



**HAL**  
open science

## Data augmentation-based prognostics for predictive maintenance of industrial system

Antonin Gay, Alexandre Voisin, Benoît Iung, Phuc Do Van, Rémi Bonidal, Ahmed Khelassi

### ► To cite this version:

Antonin Gay, Alexandre Voisin, Benoît Iung, Phuc Do Van, Rémi Bonidal, et al.. Data augmentation-based prognostics for predictive maintenance of industrial system. *CIRP Annals - Manufacturing Technology*, 2022, 71 (1), pp.409-412. <10.1016/j.cirp.2022.04.005>. <hal-03648868>

**HAL Id: hal-03648868**

**<https://hal.science/hal-03648868v1>**

Submitted on 22 Apr 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

# Data Augmentation-based Prognostics for Predictive Maintenance of Industrial System

Antonin Gay\*, \*\*, Alexandre Voisin\*, Benoit lung\* (1), Phuc Do\*, Rémi Bonidal\*\*, Ahmed Khelassi\*\*

\* Université de Lorraine, CRAN, UMR CNRS 7039, Campus Sciences, BP 70239, Vandœuvre-lès-Nancy, 54506, France

\*\* ArcelorMittal Maizières Research, Voie Romaine, Maizières-lès-Metz 57240, France

Anticipating system failures using predictive strategies based on efficient prognostics has become an important topic in manufacturing where maintenance plays a crucial role. As such, promising prognostics approaches use data-driven machine learning techniques, though the initial data set for learning is often small as failure occurrences are rare. Therefore, this study investigates data augmentation methods for improving prognostics by increasing data set size using samples generated by altering existing ones. First, a method is proposed for quantifying the gain from additional data. Thereafter, augmentation methods are assessed through a benchmark. Finally, contributions are illustrated in a steel industry case-study.

**Keywords:** Maintenance, Machine Learning, Data Augmentation

**DOI:** 10.1016/j.cirp.2022.04.005

## 1 Introduction

Maintenance plays a key role in ensuring asset availability and optimizing productivity of manufacturing plants. Preventive maintenance is standard with regularly scheduled operations in place to prevent failures. However, this method results in maintenance which is performed too soon and is not able to prevent unusual failures [1]. Thus, predictive maintenance [2] has gained attention. This method requires monitoring the system health in a diagnostic step and predicting its evolution by estimating the remaining useful lifetime (RUL) in a prognostics step. This RUL is then used to optimize maintenance decisions knowing that its uncertainty affects the decision optimization efficiency.

Many methods are available for prognostics, including data-driven ones, which are promoted today by the introduction of the Industry 4.0 paradigm [3]. Nevertheless, data-driven approaches need a significant amount of learning examples and for prognostics those examples mostly consist of recorded time series of the system degradation. Still, occurrences of failure and degradation are often rare, and a long-recorded process is necessary to collect a representative dataset. Therefore, new systems or newly instrumentalized ones present only a few samples. These **small datasets** reduce the efficiency and precision of data-driven prognostics which may lead to an increase of the RUL uncertainty impacting the decision-making optimization (e.g. too-soon reparation, un-avoided failure) and the lack of generalization to new failure modes, less represented in historical data [4]. Finally, from a practical point of view, physical models are rarely available for industrial systems, especially new ones. Thus, prognostics approaches must be with no prior system knowledge and end-to-end prediction models need to be considered.

To combat this problem, a newly emergent approach called **Data Augmentation** can be used to increase dataset size and improve prognostics results. However, it has not yet been fully developed to support efficient prognostics on industrial datasets [5] and is less represented in prognostics literature than other approaches such as meta-models. In addition, its use for prognostics end-to-end regression on time-series and run-to-failure datasets has not currently been considered.

Therefore, this study seeks to characterize the performance loss of small datasets, then evaluate the performances of general data augmentation methods for industrial prognostics, first on a simulated dataset, then an industrial case-study from the steel industry. Finally, to improve the performances of the industrial

dataset, a specific data augmentation method based on expert knowledge is proposed.

Following from here the remainder of the paper is structured as follows: Section 2 describes a problem statement for identifying issues on data augmentation for prognostics; Section 3 presents a method to combat these issues by using a simulation-based prognostics dataset; Section 4 introduces a steel industry case study to illustrate the proposed contributions; and finally conclusions and future works are presented in Section 5.

## 2 Problem Statement on Data Augmentation for Prognostics

The robustness and performance of data-driven or data-based approaches is directly related to the size of the dataset. Indeed, it is shown that smaller datasets may reduce the prognostics performance [6], but the impact of dataset size on the prognostics performance is still an open scientific issue which should be addressed. As a first scientific issue, this study seeks to assess what performance gain can be obtained by adding data to a small prognostics dataset. This can be achieved either by waiting for further data monitoring or by applying data augmentation in an emergent way. Data augmentation consists of generating new learning samples from an available dataset by applying specific modifications to the samples and is widely used in image processing for improving classification performance. Formally, existing augmentation methods can be specified into two main categories:

(a) Firstly, if no physical or functional model of the system is available, generation can be based on generic data-based methods as presented in [7] for time series classification. This review shows that data augmentation can improve classification predictions. However, these methods do not focus on data augmentation for prognostics. Therefore, the current study intends to extend the data augmentation benchmarks to the industrial system prognostics and to assess its performance gain on maintenance decision-making. This is identified as the second scientific issue.

(b) Secondly, if models of the system are available, model-based approaches consist of generating new samples through simulations [8]. But in the case of new systems, physics-based models are rarely available, and even when they are, the industrial complexity is usually too high to have a relevant model of the whole system. However, in practice, expert knowledge is often available during the system design step [9], e.g., thanks to functional analysis or Failure Mode and Effects Analysis (FMEA) [10]. Nevertheless, exploitation of this knowledge within certain data augmentation methods has not yet been attempted and is a third scientific issue to be solved in order to obtain efficient prognostics. This last issue

is addressed in Section 4 by proposing a knowledge-based method using data augmentation for prognostics while the two first issues are studied in Section 3 from an existing prognostics dataset.

### 3 Data augmentation for prognostics

The first two issues are explored using an existing prognostics dataset based on simulated turbofans [11]. The study focuses on a dataset that considers one degradation mode and one operational condition. A state-of-the-art prognostics model is used for end-to-end modeling of this dataset, which is based on Long Short-Term Memory (LSTM) networks [12]. Even still, it requires an important amount of training data.

#### 3.1 Prognostics performance on smaller datasets

To study the link between dataset size and the loss of prognostics precision, the model has been trained on subsets randomly picked among the original dataset with size  $i \in [75\%, 50\%, 25\%, 10\%]$ . The prognostics performances depend on the subsets data distributions and by repeating experiments on twenty random picks, representativity of the results can be ensured. When comparing subsets, the picks were performed such as, with  $D_{i\%}^{(j)}$  the  $j$ -th pick of the subset of size  $i$ ,  $D_{i_1\%}^{(j)} \subset D_{i_2\%}^{(j)}, \forall i_1 < i_2, \forall j$ . The network was then trained twenty times for each pick with random weight initializations in order to evaluate the robustness of the learning. Finally, the results were evaluated on a common test set never seen during the training step [11] and the performance was measured using root mean squared errors (RMSE) similar to other works conducted on this dataset [12]. The results are shown as boxplots for each subset size in Figure 1.

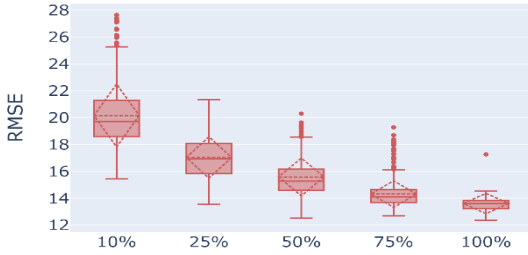


Figure 1 – Prognostics performances depending on the size of the training subset. The rectangles correspond to boxplots: middle bar is the median, edges are the first and third quartiles (Q1-Q3), bars are the span of values and points are outliers. The dashed diamonds are the  $mean \pm variance$ .

Figure 1 shows both a decrease in average precision and an increase in the dispersion between trainings as the number of samples decreases. It is noted that the increase in the dispersion may be caused by the model's hyperparameters being optimized for the full dataset.

The ability to predict the increase in performance as the number of samples increases would be useful. So, it is found relevant to fit a function  $RMSE_{median} = f(n_{samples})$  to the datasets. The median RMSE is used to reduce the impact of dispersion and outliers on small datasets. The following model is proposed in the literature for other kinds of predictions, for instance as [13] for classifiers, with  $n \in [0\%, 100\%]$  the dataset size:

$$RMSE(n) = RMSE^* + c_1 * \frac{\log(c_2 * n) + 1}{2^{c_2 * n}} (1)$$

The model was fitted on the previous median results from 0% to 75% and validated on 100%. This corresponds to how augmentation expected gain would be predicted in a new application. Thus, the following fitting was obtained:

$$RMSE^* \cong 11.5 \quad c_1 \cong 4.5 \quad c_2 \cong 0.21$$

After fitting from 0% to 75%, the RMSE value at 100% was estimated with an error of 0.013, which is an acceptable result as  $RMSE(100\%) \cong 13.6$ . In conclusion, this calculation showed the suggested model is relevant for evaluating the gain obtained by increasing the size of a dataset in prognostics. The function is then fitted when new prognostics applications must be considered.

#### 3.2 Benchmark of generic data augmentation for prognostics

To gain insights about method selection, the performance of the state-of-the-art machine learning model was used as a benchmark for the dataset augmented by several methods. This is done on the 25% subset, which presented an important performance drop and the 10% subset was not selected due to its many outliers which would reduce the ability to interpret the results. All further experiments were performed based on the protocol of subsection 3.1, where the same picks were used to ensure consistency and comparisons between results. The data augmentation methods tested were selected from the list proposed by the review in [7] among which a selection has been done based on the type of data present for turbofans application. Frequency-based methods were ignored because they could not adapt to this non-vibratory dataset. Similarly, generative models requiring large amounts of data or expert knowledge have been put aside. Three families of method have been selected:

1. Random magnitude transformations, including jittering and magnitude warping. With  $\mathbf{x} = [x_1, \dots, x_t, \dots, x_T]$  a nominal time series trajectory and  $\mathbf{x}'_{method}$  the augmented trajectory using the augmentation *method*:

$$\begin{cases} \mathbf{x}'_{jittering} = x_1 + \epsilon_1, \dots, x_t + \epsilon_t, \dots, x_T + \epsilon_T & \text{with } \epsilon_1, \dots, \epsilon_T \in \mathbb{R} \\ \mathbf{x}'_{mag. \text{warping}} = \alpha \mathbf{x} & \text{with } \alpha \in \mathbb{R} \end{cases}$$

2. Pattern mixing methods in the form of samples interpolation:

$$\mathbf{x}'_{interpolation} = \mathbf{x}_a + \lambda(\mathbf{x}_b - \mathbf{x}_a) \quad \text{with } \lambda \in [0,1]$$

3. Random time-based transformations, including time slicing and time warping; time permutation was avoided because it was not suitable to run-to-failure data:

$$\begin{cases} \mathbf{x}'_{time \text{ slicing}} = x_\phi, \dots, x_t, \dots, x_{W+\phi} & \text{with } 1 \leq \phi \leq T - W \\ \mathbf{x}'_{time \text{ warping}} = x_{\tau(1)}, \dots, x_{\tau(t)}, \dots, x_{\tau(T)} & \text{with } \tau(\cdot) \text{ warping func.} \end{cases}$$

Unlike specific data augmentation approaches, the generic nature of these methods can be applied to a sample without feature selection or system knowledge. Therefore, these methods are more relevant to prognostics with no prior knowledge.

The performance of each method was then evaluated on the 25% subset by individually applying it to the trajectories and training the model on this augmented dataset. For each original sample of the 25% dataset, two additional samples were generated with different values of the augmentation parameters. Those parameters' ranges were optimized to acquire the best performance for each method. The results presented in Figure 2 show that all but one method (magnitude warping) improved the prediction performances. Both the median and dispersion of the error are reduced, improving the prediction accuracy and the regularity among trainings, i.e., the robustness of the learning algorithms respectively. It must be noticed that magnitude warping decreased the performances and this negative effect confirms the importance of method selection. The results are analyzed to provide insight for method selection.

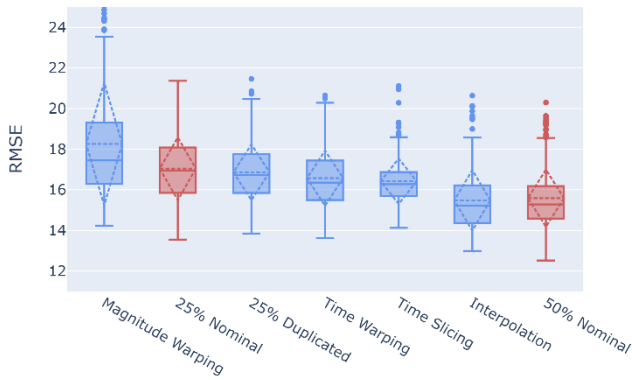


Figure 2 – Performances of data augmentation. Red ones are un-augmented datasets. “Duplicated” corresponds to an augmentation where some trajectories are just copied, to serve as reference.

Prognostics learning data is run-to-failure for multi-variate time series, where the last samples correspond to the failure states. Just before the failure, several signals approach critical thresholds. Therefore, any amplitude modification might obtain signals over this limit and generate non-physics related information. This would decrease the performance on future data. This hypothesis is verified by the poor results seen for magnitude warping, which reduces the performance. Contrarily, time warping is adapted to prognostics as it can model the increasing or decreasing of degradation speed. It is in agreement with the diversity of trajectories met in such datasets and serves to enrich them. The results validate this insight, with time warping being one of the most efficient methods tested, with an average precision improvement of 4%. Finally, pattern mixing methods can be quite efficient if the considered degradation trajectories can be combined. This property is highly dependent on the actual case and would need extra care. The results confirm this hypothesis, with exceptionally efficient results for interpolation, showing an average precision improvement of 11%. Indeed, the present dataset includes only one degradation mode and one operating condition. Additionally, model characteristics can also be considered to select methods. LSTMs predictions at time  $t$  depend on previous predictions and internal states at  $t - 1$ . Hence, learning on the first samples of a trajectory is different than the rest. Augmentation methods supplying new origins of time to LSTMs can thus be efficient. This hypothesis is verified by the good results of time slicing.

Eventually, an improvement of the average RMSE on the RUL from 17 to 15 is observed. The predicted RUL corresponds to the number of working cycles remaining before failure, hence the prognostics accuracy improved up to 2 working cycles. When applied to decision-making, this improvement is large enough to allow for maintenance decisions to be postponed to a more appropriate time. Similar results were obtained using a simpler Support Vector Machine (SVM), but to a lesser extent.

In conclusion, for prognostics applications, time-based data augmentations are quite efficient when modifying the temporality of the problems and supplying new relevant trajectories. On the contrary, amplitude-based data augmentations modify the physical properties of the problem and tend to degrade performances. Finally, pattern mixing-based methods are efficient but need extra care to ensure their physical representativity.

#### 4 Case study: descaling hydraulic circuit

The considered data augmentation methods are now illustrated on the hot strip mill (HSM) process of an ArcelorMittal steel plant (Florange, France). Each stop on the line is financially (10k€/h) and environmentally (CO<sub>2</sub>) expensive as the reheating furnaces are not shut down. So, the main maintenance objective at ArcelorMittal is

to reduce unwanted stops. The case study is focused on the descaling process of HSM which is a mandatory step for product quality and thus any component failure leads to an immediate line stop. Technically speaking, descaling involves high-pressure water projected onto hot steel plates which removes scale, a surface oxide. The hydraulic circuit is composed of standard components: pumps, valves, pipes, and nozzles with the system sensors monitoring pressure, temperature, flowrate, vibrations, etc. Regularly planned stops of various lengths are available for process and maintenance operations. Decision-making mostly consists in selecting the best time for a stop to perform maintenance operations before failure. As a result, decision-making needs consistent RUL to effectively decide whether to skip an available stop for the next one or not. Given the available instrumentation, data-based approaches are preferred for prognostics as physical models are not precise enough on such complex systems under high constraints. However, new pumps have been recently installed so only a limited amount of historical data is available. Hence, data augmentation is expected to provide improvement of the learning algorithms related to the pumps. Firstly, thanks to the previous analysis, expected improvements from additional data are estimated. Then, selective data augmentation is applied to improve the pump prognostics and start predictive maintenance before the end of failure data gathering. In addition, the third issue mentioned in section 2 regarding “knowledge-based augmentation”, is addressed by presenting a methodology of data augmentation combining knowledge from functional analysis and FMEA.

##### 4.1 Potential improvement from augmentation/additional data

The current failure dataset of the pumps consists of 16 trajectories composed of two failure modes under one operating condition. Since only a limited number of trajectories are available, a simpler Support Vector Machine (SVM) regressor model is used to predict the RUL of the pumps which also serves to reduce the model development time. Given the small amount of data, a test set could not be defined without reducing the available training data, hence the following results are obtained using cross validation. Thanks to equation (1), the expected performances are estimated. The prognostics RMSE is computed for various dataset sizes, and equation (1) is fitted as shown in Figure 3. The fitted coefficients are, with  $n$  the number of samples:

$$RMSE^* \cong 23.1 \quad c_1 \cong 1.3e15 \quad c_2 \cong 2.5e14$$

The fitting shows that the RMSE could eventually get down to 23 hours, as opposed to 47 with the 16 samples evaluated. Moreover, if data augmentation is used to triple the dataset size as done previously, the RMSE could be improved by up to 10 hours. As regular stops on the line happen every 8 hours, this gain could improve the maintenance decision range by one full working cycle. Nevertheless, those results shall be considered with care as the fitting has been done on a small number of samples. However, they justify the implementation of data augmentation on this use-case to improve prognostics before further data gathering.

##### 4.2 Data augmentation on industrial use-case

From the conclusion on data augmentation efficiency for prognostics, the previous methods have been applied and adapted to signal specificities. Pressure signals were not modified in amplitude as the pumps are regulated. Time warping was not applied to vibration signals to avoid altering frequency information. For each method, 32 additional trajectories were added, and the augmented dataset was then used to train the SVM model. The obtained performances on the augmented dataset are presented in Figure 4, which shows an improvement of average prognostics prediction up to 10 hours.

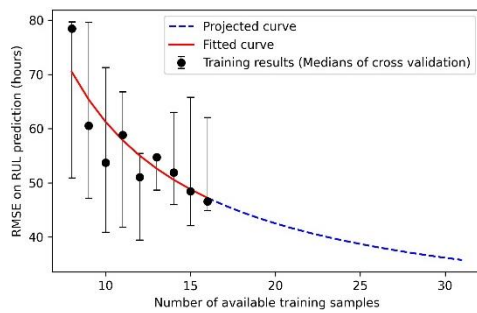


Figure 3 – Prognostics RMSE against number of samples in dataset for use-case dataset. Points are medians and error range correspond to Q1-Q3.

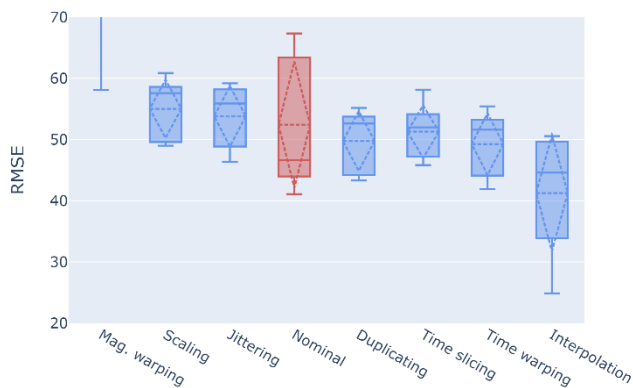


Figure 4 – Results of prognostics with augmented dataset on industrial dataset. Sorted by average RMSE except for duplicating, set as a reference.

According to the predictions from Figure 3, this improvement would equate to of 50% increase in data within the dataset. Furthermore, though the average and dispersion are significantly improved, the median is not. More importantly, data augmentation is shown to significantly reduce the dispersion of prognostics performances. For industrial companies, this provides more confidence in the prognostic's predictions.

In the case of ArcelorMittal decision-making, this gain of precision would allow more flexibility when choosing stops for maintenance operations, and thus reduce the risk of increased stop times or the risk of stopping the line specifically for maintenance.

#### 4.3 Perspective of expert knowledge-based approaches.

Finally, expert knowledge integration into data augmentation for prognostics in industrial applications is addressed.

Previous results have shown data augmentation is limited when applied to a very small industrial dataset. Consequently, the next step for prognostics data augmentation is to develop a system-specific data augmentation method. Advanced simulation models are generally not available on such systems. However, expert knowledge is normally available and can be used to improve data augmentation. For instance, functional analysis (FA) and Failure Mode Effects Analysis (FMEA) are often available.

These analysis methods contain qualitative knowledge about links between components / signals and signal evolutions for subsequent failure modes. They are not complex enough for data generation but may guide data augmentation. This is illustrated with a pipe leakage located at the pumps output. The FMEA states that, as the new pumps are pressure-regularized, the pressure would not fall but their power consumption would rise. The FA links a power increase to a vibration modification, modeled by a local physics model as a frequency shift linked to the power increase. Therefore, new trajectories corresponding to this degradation

mode would apply augmentation methods accordingly to those specificities. Therefore multiple new trajectories can be generated by using a distribution of magnitude scaling parameters. In conclusion, the proposed knowledge-based approach would provide additional data based on system properties and characteristics. Due to the lack of data, they cannot be evaluated, but do present potential for prognostics, where it was previously identified that the best augmentation methods are the ones adapted to system characteristics.

For that reason, further experiments should be conducted to validate the preliminary results given by this use-case.

## 5 Conclusions and perspectives

In this paper, the issues regarding small datasets in the case of prognostics for industrial systems were addressed through the data augmentation technique. First, the potential gain of such approaches was evaluated and a model allowing for quantification of the impact of dataset size on the prognostics accuracy was proposed. It has been tested on a turbofan dataset and used on an industrial case-study. Furthermore, several data augmentation methods have been implemented and analyzed on the turbofan dataset [11]. Their efficiency was assessed and linked to their adaptability to the dataset features and prognostics aim. In the process a subset of the dataset was augmented and reached similar performances to that of a subset twice its size. This work was validated on the industrial use-case, whose performances were significantly improved. Finally, an augmentation methodology was sketched using expert knowledge. The results illustrate the importance of augmentation methods parametrization. Further work will focus on a deeper characterization of the parametrization impact of data augmentation as well as the development of the knowledge-based augmentation methodology.

## References

- [1] Roy, R., Stark, R., Tracht, K., Takata, S., Mori, M., 2016, Continuous Maintenance and the Future – Foundations and Technological Challenges, *CIRP Annals - Manufacturing Technology*, 65/2:667–688.
- [2] Sahli, A., Evans, R., Manohar, A., 2021, Predictive Maintenance in Industry 4.0: Current Themes, *Procedia CIRP*, 104:1948–1953.
- [3] Gao, R.X., Wang, L., Helu, M., Teti, R., 2020, Big Data Analytics for Smart Factories of the Future, *CIRP Annals*, 69/2:668–692.
- [4] Tiddens, W.W., Braaksma, A.J.J., Tinga, T., 2015, The Adoption of Prognostic Technologies in Maintenance Decision Making: A Multiple Case Study, *Procedia CIRP*, 38:171–176.
- [5] Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., Ducoffe, M., 2020, Potential, Challenges and Future Directions for Deep Learning in Prognostics and Health Management Applications, *Engineering Applications of Artificial Intelligence*, 92:103678–103692.
- [6] Catal, C., Diri, B., 2009, Investigating the Effect of Dataset Size, Metrics Sets, and Feature Selection Techniques on Software Fault Prediction Problem, *Information Sciences*, 179/8:1040–1058.
- [7] Iwana, B.K., Uchida, S., 2021, An Empirical Survey of Data Augmentation for Time Series Classification with Neural Networks, *PLOS ONE*, 16/7:e0254841.
- [8] Ozdagli, A.I., Koutsoukos, X., 2021, Model-based Damage Detection through Physics Guided Learning, *Annual Conference of the PHM Society 2021*, 13/1. DOI: 10.36001/phmconf.2021.v13i1.3012
- [9] Xiahou, T., Zeng, Z., Liu, Y., 2021, Remaining Useful Life Prediction by Fusing Expert Knowledge and Condition Monitoring Information, *IEEE Transactions on Industrial Informatics*, 17/4:2653–2663.
- [10] Laloix, T., Voisin, A., Deeb, S., Romagne, E., Lung, B., Lorange, F., 2017, Industrial System Functioning/Dysfunctioning-Based Approach for Indicator Identification to Support Proactive Maintenance, *IFAC-PapersOnLine*, 50/1:13704–13709.
- [11] Saxena, A., Goebel, K., 2008, Turbofan Engine Degradation Simulation Data Set, NASA Ames Prognostics Data Repository (<http://ti.arc.nasa.gov/project/prognostic-data-repository>).
- [12] Chaoub, A., Voisin, A., Cerisara, C., Lung, B., 2021, Learning Representations with End-to-End Models for Improved Remaining Useful Life Prognostic, *PHM Society European Conference*, 6/1:1–8.
- [13] Vapnik, V., Levin, E., Cun, Y.L., 1994, Measuring the VC-Dimension of a Learning Machine, *Neural Computation*, 6/5:851–876.