Deep Variational Autoencoder: An Efficient Tool for PHM Frameworks

Ryad Zemouri, Melanie Levesque, Normand Amyot, Claude Hudon, Olivier Kokoko

Deep Variational Autoencoder: an efficient tool for PHM frameworks.

Ryad Zemouri*, Mélanie Lévesque†, Normand Amyot†, Claude Hudon† and Olivier Kokoko†
*Cedric-Lab, CNAM, HESAM université, Paris (France), ryad.zemouri@cnam.fr
† Institut de Recherche d'Hydro-Québec (IREQ), Varennes, QC, J3X1S1 Canada

***Abstract*—Deep learning (DL) has been recently used in several applications of machine health monitoring systems. Unfortunately, most of these DL models are considered as black-boxes with low interpretability. In this research, we propose an original PHM framework based on visual data analysis. The most suitable space dimension for the data visualization is the 2D-space, which necessarily involves a significant reduction from a high-dimensional to a low-dimensional data space. To perform the data analysis and the diagnostic interpretation in a PHM framework, a Variational Autoencoder (VAE) is used jointly with a classifier. The proposed model was evaluated to automatically recognize individual Partial Discharge (PD) sources for hydro generators monitoring.**

***Index Terms*—Variational Autoencoder, data visualization, system health management, diagnosis analysis.**

I. Introduction

Deep Neural Networks (DNN) and Deep learning (DL) represent, nowadays, the most effective machine learning technology in recent applications of Machine Health Monitoring Systems (MHMS) and fault diagnosis [1], [2], [3], [4], [5]. The quality of the training data is an important factor that affects the performance of these DL architectures. In spite of their superior discrimination power in many fields, the DNN often lacks interpretability. Usually, it is very hard to understand why the classifier makes a given decision, and in what situations it is reliable [6]. The interpretability of a classifier for an efficient diagnosis is especially important in machine health monitoring systems. Therefore, most of the DNN classification models are considered as black-boxes. The interpretability of their internal processing mechanism through the hidden layers is usually challenging. Indeed, the high-dimensional space data transformation goes beyond the capacity of human interpretation [7].

In this paper, a new PHM framework based on visual data analysis and interpretation is presented (Fig 3). More reliable interpretations and analysis of classifier performance can be made by visualizing the 2D latent space of a VAE. This work is the progression of the results previously published [8] where the reader can find additional explanations. The paper is organized in three main sections. First, some theoretical backgrounds of the VAE are given. Then, the deep learning PHM framework is presented. Finally, some practical results obtained on hydro generators monitoring are detailed.

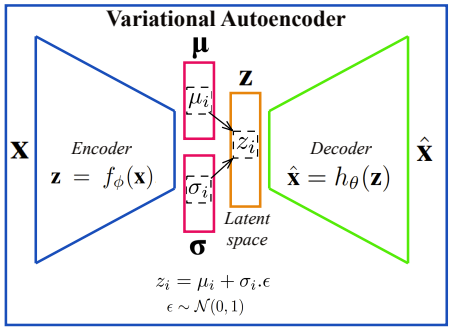

Fig. 1. Schematic architecture of a variational deep autoencoder.

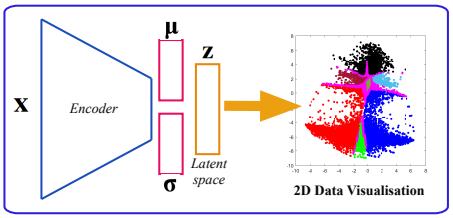

Fig. 2. Data visualization by dimension reduction.

II. Deep Variational Autoencoder for data visualization

For the data visualization, the most suitable space dimension for a human brain is the 2D-space. This necessarily involves data dimension reduction from a high-dimensional to a low-dimensional space. Classical methods from statistics such as principal components analysis (PCA) [9] and t-SNE [10] has been an important topic in visual analytics. More recently, the deep variational autoencoders (VAE) have been successfully used for feature dimension reduction [8], [11], [5].

A VAE is a special extension of the autoencoder (AE) which is an unsupervised Neural Network (NN) trained to reproduce the input vector $\mathbf{x}$ [12], [8]. The VAE is composed by two separate multilayered NNs: an encoder and a decoder as illustrated in Fig. 1, parameterized by ϕ and θ, respectively. The first NN encodes the input data $\mathbf{x}$ into a latent representation $\mathbf{z}$ by the encoder function $\mathbf{z} = f_\phi(\mathbf{x})$, whereas the second one decodes this latent representation onto $\hat{\mathbf{x}} = h_\theta(\mathbf{z})$ which is a reconstruction of the original data. In a VAE, an equal number of units are used in the input/output layers while less units are used in the latent space.

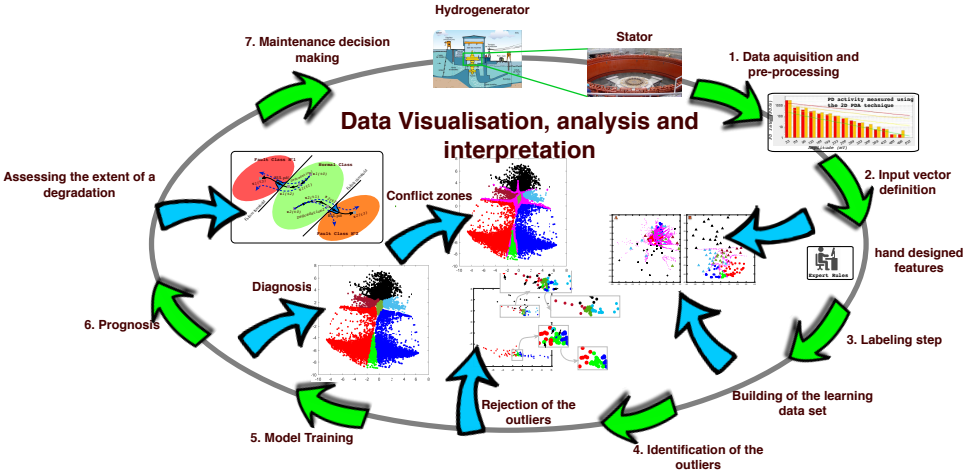

Fig. 3. The PHM framework used by Hydro-Quebec for hydro generator diagnosis and prognosis.

VAE becomes a popular generative model by combining Bayesian inference and the efficiency of the NNs to obtain a nonlinear low-dimensional latent space [8]. The Bayesian inference is obtained by an additional layer used for sampling the latent vector $\mathbf{z}$ with a prior specified distribution $p(\mathbf{z})$, usually assumed to be a standard Gaussian $\mathcal{N}(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. Each element z_i of the latent layer is obtained by $z_i = \mu_i + \sigma_i.\epsilon$, 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 $\mathbf{z}$, the VAE generates vector of means μ_i and standard deviations σ_i. This a major advantage that gives more continuity in the latent space than the original AE.

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 as seen in the 2D Data visualization in Fig. 2.

III. Deep Learning PHM Framework

Figure 3 shows the PHM framework used by Hydro-Quebec for hydro generator diagnosis and prognosis. This PHM framework is based on a main central function which is the visual data analysis and interpretation. Seven main steps are considered: 1. Data acquisition and pre-processing, 2. Input vector definition, 3. Labelling step, 4. Outliers identification, 5. Diagnosis model training, 6. Assessing the extent of a degradation, 7. Maintenance decision making. Some of these functions are detailed below.

A. Hand-designed features definition

The features definition in the machine health monitoring systems are categorized into conventional hand-crafted feature design against deep learning end-to-end feature design. The deep learning end-to-end structure enables the construction of the MHMS framework with less expert knowledge [1], but needs to have a significant amount of labeled data. The whole model including the feature extraction and the classification module are usually trained jointly. Conversely, the conventional hand-crafted feature design requires a significant amount of expertise from the practitioner, especially for complex systems, but needs less labeled data. We have seen in our previous study [8] that the more we put knowledge in the input vector, the more the data space becomes discriminant, thus making the classification easier. Separation between classes is transformed from non-linear separability with high overlapping clusters, as illustrated in Fig. 4.A, into a pseudo-linear one with less overlapping clusters as seen in Fig. 4.B. This second representation based on the conventional hand-crafted design with expert knowledge reduces the ambiguities and conflict areas between classes.

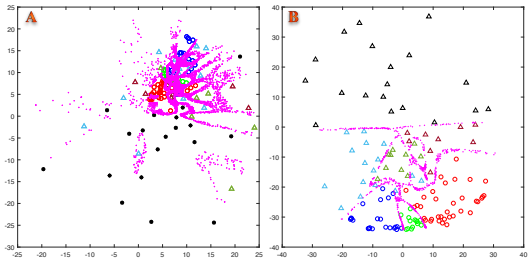

Fig. 4. The latent space obtained by a variational autoencoder [8]: A. deep learning end-to-end feature design without expert knowledge, B. Conventional hand-crafted feature design with expert knowledge.

B. Labeling the training dataset

One of the main problems facing all industries is the massive high dimensionality unlabeled data. To perform an intelligent classification of such amount of data, experts must first label several measurement files for the model training process. To do this, experts faced two challenging problems: how to select the most significant data for labeling? and what is the minimum data size required to complete the training process that would be sufficient to define each class?

One efficient way to help the PHM expert for selecting the most suitable instances for the labeling step is to visualize and analyze the spatial distribution of the available dataset. Selecting two instances with too close similarities, i.e. two close neighbor points in the 2D-space, is not very efficient for generalization during the model training. It is more efficient to select the training instances with a better spatial distribution, as we can see in the example in figure 5. In this illustration, there is more data in the first set comparing to the second one, but the spatial distribution of the second set is more efficient for the training process. Indeed, the area between the classes not covered by the training points is thinner in the second case, which gives less conflicting decisions for the classification. It should be noted that most of the false positive and false negative are located nearby or within the conflict areas (i.e. the boundaries between the classes).

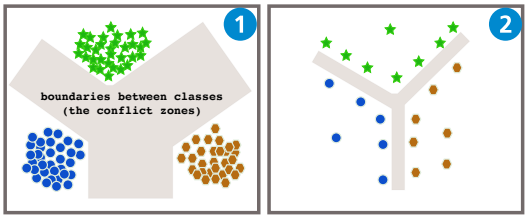

Fig. 5. Basis illustration of two training datasets with two different spatial distribution.

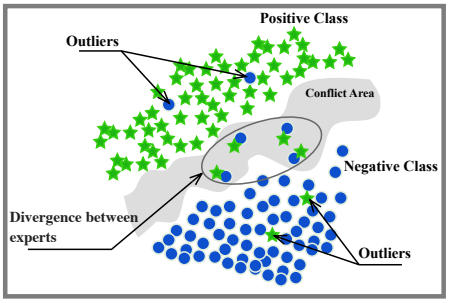

Fig. 6. Identification of the Outliers and the conflict zone.

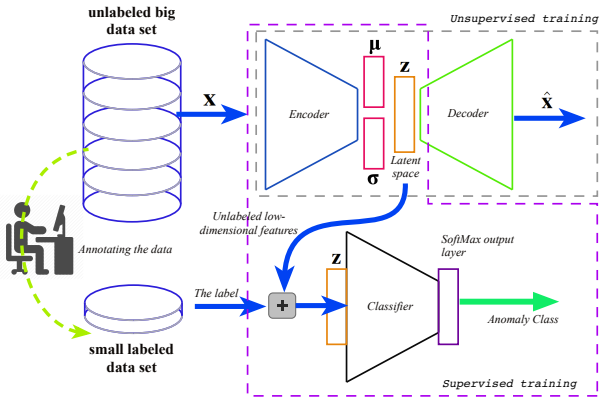

Fig. 7. Architecture of the proposed model.

C. Outliers identification

For an efficient MHMS, it is essential to analyze the quality of the training dataset in order to identify the outliers. These incoherencies are usually due to several problems with data acquisition, such as missing or bad attributes, where some of the parameters may have been incorrectly collected [13]. The second reason that can cause outliers is the human factor. Indeed, when labeling several measurement's files, it is not excluded that the expert may make errors of judgment for certain complex cases. In addition, contradictions between experts are generally encountered nearby or within the conflict zone [8]. These unusual data points are easy to identify in a low-dimensional space, as illustrated in figure 6. The classifier's predictions may be less reliable nearby these outliers. For these reasons, the identified outliers should be rejected from the training dataset.

D. Model training for data visualization and fault detection

Figure 7 shows the proposed architecture used for the data visualization and fault detection. Training of the whole model is split in two successive steps:

- (a) an unsupervised training step which consists of training the VAE, i.e the encoder and the decoder, to capture the input data features for data visualization,
- (b) followed by a supervised training step, which consists of training the combined architecture, i.e. the encoder and the classifier, for fault detection.

As the first unsupervised training step of the VAE does not require annotated instances, all unlabeled dataset could be used. For a better encoding of the original feature space into a low-dimensional latent space by the encoder, it is essential to have a database that is as representative as possible of the system's operating modes. For the second supervised training step, labeled data are required. The advantage of the proposed architecture is to consider the label information when training the encoder. Thus, the obtained latent space is better segmented with less overlapping between opposite clusters. For the first unsupervised training step, the Kullback-Liebler VAE loss function is used [8], while the categorical cross entropy loss function is used for the second supervised training step.

E. Assessing the extent of a degradation

In predictive maintenance, the system is continuously monitored over time t with multiple sensors. Considering that $\mathbf{x}(t)$ is the input feature vector and $\mathbf{z}(t)$ is its corresponding latent vector in the low- dimensional space. Except for a sudden breakdown, a degradation is characterized by a gradual loss of operating performances. Starting from an initial healthy state $\mathbf{x}(t0)$, respectively $\mathbf{z}(t0)$, the system will pass through several successive states $\mathbf{x}(ti)$, respectively $\mathbf{z}(ti)$, before the transition to an unhealthy state. This degradation phenomenon can be viewed on the latent space where the boundaries between the normal class and other unhealthy classes can be considered as the failure threshold. The RUL is then predicted with

the temporal analysis of the degradation. Figure 8 shows an illustration of assessing the extent of two degradations in the latent space. The more the system degrades and becomes close to the failure, the more the point $\mathbf{z}(ti)$ gets closer to the border with the unhealthy state.

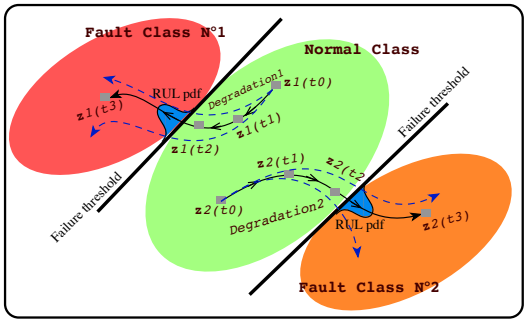

Fig. 8. Assessing the extent of a degradation for RUL prediction.

IV. Real application case study

A. Hydrogenerator monitoring

Hydro generators are strategic assets for power utilities. Their reliability and availability can lead to significant benefits. For decades, monitoring and diagnosis of hydro generators 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 hydro generator breakdown comes from failure of its high voltage stator, which is a major component of hydro generators. A study of all stator failure mechanisms reveals that more than 85 % of them involve the presence of PD activity [14]. PD are minute sparks that occur within voids inside the 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 [15], [16]. The impulses can be detected online from sensors connected to hydro generators, 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. Differentiation between PD sources is not straightforward and cannot only rely on simple quantification rules. In the present work, a methodology to automatically recognize individual PD sources from 2D PDA files was implemented using deep learning techniques.

B. Visual Partial Discharge analysis

In hydro generator diagnosis, it is important to determine if PD signal is coming from internal discharges (symmetry between positive and negative PD pulses) or from slot PD or corona at the junction of the stress grading coating (asymmetric in favor of positive PD pulses occurring during the negative voltage half-cycle).

In addition to the symmetry factor, an additional rule is used to determine if gap type discharges are active or not. Such activity is known to give a cluster of PD activity at higher amplitude thus resulting in bumps on the 2D plot. It should be pointed out that the bumps at higher amplitude sometimes affect both polarities, but other times can be more prevalent in one polarity. Thus, seven classes are defined according to the distribution of the PD pulses:

- PD source 1 (C_1): Negative Asymmetry, asymmetric in favor of negative PD pulses
- PD source 2 (C_2): Positive Asymmetry, asymmetric in favor of positive PD pulses,
- PD source 3 (C_3): Symmetry between positive and negative PD pulses,
- PD source 4 (C_4): Negative Asymmetry with Gap,
- PD source 5 (C_5): Positive Asymmetry with Gap,
- PD source 6 (C_6): Symmetry with Gap,
- PD source 7 (C_7): Gap.

A small part of PD measurements has been visually selected from the whole unlabeled PD database [8] and annotated by the experts of Hydro-Québec. The whole unlabeled database combined with the selected labeled PD measurements have been used to train the proposed architecture. The labeled PD measurements are illustrated in figure 9 in a 2D space given by the output of the VAE's latent layer. For some representative's points of the 2D-latent space, the corresponding PD representations are shown as a histogram of the discharge rate (PD/s) according to 16 channels of amplitudes (mV) where the positive PDs are in red and the negative PDs in yellow. The shape of each PD activity is given by the red and the yellow curve for respectively positive and negative PDs.

It is interesting to note that the latent space has been judiciously segmented by the VAE into several areas according to the shape of PD pulses distribution. Two main clusters are completely disjointed: a cluster of PD without the presence of Gap (classes C_1, C_2, C_3 which corresponds to histograms from #1 and #8) and a second cluster of PD where the Gap type occurs (classes C_4, C_5, C_6 and C_7 which corresponds to histograms from #9 and #18). Within each cluster, asymmetries in favor of negative PD pulses are located on the right side of the 2D space while the asymmetries in favor of positive PD pulses are located on the left side. Symmetries between positive and negative PD pulses are rather located in the middle of each cluster. By taking the example of the two extreme cases, i.e. histograms #1 and #8, where a great asymmetry in favor of negative PD pulses for the first case and in favor of the positive PD pulses for the second case. The closer we get to the center, i.e. the green points, the more this asymmetry disappears. The same behavior is obtained for asymmetry with Gap (see histograms from #9 and #14).

When the model training has been successfully done, the combined structure, i.e. encoder/classifier, was tested overall the unlabeled database. Figure 10 shows the 2D distribution

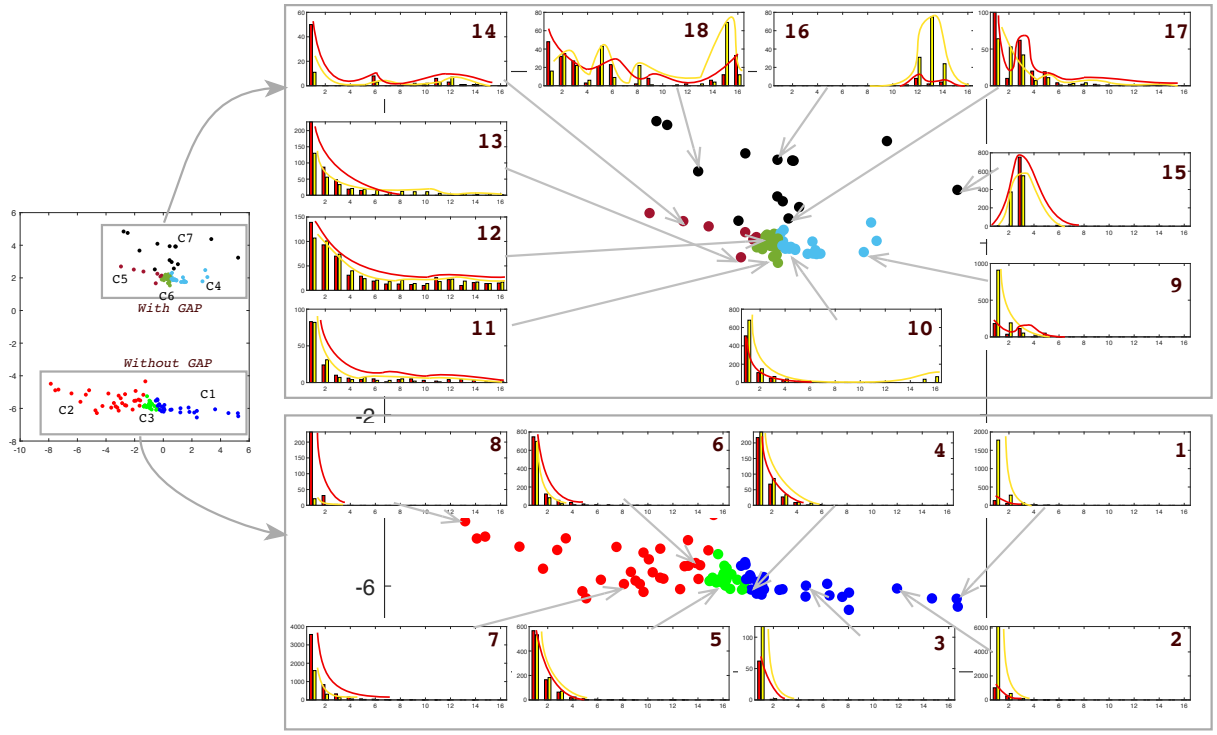

Fig. 9. Analyzing the training dataset through the 2D latent space.

of this database obtained by the encoder. The different colors correspond to the classes obtained by the classifier. Viewing the classification results on a 2D space is very comfortable for a human's brain. Each portion of this 2D space can be analyzed by a PHM expert in order to evaluate the performances of the given diagnosis. We focused on certain regions of this space by giving the histograms of the discharge rates. We can see that, depending on the position on the 2D latent space, the histogram has a particular signature. This signature is characteristic of its operating state. Degradation could be detected based on different positions at different times for a given hydro generator. Thus, by considering the 2D position jointly with the classifier response, this model offers to the PHM expert an efficient and a powerful tool for diagnosis and for maintenance decision-making.

V. Conclusion

In this paper, the use of a VAE used jointly with a classifier has been investigated for a new PHM framework. The proposed model has been used to classify Hydro-Quebec's entire PD database in seven classes. This paper presents an innovative approach to improve the interpretability of the diagnosis for machine health monitoring systems. This is achieved by utilizing the low dimension reduction of the VAE trained jointly with a classifier. The results show that the Deep Variational Autoencoder seems to be an efficient and a promising tool for PHM frameworks.

References

[1] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep learning and its applications to machine health monitoring," *Mechanical Systems and Signal Processing*, vol. 115, pp. 213 – 237, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327018303108

[2] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to rul prediction," *Mechanical Systems and Signal Processing*, vol. 104, pp. 799 – 834, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327017305988

[3] J. Zhang, P. Wang, R. Yan, and R. X. Gao, "Long short-term memory for machine remaining life prediction," *Journal of Manufacturing Systems*, vol. 48, pp. 78 – 86, 2018, special Issue on Smart Manufacturing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278612518300803

[4] H. Liu, J. Zhou, Y. Xu, Y. Zheng, X. Peng, and W. Jiang, "Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks," *Neurocomputing*, vol. 315, pp. 412 – 424, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231218308695

[5] S. Lee, M. Kwak, K.-L. Tsui, and S. B. Kim, "Process monitoring using variational autoencoder for high-dimensional nonlinear processes," *Engineering Applications of Artificial Intelligence*, vol. 83, pp. 13 – 27, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197619300983

[6] K. Xu, D. H. Park, C. Yi, and C. A. Sutton, "Interpreting deep classifier by visual distillation of dark knowledge," *CoRR*, vol. abs/1803.04042, 2018. [Online]. Available: http://arxiv.org/abs/1803.04042

[7] J. Wang, L. Gou, W. Zhang, H. Yang, and H. Shen, "Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 6, pp. 2168–2180, June 2019.

[8] R. Zemouri, M. Lévesque, N. Amyot, C. Hudon, O. Kokoko, and S. A. Tahan, "Deep convolutional variational autoencoder as a 2d-visualization tool for partial discharge source classification in hydrogenerators," *IEEE Access*, vol. 8, pp. 5438–5454, 2020.

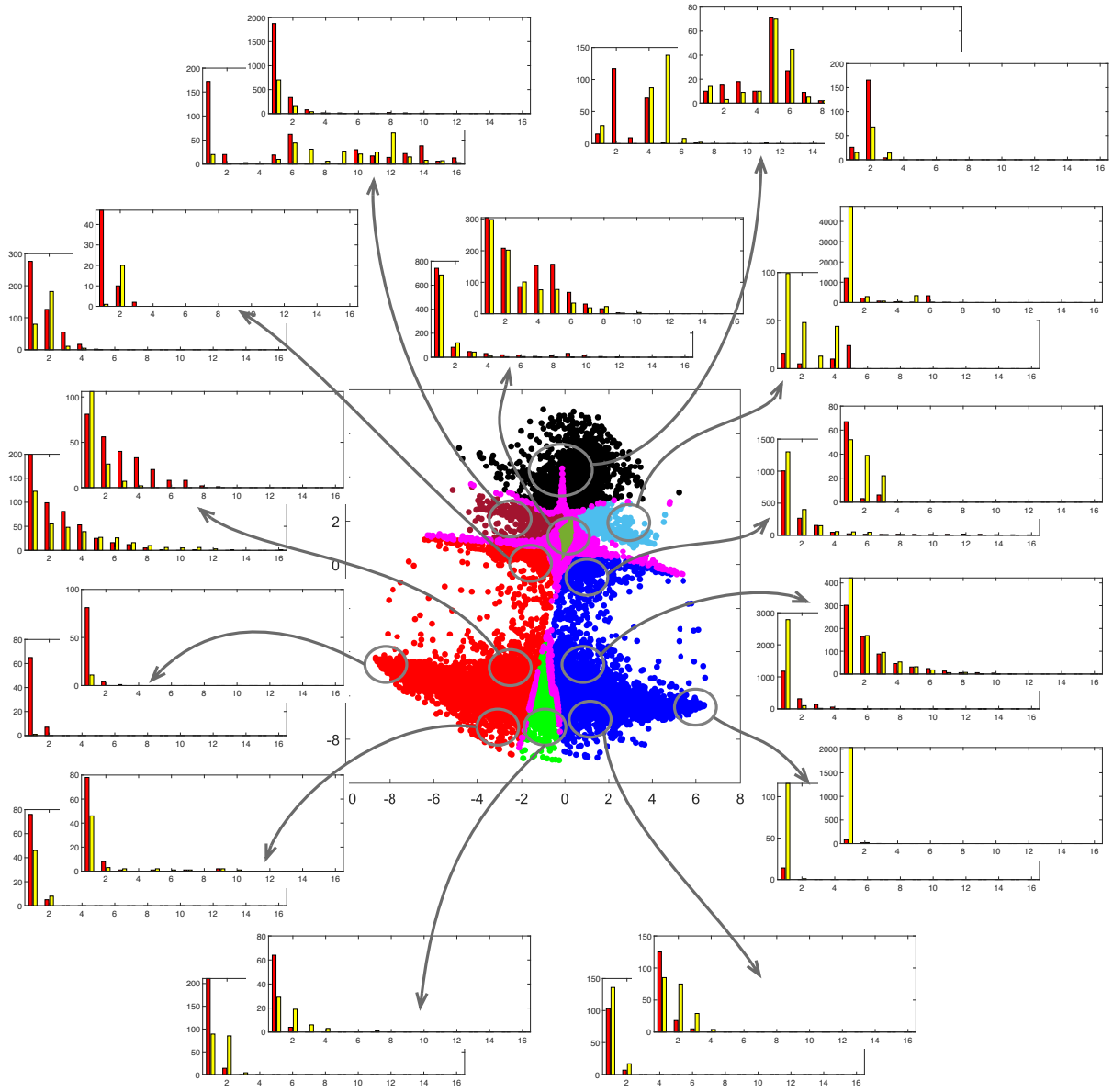

Fig. 10. 2D visualization and interpretation of the classification results obtained on the whole database.

[9] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1, pp. 37 – 52, 1987, proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0169743987800849

[10] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, p. 2579–2605, 2008.

[11] D. Kingma, "Variational inference & deep learning: A new synthesis," Ph.D. dissertation, Faculty of Science (FNWI), Informatics Institute (IVI), University of Amsterdam, Oct. 2017. [Online]. Available: https://hdl.handle.net/11245.1/8e55e07f-e4be-458f-a929-2f9bc2d169e8

[12] 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," *Structural Health Monitoring*, vol. 0, no. 0, p. 1475921718788299, 2018. [Online]. Available: https://doi.org/10.1177/1475921718788299

[13] S. Khan and T. Yairi, "A review on the application of deep learning in system health management," *Mechanical Systems and Signal Processing*, vol. 107, pp. 241 – 265, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0888327017306064

[14] M. Lévesque, N. Amyot, C. Hudon, M. Bélec, and O. Blancke, "Improvement of a hydrogenerator prognostic model by using partial discharge measurement analysis," in *Annual Conference of the Prognostics and Health Management Society 2017*, vol. 8, 2017, p. 7.

[15] Y. Luo, Z. Li, and H. Wang, "A review of online partial discharge measurement of large generators," *Energies*, vol. 10, no. 11, 2017. [Online]. Available: https://www.mdpi.com/1996-1073/10/11/1694

[16] C. Hudon and M. Belec, "Partial discharge signal interpretation for generator diagnostics," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 12, no. 2, pp. 297–319, April 2005.