pp. 1092–1128, Jul. 2019, doi: 10.1177/1475921718788299.
[30] S. W. Canchumuni, A. A. Emerick, and M. A. C. Pacheco, "Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother," *Comput. Geosci.*, vol. 128, pp. 87–102, Jul. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098300419300378
[31] S. Lee, M. Kwak, K.-L. Tsui, and S. B. Kim, "Process monitoring using variational autoencoder for high-dimensional nonlinear processes," *Eng. Appl. Artif. Intell.*, vol. 83, pp. 13–27, Aug. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197619300983
[32] Z. Zhang, T. Jiang, C. Zhan, and Y. Yang, "Gaussian feature learning based on variational autoencoder for improving nonlinear process monitoring," *J. Process Control*, vol. 75, pp. 136–155, Mar. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095915241930037X
[33] X. Hou, K. Sun, L. Shen, and G. Qiu, "Improving variational autoencoder with deep feature consistent and generative adversarial training," *Neurocomputing*, vol. 341, pp. 183–194, May 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231219303157
[34] K. Wang, M. G. Forbes, B. Gopaluni, J. Chen, and Z. Song, "Systematic development of a new variational autoencoder model based on uncertain data for monitoring nonlinear processes," *IEEE Access*, vol. 7, pp. 22554–22565, 2019.
[35] J. Sun, X. Wang, N. Xiong, and J. Shao, "Learning sparse representation with variational auto–encoder for anomaly detection," *IEEE Access*, vol. 6, pp. 33353–33361, 2018.
[36] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," Dec. 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114
[37] D. P. Kingma, "Variational inference & deep learning: A new synthesis," Ph.D. dissertation, Fac. Sci., Inform. Inst., Univ. Amsterdam, Amsterdam, The Netherlands, Oct. 2017. [Online]. Available: https://hdl.handle.net/11245.1/8e55e07f-e4be-458f-a929-2f9bc2d169e8
[38] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," Jan. 2016, *arXiv:1601.00670*. [Online]. Available: https://arxiv.org/abs/1601.00670
[39] F. Cheng, Q. P. He, and J. Zhao, "A novel process monitoring approach based on variational recurrent autoencoder," *Comput. Chem. Eng.*, vol. 129, Oct. 2019, Art. no. 106515. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098135418309219
[40] Z. Zhang, T. Jiang, S. Li, and Y. Yang, "Automated feature learning for nonlinear process monitoring—An approach using stacked denoising autoencoder and k-nearest neighbor rule," *J. Process Control*, vol. 64, pp. 49–61, Apr. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095915241830026X
[41] Y.-R. Wang, Q. Jin, G.-D. Sun, and C.-F. Sun, "Planetary gearbox fault feature learning using conditional variational neural networks under noise environment," *Knowl.-Based Syst.*, vol. 163, pp. 438–449, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705118304568
[42] M. Mastroleo, R. Ugolotti, L. Mussi, E. Vicari, F. Sassi, F. Sciocchetti, B. Beasant, and C. Mcilroy, "Automatic analysis of faulty low voltage network asset using deep neural networks," *J. Eng.*, vol. 2018, no. 15, pp. 851–855, Oct. 2018.

[43] Y. Luo, Z. Li, and H. Wang, "A review of online partial discharge measurement of large generators," *Energies*, vol. 10, no. 11, p. 1694, Oct. 2017. [Online]. Available: https://www.mdpi.com/1996-1073/10/11/1694
[44] C. Hudon and M. Belec, "Partial discharge signal interpretation for generator diagnostics," *IEEE Trans. Dielectr. Electr. Insul.*, vol. 12, no. 2, pp. 297–319, Apr. 2005.
[45] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[46] D. Coomans and D. L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules," *Analytica Chimica Acta*, vol. 136, pp. 15–27, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0003267001953590
[47] S. Barrios, D. Buldain, M. P. Comech, I. Gilbert, and I. Orue, "Partial discharge classification using deep learning methods—Survey of recent progress," *Energies*, vol. 12, no. 13, p. 2485, Jun. 2019.
[48] G. Li, X. Wang, X. Li, A. Yang, and M. Rong, "Partial discharge recognition with a multi–resolution convolutional neural network," *Sensors*, vol. 18, no. 10, p. 3512, Oct. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/10/3512
[49] R. Zemouri, N. Omri, F. Fnaiech, N. Zerhouni, and N. Fnaiech, "A new growing pruning deep learning neural network algorithm (GP-DLNN)," *Neural Comput. Appl.*, May 2019, doi: 10.1007/s00521-019-04196-8.
[50] B. Pérez-Sánchez, O. Fontenla-Romero, and B. Guijarro-Berdiñas, "A review of adaptive online learning for artificial neural networks," *Artif. Intell. Rev.*, vol. 49, no. 2, pp. 281–299, Feb. 2018.

RYAD ZEMOURI was born in Algiers, Algeria, in 1974. He received the engineering degree in electronics and control science from the University Mouloud Mammeri of Tizi-Ouzou, Algeria, in 1998, and the post-graduated Diploma, and the Ph.D. degree from the Franche-Comté University of Besançon, France, in 2000 and 2003, respectively.

Since 2003, he has been an Associate Professor with the Engineering Department of Automatic Control, Electronics and Computer Science, Conservatoire National des Arts et Métiers, Paris, France. He was a Research Fellow with the Institut de recherche of Hydro-Québec (IREQ), Varennes, Canada, and the École de Technologie Supérieure, Montréal, Canada, for six months, from February 2019 to August 2019. His research interests include machine learning for safety-critical prognosis, diagnosis, and health management.

MÉLANIE LÉVESQUE was born in Chicoutimi, Québec, Canada, in 1980. She received the Ph.D. degree in electrical engineering from the École de Technologie Supérieure (ÉTS), Montréal, in 2012. She is currently working as a Research Scientist on generator diagnostics and prognostics with the Research Institute of Hydro-Québec (IREQ).

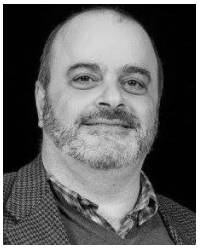
NORMAND AMYOT was born in Sorengo, Switzerland, in 1964. He received the M.Sc.A. degree in physics engineering from the École Polytechnique de Montréal, in 1990. He has worked one year at the Université des Sciences et Techniques (USTM) de Masuku, Gabon, after which he joined the Research Institute of Hydro-Québec (IREQ), where he is currently employed as a Senior Research Scientist. He is the author or co-author of more than 40 scientific articles. His fields of expertise include electrical insulation aging and characterization, diagnostic techniques, condition-based maintenance, and more recently, prognostics. He is an active member of CIGRÉ working groups and a member of the IEEE DEIS.

CLAUDE HUDON was born in Montreal, Québec, Canada, in 1963. He received the Ph.D. degree in engineering physics from the École Polytechnique de Montréal, in 1993.

He worked for two years at the Corporate Research and Development Center of General Electric, after which he joined the Hydro-Québec's Research Institute, Institut de recherche d'Hydro-Québec (IREQ), where he is currently employed as a Senior Researcher. His fields of interest are generator and motor diagnostics, the development of diagnostic tools, and the measurement and analysis of partial discharges. He is the author of more than 45 scientific articles.

OLIVIER KOKOKO graduated from the Electrical Department, Ecole Polytechnique de Montréal and Ecole de Technologie Supérieur.

He is currently a Researcher with the Institut de recherche d'Hydro-Québec (IREQ). His research area includes FEM modeling of hydrogenerator for diagnostic and prognostic.



ANTOINE TAHAN currently a Professor with the Department of Mechanical Engineering, École de Technologie Supérieure (ÉTS), Montreal, QC, Canada, since 2004. He is an active member of the Engineering Laboratory for Products, Processes, and Systems (LIPPS). He is specialized in dimensional metrology and geometrical characterization, tolerance management, and propagation of uncertainties. His research focuses on the aerospace and energy industries. He is working on the propagation of uncertainties in engineering models, diagnostic, and prognostic tools for complex electromechanical systems. He contributes to the development of probabilistic approaches to detect the state of degradation and damage in order to improve the reliability and design.