engineering proceedings

Proceeding Paper

Getting a Better Sense of Data Drift in Dynamic Systems: Sequence-Based Deep Learning for Monitoring Slowly Evolving Degradation Processes [†]

Tarek Berghout [1,*] and Mohamed Benbouzid [2,3]

[1] Laboratory of Automation and Manufacturing Engineering, University of Batna 2, Batna 05000, Algeria
[2] UMR CNRS 6027 IRDL, University of Brest, 29238 Brest, France; mohamed.benbouzid@univ-brest.fr
[3] Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China
[*] Correspondence: t.berghout@univ-batna2.dz

[†] Presented at the 10th International Electronic Conference on Sensors and Applications (ECSA-10), 15–30 November 2023; Available online: https://ecsa-10.sciforum.net/.

Abstract: Deep Learning (DL) for monitoring slowly evolving degradation processes typically involves overcoming data drift, complexity, and unavailability issues resulting from dynamic and harsh conditions, and rarity of labeled failure patterns, respectively. While degradation patterns are mostly hidden in such complex data, observation-based DL leans towards producing uncertain predictions and/or overfit the model during training process. This problem is usually caused by the insignificance of certain data representations. Therefore, and particularly due to the sequential nature of data in such a degradation process, it is necessary to consider neighboring observations to judge the accuracy of its representation or improving it. In this context, instead of traditional observation-based learning philosophy, this paper presents data-driven sequential mapping, while health indices can also be represented as a vector of sequential data and not as a single regressor output changing the model's architecture. Using a dataset generated from a mathematical model mimicking bearing degradation life cycles and responding to the aforementioned three main challenges, a comparative study is built on investigating observation-based and sequence-based learning paths. According to a well-defined visual and numerical evaluation criterion, a sequence-based methodology reflects a better understanding of data representations through parameter tuning reaching better approximation and generalization. Such results support the necessity to such learning mechanism, especially for sequential data, dealing with some sort of correlation, and degrade controversially. Necessary files to reproduce the findings of this work are made available at: https://doi.org/10.5281/zenodo.8142676.

Keywords: bearing; deep learning; degradation; prognostics and health management; remaining useful life; sequential data; vibration

Citation: Berghout, T.; Benbouzid, M. Getting a Better Sense of Data Drift in Dynamic Systems: Sequence-Based Deep Learning for Monitoring Slowly Evolving Degradation Processes. *Eng. Proc.* **2023**, *56*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2023

1. Introduction

Monitoring of slowly degradation processes of a dynamic systems under real conditions based on DL is generally a problem of building a regression model where a specifically reconstructed health index need to be predicted accurately for unseen health indicators [1]. This usually poses a problem of data drift, complexity and unavailability [2]. The concept drift refers to massive changes in historical data features of a specific system lifecycle (i.e., run-to-failure data) [3]. Similarly, data complexity, and unavailability refers to different kind of distortions and rarity of failures patterns [4]. Such distortions could be the results of presence of noise and different outliers/anomalies in data affected by environmental conditions or physical damage propagation of the system itself. In the meanwhile, rarity of failure patterns is generally due to the fact that data is most of time generated from physics-based models or accelerated aging experiments and not true degradation phenomena for many reasons including financial and critical safety-related issues reducing emulation quality to reality [5].

It should be mentioned accordingly that this paper focuses on data drift problem while the DL model is required to be continuously update to meet up new changes in data and generalize better for unseen samples. As a result, research gaps in this paper will be revealed based upon analysis from this perspective. Basically, DL models for health monitoring are generally constructed based on ordinary training process of mapping each observation features separately towards outputs driven at each time instant. This means when for instance an observation is miss presented due to any possible data distortion, sensors malfunction, or any other possible disturbances, the model will automatically be affected and may lead to bias, misprediction, overfitting, etc. [3]. In this case, it is necessary to mitigate such misleading information to maintain both approximation and generalization process of the DL model [6].

1.1. Research Gaps

A according to the brief previous analysis, gap in research in this case can consequently be highlighted as follow:

1. Observation-based learning doesn't consider correlation between times series data which could lead the model to bias if samples are mispresented due many aforementioned reasons;
2. Observation-based learning doesn't reflect the actual monitoring of concept drift and its detection at some point while data is subject to continuous change.

Overall, a single-observation even if driven in a form of chunk-by-chunk is not expected to carry information to the learning model itself about neighboring samples. This is a true learning problem especially when slowly evolving degradation process monitoring is a time series analysis problem and should considers this fact [7].

1.2. Contributions

Based on highlighted analysis criteria of concept drift in dynamic systems for monitoring slowly evolving degradation processes, the following contributions are proposed in this paper.

1. **Considering a sequence-based learning methodology:** one of the main solutions that this paper proposes is to follow a sequence-based learning methodology for such mission. In this case, a sequence of observations of a specific length will be flattened and used as an input to the DL network. It should be mentioned that this is different from sequence-to-sequence learning presenting a series of encoding-decoding patterns and processes as proposed in [8]. Therefore, the output of the DL regressor will be a vector instead of a single health index during sequential mapping. The tuning mechanism of the DL model will make it possible to get a sense of data changes and to improve its representations taking into account the loss result;
2. **Considering adaptive deep learning**: an additional step of adaptive deep learning is taken into account in this case by introducing long-short term memory (LSTM) neural network. LSTM has a strong advantage as it allows for considering correlation between driven sequences of time series data. In another way, data drift in this case will be treated into two main steps, (i) preprocessing step where data is organized in sequences instead of observations, and (ii) where the learning algorithm itself considers adaptive learning. This will further strength the learning processes an introduces more accurate adaptive learning;
3. **Using data generated from a mathematical model:** an experiment will be conducted on data generated from a mathematical model mimicking health degradation trajectories of bearings responding to the three aforementioned health monitoring issues of slowly evolving degradation processes [4]. Compared to traditional observation-

based DL, experiments encourage such data mapping especially for sequential data as of similar degradation behavior. Necessary files to reproduce the findings of this work are made online available at [9].

1.3. Outlines

In attempt to make sure that contributions of this paper are clear and well-illustrated making this study reproducible, this paper is organized as follow. Section 1 is dedicated to data description. Section 2 is devoted to proposed methodology and used methods. Section 3 is dedicated to results and discussion. Finally, Section 4 concludes this work.

2. Materials

This work used a bearing degradation dataset extracted from a mathematical model describing an exponentially growing sinusoidal with additive noise and distortions in an attempt to mimic real-world conditions [10]. The dataset contains two subsets dedicated to outer race and inner race faults where each subset contains 20 sequences with different degradation rate. The degradtion rate is defined using different number of vibration window sizes with 30, 50, 65, 80, and 100 time cycles, respectively. For each speed profile, 4 life cycles are generated modifying the parameters influencing the fault signatures randomly to +/− 5%. Each window has 16,348 number of samples with a sampling frequency of 51.2 kHz. Meaning that we have 40 life cycles in total. Row-data from a single life cycle from the dataset is presented in Figure 1a clearly showing data exponential shift towards failure mode.

In this work, data is subjected to preprocessing making it easier to extract any possible degradation signs at first glance. In this context, 11 time domain features same to the ones used in previous work in [6] (see [6], Section 2.3) are extracted. Similarly, the same denoising, outlier removing, and scaling steps in [6] are also followed to make sure the data is ready for DL model training. Accordingly, Figure 1b is an example of extracted features from life cycle in Figure 1a as some degradation patterns clearly can be seen in this situation. Based on data visualization in Figure 1a,b, the health index function is defined as an exponential degradation function (see Equation (2) from [11]) better reflecting degradation mechanism then linear trends which also can be seen in Figure 1c.

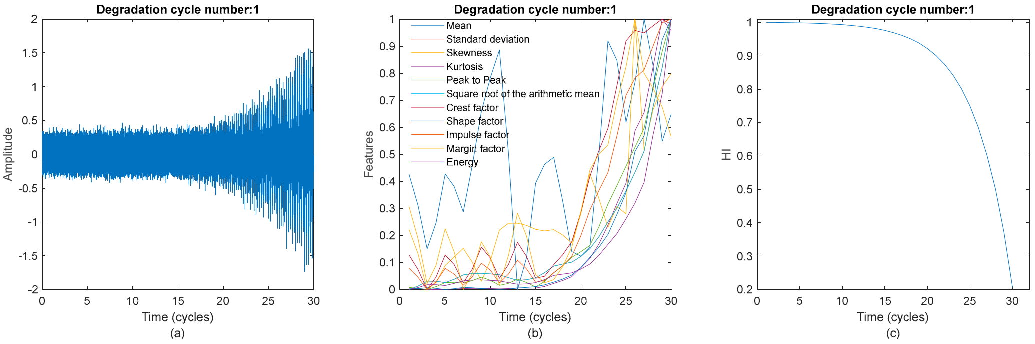

Figure 1. Dataset and preprocessing: (**a**) raw data of a single degradation life cycle; (**b**) prepared data for a single degradation life cycle; (**c**) health index of an entire life cycle.

3. Methods

In this work a long-short term memory (LSTM) network is involved in training process as it is recommended for such data drift and complexity problems (see Section 6.2 from [2]). In this context, for observation-based learning, a single layer LSTM with error-trial tuned parameters of 10 neurons, l_2 regularization parameter equal to 0.01, learning rate of 0.01 is used in this case. The same parameters are kept for sequence-based training while a sequence length was fixed to 6 observations. The only thing that changes in this case is the input and output layers sizes to fit changes in data mapping and sequence length. Figure 2a is and example of an ordinary deep network familiarly used within degradation process monitoring. While Figure 2b is the new network architecture reflecting sequence-based learning.

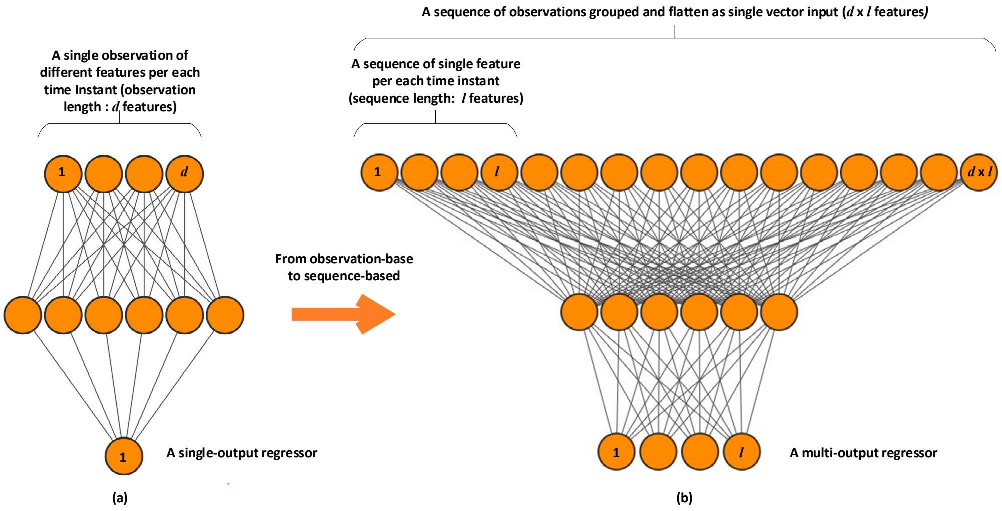
A sequence of observations grouped and flatten as single vector input (*d* x *l* features)
A single observation of different features per each time Instant (observation length : *d* features)
A sequence of single feature per each time instant (sequence length: *l* features)
1
d
1
l
d x *l*
From observation-base to sequence-based
1
A single-output regressor
1
l
A multi-output regressor
(a)
(b)

Figure 2. DL model architecture: (**a**) observation-based DL model; (**b**) sequence-based DL model.

4. Results and Discussion

In this work, from each previously mentioned speed profiles we selected 3 life cycles for training and 1 file for testing. Meaning that 30 degradation profiles are used for training and 10 degradation profiles are used for testing. This includes files of both subsets for both inner and outer race fault scenarios. Mini-batch size, maximum number of epochs, and iterations are fixed to 10, 1000, and 3000 respectively for both DL networks under same tuning mechanism of error-trial basis. Two types of metrics are used in this case to judge the accuracy of training process; visual and numeric. Visual metrics including curve fitting examples, loss function behavior, and some scoring functions behavior as will be illustrated in the following numerical metrics. Numerical metrics include the root mean squared error (RMSE) in Equation (1) and the a Score function in Equation (2) which usually used to evaluate data-driven models for bearing degradation analysis [11]. n,y and $\tilde{y}$ are number of samples, desired health index, and predicted health index respectively. The score function penalizes early and late predictions differently to satisfy some decision making constraints related to maintenance planning [12]. In the meantime, the RMSE designed to study the actual distance between prediction reflecting a real measurement meaning.

$$RMSE = \frac{1}{n}\sum_{i=1}^{n} y_i - \tilde{y}_i \tag{1}$$

$$Score = \begin{cases} e^{-\ln(0.5)\left(\frac{100(y-\tilde{y})}{5}\right)}, & \frac{100(y-\tilde{y})}{5} < 0 \\ e^{+\ln(0.5)\left(\frac{100(y-\tilde{y})}{5}\right)}, & \frac{100(y-\tilde{y})}{20} \geq 0 \end{cases} \tag{2}$$

Figure 3a is showcasing the loss function behavior. For observation-based learning we can observe a faster convergence and less loss values. However, the model sooner stacks in overfitting problem showcasing fluctuations in loss values (e.g., iteration 1000, 1500, 2500). In the meanwhile, despite late convergence and a bit bigger values of the loss function for sequence-based learning, the DL model shows better stability with no signs of overfitting. The curve fit examples of the test set for both inner race and out race profiles for speed profile of 100 time cycles in Figure 3b,c shows that sequence-based results are closer and smother leading to better predictions.

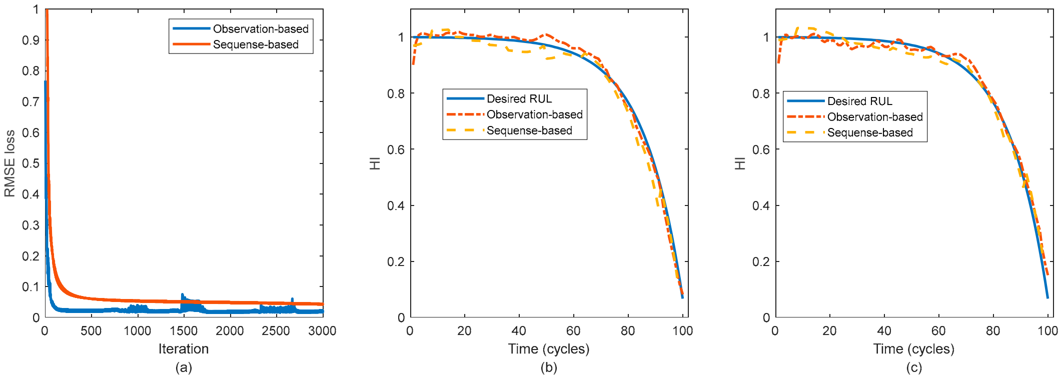

Figure 3. Obtained results: (**a**) loss function behavior; (**b**) curve-fit results example for inner race fault degradation cycles; (**c**) curve-fit results example for outer race fault degradation cycles.

Figure 4 is dedicated to address the behavior of suggested scoring function. The score function designed to explain both early predictions scores (percentage error > 0) and late predictions (percentage error < 0) scores. These predictions have relation with maintenance decisions. What we see in Figure 4 is observation-based prediction scores are further dispersed compared to sequence-based ones when approaching value 1. This means that the DL model in the latter has a better generalization (prediction on the test set). This proves necessity to sequential learning in improving data presentation consequently improving approximation and generalization through accurate tuning.

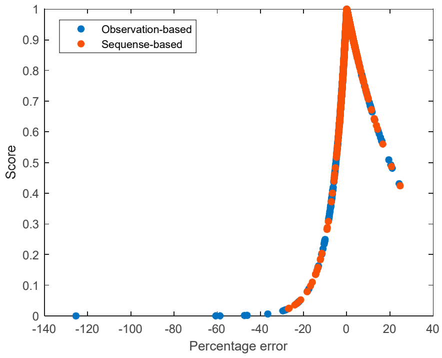

Figure 4. Score function behavior.

For numerical evaluation, the RMSE and Score results are showcased in Table 1. Results encourage using sequence-based mapping when dealing with such slowly evolving degradation process better then observation-based due to the clear gap between them in term of performances. Also, sequence-based methodology seems less computationally expensive than observation-based one, especially when computationally time results in Table 1 confirming such information.

Table 1. Final numerical evaluation results.

Method	RMSE	Score	Training Time (s)
Sequence-based LSTM	0.0243	0.8401	16.4235
Observation-based LSTM	0.0246	0.7989	30.9235

5. Conclusions

This work introduced an experiment of health index assessment using deep learning under slowly evolving degradation processes. It discussed the use of sequential learning philosophy versus traditional observation philosophy when training a DL model for approximating a degraded function. A bearing degradation dataset of multiple faults scenarios is used in these cases while adopting LSTM learning rules for DL model reconstruction. Many visual and numeric assessment metrics are used to evaluate performances of investigated approaches. Results encourage adopting sequence-based methodology as it allows mitigating mispresented observations resulted due to harsh operating conditions. As a perspective, further highly dynamic systems need to be studied for such problem including deeper architecture and targeting other problems of data complexity and availability and not only data drift problems. By doing so, further performances details about such methodology will be revealed.

Author Contributions: Conceptualization, T.B.; methodology, T.B. and M.B.; T.B. and M.B.; validation, T.B. and M.B.; formal analysis, T.B. and M.B.; investigation, T.B.; resources, T.B.; data curation, T.B. and M.B.; writing—original draft preparation, T.B.; writing—review and editing, T.B. and M.B.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data used in this work can be downloaded at https://figshare.com/articles/dataset/Simulated_Bearing_Degradation_Data_mat/12554690. Necessary files to reproduce the findings of this work can be downloaded at: https://doi.org/10.5281/zenodo.8142676.

Acknowledgments: Authors thank Koceila, A.; Mouchaweh, M.S.; Cornez, L.; Chiementin, X for making their simulated bearing degradation dataset public and easily accessible, enabling the results presented in this work to be produced.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Berghout, T.; Benbouzid, M.; Bentrcia, T.; Lim, W.H.; Amirat, Y. Federated Learning for Condition Monitoring of Industrial Processes: A Review on Fault Diagnosis Methods, Challenges, and Prospects. *Electronics* **2022**, *10*, 158. https://doi.org/10.3390/electronics12010158.
2. Berghout, T.; Benbouzid, M. A Systematic Guide for Predicting Remaining Useful Life with Machine Learning. *Electronics* **2022**, *11*, 1125. https://doi.org/10.3390/electronics11071125.
3. Khamassi, I.; Sayed-Mouchaweh, M.; Hammami, M.; Ghédira, K. Discussion and Review on Evolving Data Streams and Concept Drift Adapting. *Evol. Syst.* **2018**, *9*, 1–23. https://doi.org/10.1007/s12530-016-9168-2.
4. Berghout, T.; Benbouzid, M. PrognosEase: A Data Generator for Health Deterioration Prognosis. *SoftwareX* **2023**, *23*, 101461. https://doi.org/10.1016/j.softx.2023.101461.
5. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery Health Prognostics: A Systematic Review from Data Acquisition to RUL Prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834. https://doi.org/10.1016/j.ymssp.2017.11.016.
6. Berghout, T.; Benbouzid, M.; Amirat, Y. Towards Resilient and Secure Smart Grids against PMU Adversarial Attacks: A Deep Learning-Based Robust Data Engineering Approach. *Electronics* **2023**, *12*, 2554. https://doi.org/10.3390/electronics12122554.
7. Ding, C.; Zhao, J.; Sun, S. Concept Drift Adaptation for Time Series Anomaly Detection via Transformer. *Neural Process. Lett.* **2023**, *55*, 2081–2101. https://doi.org/10.1007/s11063-022-11015-0.
8. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *4*, 3104–3112.
9. Berghout, T. Sequence-Based Instead of Observation-Based Deep Learning: Monitoring Slowly Evolving Degradation Processes. 2023. https://doi.org/10.5281/zenodo.8142676.
10. Koceila, A.; Mouchaweh, M.S.; Cornez, L.; Chiementin, X. Simulated Bearing Degradation Data. 2020. https://doi.org/10.6084/m9.figshare.12554690.v2.
11. Berghout, T.; Mouss, L.-H.; Bentrcia, T.; Benbouzid, M. A Semi-Supervised Deep Transfer Learning Approach for Rolling-Element Bearing Remaining Useful Life Prediction. *IEEE Trans. Energy Convers.* **2022**, *37*, 1200–1210. https://doi.org/10.1109/TEC.2021.3116423.
12. Gouriveau, R.; Medjaher, K.; Ramasso, E.; Zerhouni, N. PHM–Prognostics and Health Management De La Surveillance Au Pronostic de Défaillances de Systèmes Complexes. *Tech. l'ingénieur Fonct. Strat. la Maint.* **2013**, *9*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.