Towards a Taxonomic Benchmarking Framework for Predictive Maintenance: The Case of NASA's Turbofan Degradation

Patrick Zschech
Technische Universität Dresden, patrick.zschech@tu-dresden.de

Jonas Bernien
Technische Universität Dresden, jonas.bernien@mailbox.tu-dresden.de

Kai Heinrich
Technische Universität Dresden, kai.heinrich@tu-dresden.de

Towards a Taxonomic Benchmarking Framework for Predictive Maintenance: The Case of NASA's Turbofan Degradation

Short Paper

Patrick Zschech, Jonas Bernien, Kai Heinrich
Technische Universität Dresden
Dresden, Germany
{patrick.zschech, jonas.bernien, kai.heinrich}@tu-dresden.de

Abstract

The availability of datasets for analytical solution development is a common bottleneck in data-driven predictive maintenance. Therefore, novel solutions are mostly based on synthetic benchmarking examples, such as NASA's C-MAPSS datasets, where researchers from various disciplines like artificial intelligence and statistics apply and test their methodical approaches. The majority of studies, however, only evaluate the overall solution against a final prediction score, where we argue that a more fine-grained consideration is required distinguishing between detailed method components to measure their particular impact along the prognostic development process. To address this issue, we first conduct a literature review resulting in more than one hundred studies using the C-MAPSS datasets. Subsequently, we apply a taxonomy approach to receive dimensions and characteristics that decompose complex analytical solutions into more manageable components. The result is a first draft of a systematic benchmarking framework as a more comparable basis for future development and evaluation purposes.

Keywords: Prognostics, Maintenance, Data Science, Evaluation Framework, Taxonomy, Review

Introduction

Given the evolving complexity of production systems and machinery, the maintenance function plays a crucial role in today's industrial value creation as it helps manufacturing companies to guarantee high reliability, human safety and low environmental risks (Muchiri et al. 2011; Peng et al. 2010). For this purpose, modern production environments increasingly focus on proactive maintenance strategies like predictive maintenance based on data-driven prognostic solutions to make efficient use of given resources and avoid redundant expenditures. The main goal of anticipatory approaches is to predict faults and failures before they occur and determine the remaining useful life (RUL) of technical assets by identifying relationships between extensive monitoring data and critical events (Bousdekis et al. 2018; Elattar et al. 2016). This situation is favored by the ubiquitous use of advanced information and communication technology that simplifies the collection of large and multifaceted data and the fast developments in big data and analytics methods that help to unveil valuable insights for better decision support (Zschech 2018).

In contrast to physical models and experience-based methods, data-driven approaches, that are based on techniques from disciplines like machine learning (ML) or statistics, have the advantages that i) they do not require comprehensive system knowledge, ii) they are fast to implement, iii) they can be tuned for similar systems, and iv) they are able to exploit hidden relations and nuances within the data records (Elattar et al. 2016; Peng et al. 2010). However, a prevailing disadvantage, and thus a common bottleneck of data-driven approaches, is the availability of representative data that reflect a system's symptomatic behavior from normal and faulty operations to degradation patterns under certain operating conditions.

Such run-to-failure data are often scarce in industrial practice and can only be procured at great expense due to zero-downtime policies (Dragomir et al. 2009; Zschech et al. 2019). For this reason, there have been several initiatives to generate synthetic datasets, covering a variety of technical settings, such as *milling machines*, *bearings*, *li-ion batteries* or *turbofan engines* (Eker et al. 2012). Derived from laboratory experiments or advanced simulations, such synthetic datasets show realistic properties and therefore provide a fundamental basis for the development and assessment of data-driven prognostic solutions (Eker et al. 2012; Ramasso and Saxena 2014). Accordingly, they are in use by a large scientific community, where researchers from various disciplines continuously compete against each other in order to propose novel solutions that consist of multiple processing components, while simultaneously striving to achieve the best benchmark result for a specific setting. The majority of studies, however, only consider their solution as a whole and evaluate the overall approach against a final prediction score to assess whether it performs better or worse than existing benchmarks. While a single metric proves to be a good choice for a quick comparison, we argue that a much more fine-grained consideration is required that distinguishes between detailed aspects of an overall solution along its analytical development process and measures the impact of particular method components, such as specific pre-processing or modelling steps.

Thus, our research goal is to *identify characteristic components of data-driven prognostic solutions* towards the establishment of a systematic benchmarking framework that reduces the complexity of prognostic solutions by decomposing them into their inherent parts. This should help to increase transparency and guide the selection of different components during the development of novel solutions and their embedding into analytical information systems, as it allows an evaluation on a more fine-grained basis. To carry out our research, we apply a taxonomy development approach guided by the method of Nickerson et al. (2013). Hereby, we demonstrate how a taxonomic approach can help to decompose complex analytical solutions into more manageable pieces. For the identification of a large corpus of studies, which serve as the foundation for the taxonomy development, we rely on a systematic literature review (vom Brocke et al. 2009). Particularly, in this research-in-progress paper, we start our research with a narrow focus on studies dealing with a specific maintenance scenario in order to keep the complexity manageable and obtain a benchmarking framework that applies to a distinct class of prognostic problems. Here, we choose the *turbofan degradation scenario* based on *NASA's C-MAPSS data* as a predominantly applied scenario within the prognostics community (Ramasso and Saxena 2014), whereas in further research, we will extend the scope to additional settings (e.g., Eker et al. 2012) in order to validate our results and possibly improve them towards higher generalizability. Following this line, the rest of this paper is organized as follows: First, we briefly describe the background of NASA's turbofan degradation scenario and refer to related work. We then describe our research method in detail and subsequently present the taxonomic framework, followed by a thorough discussion of the results. Finally, we draw a conclusion and give an outlook on how the work will be continued in subsequent steps.

Case Background and Related Work

The availability of suitable datasets, which researchers can use to develop and assess novel solutions, is a common bottleneck in data-driven prognostics due to the fact that critical assets are most often not allowed to fail and thus are replaced well in advance without recording relevant events (Ramasso and Saxena 2014; Zschech et al. 2019). Therefore, synthetic benchmarking datasets generated in different experimental settings, such as *milling machines*, *bearings*, *li-ion batteries* or *turbofan engines* (Eker et al. 2012), play a crucial role within the field. Among these examples, the turbofan scenario based on NASA's C-MAPSS data (commercial modular aero-propulsion system simulation) is one of the most frequently applied scenarios for prognostic solution development (Elattar et al. 2016; Zschech et al. 2019). The data collection encompasses five similar datasets generated by a simulation environment from the NASA Ames Research Center that replicates the behavior of turbofan engines under a variety of operating conditions and a continuous degradation due to varying fault injection parameters. As such, each dataset is composed of multivariate time series containing parameters and sensor measurements of operating cycles from different turbofan engines (Saxena, Goebel, et al. 2008). The data were used for the first time as part of a prognostic challenge at the *PHM'08* conference. Thereafter, another four datasets with varying degrees of complexity were released *(FD001-FD004)*. Due to their realistic characteristics in terms of i) multi-dimensional sensor measurements from a complex non-linear system, ii) high level noise to represent natural variability and iii) masked fault effects through different operating conditions, the datasets were used by numerous researchers for diagnostic and prognostic development purposes (Ramasso and

Saxena 2014). To draw conclusions from these efforts, Ramasso and Saxena (2014) conducted a first review, in which they compared and classified several aspects of existing solutions identified in forty unique studies. Meanwhile, the number of studies has more than doubled and developments are even steadily increasing with a push of new achievements, especially from the field of ML and deep learning.

These developments shall be given a new orientation by proposing a systematic, taxonomic benchmarking framework. In general, taxonomies serve as a viable research approach for organizing knowledge in a structured manner and manifesting descriptive theories (Gregor 2006). For this purpose, they enable researchers to study the relationship among concepts and help to analyze and understand complex domains (Nickerson et al. 2013). Zschech (2018), for example, already applied a taxonomy development approach within the field of data-driven maintenance in order to decompose dimensions and characteristics of recurring data analysis problems. The author discusses the merits of this methodical approach as a systematic procedure to decompose complexity of analytical solutions. The focus of this paper, however, is on a more detailed sublevel, as it exclusively concentrates on prognostic method components along the overall analytical development process. This is exemplified on the C-MAPSS datasets as a representative foundation for prognostic solution development in order to receive taxonomic benchmarking dimensions and characteristics from a manageable amount of prognostic studies.

Research Method

To identify relevant studies that dealt with the development of prognostic solutions using C-MAPSS datasets, we carried out a systematic literature review (vom Brocke et al. 2009). Specifically, we applied a database search using the following libraries: *AIS Electronic Library, EBSCOhost, IEEE Xplore, ScienceDirect* and *SpringerLink*. As search terms, we combined the keywords 'C-MAPSS' and 'dataset' and applied them with alternative spellings (e.g., 'CMAPSS') and synonyms (e.g., 'PHM 2008', 'NASA turbofan degradation'), leading to 128 unique items. Additionally, we performed a forward search based on the C-MAPSS introduction provided by Saxena, Goebel, et al. (2008) (+52 items), searched the websites of the PHM Society and the NASA Prognostics Center of Excellence (+40 items), and performed a backward search based on the review conducted by Ramasso and Saxena (2014) (+7 items). Thus, it was possible to obtain 227 unique hits (day of search: 2018-09-24), which had to be further reduced by appropriate filter criteria. Here, we ensured that the studies i) were written in English/German (-1 item), ii) were based on one of the five C-MAPSS datasets (-68 items), iii) dealt with a prognostic approach as opposed to a diagnostic orientation (-30 items), iv) applied a data-driven approach (-4 items), and v) proposed a previously unknown solution (-18 items). This resulted in 106 relevant studies, which were then used for the subsequent step of the taxonomy development (cf. Appendix, Table 2).

To carry out the taxonomy development, we applied the method proposed by Nickerson et al. (2013) as it provides systematic guidance. It basically consists of three steps: i) determining a meta-characteristic, ii) specifying ending conditions, and iii) identifying dimensions and characteristics of the taxonomy. The meta-characteristic is the root element, as it serves as a foundation for the choice of all other characteristics. As such, it was defined in accordance with our research goal, i.e., to identify *characteristic components of data-driven prognostic solutions*. The specification of ending conditions, on the other hand, is required due to the iterative method character. For this purpose, Nickerson et al. (2013) define certain subjective criteria that must be fulfilled, e.g., that a taxonomy is sufficiently robust to contain enough dimensions and characteristics to separate between the objects of interest, while it is sufficiently concise to not exceed the cognitive load of the taxonomy user. Moreover, the method requires the specification of objective ending conditions, e.g., that every characteristic within its dimension is unique and not repeated. At this point, we adopted the following four criteria for our approach: *i) all objects have been examined, ii) at least one object can be assigned for each characteristic across all dimensions, iii) no new dimensions or characteristics were added in the last iteration*, and *iv) no dimensions or characteristics were modified in the last iteration*. The actual step of identifying dimensions and characteristics can then be carried out either with an *empirical-to-conceptual (E2C)* or a *conceptual-to-empirical (C2E)* path. We applied a combination of both paths by running several iterations until all ending conditions were met.

In a **first iteration**, all solutions were differentiated according to the underlying properties of the respectively applied C-MAPSS datasets. Here, it was possible to distinguish between different levels of complexity among all five datasets, mainly expressed by the two dimensions '*simultaneous fault modes'* and '*number of operational conditions'*. While some solutions were exclusively evaluated on a particular dataset

with its inherent degree of complexity, others were also tested on multiple instances with their respective properties. In some studies, however, it was not indicated at all which datasets or complexity levels were considered. In the **next four iterations**, we identified a broad range of pre-processing methods, which could be organized within the following four dimensions: *normalization, noise reduction, dimensionality reduction*, and *feature selection*. For the first three dimensions, we applied E2C-paths classifying empirically derived methods for each step. For the 'feature selection' dimension, on the other hand, we applied the categorization suggested by Wang (2010) distinguishing between filter and wrapper methods, which could also be confirmed empirically. Additionally, we had to add a third category covering all approaches where the selection step was carried out manually. Moreover, for several studies, a precise classification was not possible, since no exact information was given on the details of the pre-processing steps. In the **sixth iteration**, we considered all *prognostic modelling approaches* starting with an E2C-path. At this point, we were faced with some major abstraction difficulties when trying to organize a variety of heterogenous modelling approaches from the fields of statistics, mathematical modelling and artificial intelligence into a flat-dimensioned and non-hierarchical representation. To this end, we proceeded with an C2E-path applying a categorization by Ramasso and Saxena (2014) and additionally listed each individual modelling approach separately to avoid an inflated, imbalanced and partially overlapping set of modelling characteristics within the taxonomy. This categorization encompasses three groups of generic approaches, namely direct, indirect, and similarity-based modelling, which could all be confirmed empirically by the approaches at hand. In a **seventh iteration**, we extracted a series of *performance metrics* for prognostic model assessment. However, due to a high number of individual measures, we applied a categorization proposed by Saxena, Celaya, et al. (2008) that distinguishes between metrics based on accuracy, robustness, precision, trajectory, and prognostic-specific properties. Of these five groups, accuracy, precision and prognostics-specific metrics could be confirmed empirically and thus they were kept within the taxonomy. In a **last iteration**, all studies were screened again and since no more modifications occurred, all ending conditions were met to complete the taxonomy development process.

Results

In this section, we present our results and briefly describe the final taxonomy (cf. Table 1). Additionally, we list all 106 examined studies with their respective characteristics in Table 2 of the Appendix, where we also report the specific prognostic approaches for the modelling step as well as the datasets used for solution development. Furthermore, to view the full source reference for each study, please refer to the following link: https://www.researchgate.net/publication/335611604

Dimension	Characteristics		
Fault Modes	Single Fault Mode	Multiple Fault Modes	
Operational Conditions	Single Condition	Multiple Conditions	
Normalization	Standardization	Rescaling	
Noise Reduction	Moving Average	Exponential Smoothing	Polynomial Smoothing
Feature Selection	Manual Selection	Filter	Wrapper
Dimensionality Reduction	Hierarchical	Non- Hierarchical	
Prognostic Approach	Direct RUL-Mapping	Indirect RUL-Mapping via HI	Similarity-based Matching
Performance Metric	Accuracy-based	Precision-based	Prognostic-specific Metric

Table 1. Taxonomy for Data-Driven Prognostic Solutions Based on C-MAPSS Studies

Fault Modes: As described previously, the five C-MAPSS datasets used for solution development show different levels of complexity, where the first dimension refers to the degree of simultaneous fault modes. As such, the datasets *FD003* and *FD004* are subject to *multiple fault modes*, while the remaining datasets *FD001*, *FD002* and *PHM08* only contain a *single fault mode* and therefore show a lower level of complexity. However, since the majority of the studies are based on *FD001* and *PHM08*, the single fault setting has played so far a much larger role in the development of prognostic solutions within the turbofan scenario (97 studies), whereas studies based on multiple faults are severely underrepresented (22 studies).

Operational Conditions: The second dimension describing dataset complexity is given by different numbers of operating conditions that mask the general effect of faults. Here, it can be distinguished between a *single operational condition*, as given in the datasets *FD001* and *FD003,* and *multiple conditions* representing a higher complexity, which can be found in the datasets *FD002*, *FD004* and *PHM08*. In this dimension, both characteristics are almost equally considered among all studies (single: 60/multiple: 57).

Normalization: The next dimension refers to a pre-processing step called normalization, in which measured values of different scales are adjusted to a notionally common scale. It could basically be distinguished between a *standardization* approach (i.e., values are transformed via mean and standard deviation, also called z-score) and a *rescaling* approach (i.e., values are adjusted to [-1,1] or [0,1] intervals primarily based on a min-max transformation). Normalization is only mentioned by 32 studies, with standardization being the preferred approach in comparison to rescaling (20/12).

Noise reduction: This pre-processing step is used to reduce noise from a signal. Here, it was possible to differentiate between *moving average* (i.e., mean calculation on rolling subsets), *exponential smoothing* (i.e., usage of exponential function to assign decreasing weights over time), and *polynomial smoothing* (i.e., usage of polynomial function). Overall, only 15 studies explicitly mention a noise reduction approach, with all three groups being almost equally represented within the studies (6/6/5).

Feature selection: Methods in this dimension support the process of identifying relevant features from uninteresting/redundant ones, which might affect the performance of learning algorithms. Based on the results, it could be differentiated among wrapper, filter, and manual selection methods. *Wrapper methods* utilize a learning algorithm as a black box to score subsets of features according to their predictive power. *Filter methods*, on the other hand, use local performance criteria such as correlation or mutual information criteria for feature subset selection independently of a chosen predictor (Wang 2010). Further on, *manual selection* takes place, when no automatic procedure is used, but instead features are selected, for example, based on visual investigations. Manual selection is by far the most widely mentioned approach within the studies in comparison to wrapper and filter methods (44/18/10).

Dimensionality reduction: Similar to feature selection, dimensionality reduction aims at preserving discriminatory information while shrinking the extent of features. However, the idea behind this step is the transformation of features into a more compact representation. The most frequently used approach, which is mentioned in nine studies, is based on principal component analysis (PCA), where a set of features is projected onto a set of linearly uncorrelated features via an orthogonal transformation. A few studies also consider several PCA variants such as kernel PCA (e.g., Lim et al. 2016) or greedy kernel PCA (e.g., Chen et al. 2016). Another approach is based on a hierarchical procedure, in which the feature reduction is carried out over several levels so that a full correlation/covariance matrix is not required (Krishnan et al. 2017). Against this background, the taxonomy distinguishes between *hierarchical* and *non-hierarchical* approaches, with the latter category being more frequently used (1/13).

Prognostic approach: For the modelling step, the categorization by Ramasso and Saxena (2014) distinguishes between three groups of prognostic approaches, i.e., direct, indirect and similarity-based approaches. In the *direct approach*, a functional mapping between a multidimensional feature space and the RUL is established. Overall, 36 studies apply this group, where the majority of studies are based on different types of neural networks with an increasing tendency towards deeper architectures, such as deep belief networks (e.g., Zhang et al. 2018), convolutional neural networks (CNN) (e.g., X. Li et al. 2018) or long short-term memory networks (LSTM) as a specific variant of recurrent networks (e.g., Y. Wu et al. 2018). The *indirect approach*, on the other hand, requires two mapping functions. The first one maps input data to a health index (HI) and the second one maps HI values to the RUL. In this way, a library of degradation models is constructed that serves as prior knowledge for model estimation. The generation of the HI can be based, for example, on hidden Markov models (e.g., Juesas and Ramasso 2016) or stochastic process models, such as Wiener or Gamma processes (e.g., N. Li et al. 2018). With 54 studies, the indirect mapping is most frequently pursued, whereas the last group of *similarity-based matching* is only employed by 19 studies. Here, a library of trajectories is created based on historical instances labeled with known failure times. The RUL of new instances is then estimated by evaluating their similarity with instances in the library. To assess similarity, different approaches can be used, such as Euclidean distance (e.g., Wang et al. 2008) or polygon coverage based on computational geometry (e.g., Ramasso 2014).

Performance metric: The last dimension is divided in different types of performance metrics using the categorization by Saxena, Celaya, et al. (2008): *Accuracy-based metrics* compute statistics over the error between the actual failure time and the predicted value. This group is predominantly used (93 studies) and includes metrics like root mean squared error or mean absolute percentage error. *Precision-based metrics*, on the other hand, measure the dispersion of the prediction error, such as mean absolute deviation, which is only applied by six studies. The third group encompasses *prognostics-specific metrics*, addressing domain-specific peculiarities, which are also rarely applied (11 studies).

Discussion

Reflecting the results of this research, we contribute to the field of data-driven solution development for predictive maintenance and analytical information systems in several ways: **First**, we extended the review of Ramasso and Saxena (2014) by a more recent scope with an exclusive focus on prognostic approaches. This helps the community to get an overview of current trends, where it becomes apparent, for example, that although the largest part of prognostic modelling steps is taken up by indirect RUL-mappings, the number of direct mappings has grown enormously, where recent efforts are increasingly directed towards the application of deep neural networks, such as LSTM or CNN. In subsequent research, it is also planned to reveal further descriptive insights to the reviewed studies, which was not part of the current investigation so far. **Second**, by applying a taxonomic approach, it was indeed possible to decompose the complexity of analytical solutions to a certain degree, which is necessary for future efforts to compare novel solutions on a more fine-grained basis. By employing the method of Nickerson et al. (2013), a number of useful dimensions and characteristics could be extracted, including *dataset characteristics*, various *pre-processing components*, a high-level distinction of *prognostic modelling approaches* and various groups of *performance metrics*. Thus, the taxonomy delivers an overview about different design options (including frequencies of different methods and method combinations), provides structured access to knowledge in terms of a method catalog and provides a setting to position individual configurations of novel solutions on a more comparable basis. **Third**, as pursued by the research goal, the taxonomy provides the foundation for a systematic framework, which is intended to be used in subsequent research for benchmarking purposes. Here, it is planned to reconstruct existing solutions and iteratively modify the characteristics of the dimensions to measure and evaluate their impact on the overall performance. On this basis, more stable statements can be made as to whether, for example, the performance of new proposed solutions primarily depends on specific modelling steps or whether particular pre-processing steps contribute to their impact. This can be of particular interest, for example, against the background of the aforementioned deep neural networks, which claim to be capable of automatically extracting relevant features without the need for corresponding feature engineering (LeCun et al. 2015).

On the downside, our approach has also some limitations: **First**, the taxonomy was developed in such a way that only characteristics explicitly mentioned within the studies were extracted. However, it can be assumed that far more studies show corresponding characteristics in several dimensions, especially in those for pre-processing, since, as in the case of normalization for example, certain prognostic models like neural networks require normalized input values for feasible calculations within their activation functions. Therefore, no distinction was made as to whether certain characteristics were simply not discussed or whether they have indeed a 'blank option'. Thus, for the benchmarking framework, it is planned to introduce such blank options whenever they are applicable. **Second**, the specific order of pre-processing steps has not yet been taken into account, which of course can have an impact. This is also subject of further research. **Third**, during the extraction of several method components, we were faced with heterogenous wordings and concept names. For example, some authors apply the term 'feature extraction' in the sense of dimensionality reduction (e.g., Lim et al. 2016), while others use it for feature selection (e.g., Elattar et al. 2018). Another example is the term 'filter', which is either used for noise reduction (e.g., Ramasso and Saxena 2014), feature selection (e.g., Wang 2010) or specific modelling approaches (e.g., Peel 2008). As such, the results possibly suffer some ambiguous categorization, especially when highly specific approaches were applied. In further research, this issue should be addressed by taking a more generalized view on method categorizations beyond the scope of maintenance studies to avoid imprecise classifications. **Fourth**, for some dimensions, a deeper partitioning was conceivable, such as for the performance metrics, where each metric could form a single characteristic. However, in the course of the development method from Nickerson et al. (2013), a balance had to be struck between *robustness* on one hand and *conciseness/parsimony* on the other hand to obtain a suitable level of abstraction. This also applies to the evaluation framework, where it is not necessarily desired to examine every slight change of characteristics. Thus, while an acceptable level of abstraction could currently be found for the pre-processing dimensions, this proved to be more difficult for the modelling dimension. Facing heterogenous approaches like instance-based learning, deep learning, stochastic process models or optimization and simulation approaches, it was difficult to abstract their inherent characteristics (e.g., model assumptions, learning/approximation concepts, target functions to be optimized, treatment of uncertainty, etc.) in a reasonable way so that they could be transferred into a flat taxonomy structure. This issue was also noted by Zschech (2018) during the extraction of taxonomic characteristics for a wide range of analytical tech-

niques. Consequently, it seems more reasonable to allow hierarchical, tree-like categorizations or to create sub-taxonomies for each class of prognostic modelling approaches. Here, we already reported several candidates that appear to be useful sub-dimensions in subsequent research (e.g., division of direct modelling approaches into different sub-types of deep neural networks, such as deep belief, CNN and recurrent networks). **Fifth**, to take this research to a more generic level, there is currently a certain bias, as the extracted characteristics depend on the C-MAPSS scenario. This limitation was useful, as it allowed to concentrate on an extensive but manageable amount of studies representing a scenario with highly realistic properties. However, a preliminary screening of several C-MAPSS studies, in which prognostic solutions are also applied to other scenarios (e.g., Krishnan et al. 2017; Malhotra et al. 2016; Mosallam et al. 2015; Xi et al. 2018), reveals that the extracted taxonomy framework is basically applicable for broader contexts. Thus, it might only require smaller adjustments towards higher generalizability, such as the consideration of further data characteristics or pre-processing steps (e.g., treatment of missing values or relaxation effects as in the case of the *li-ion battery scenario*, cf. Xi et al. 2018). Following this line, we will extend our research in subsequent work to additional maintenance settings to cover a broader variability and validate our results. Similarly, it is planned to apply the overall methodical approach to further contexts in completely different domains (e.g., predicting behavior of business processes or visual object recognition in enterprise applications) to demonstrate our approach's general feasibility and expediency, as we are convinced that such a methodical application is not necessarily limited to maintenance scenarios.

Conclusion and Outlook

In this paper, we developed a first draft of a benchmarking framework for data-driven prognostic solutions. To this end, we applied a taxonomy approach to reach better transparency and decompose complex solutions into more manageable pieces. In subsequent steps, it is planned to refine our results by addressing the discussed limitations - especially towards extending the investigation to other maintenance settings and related datasets as well as extracting further aspects for the three prognostic modelling classes in which more specific subdimensions are conceivable. After the refinement, the results will be used for the implementation of the benchmarking framework, where the extracted dimensions and characteristics serve as evaluation options to be iteratively modified under ceteris paribus conditions. Thus, by using dynamic programming pipelines, a varying combination of pre-processing and modelling approaches (e.g., *no noise reduction vs. exponential smoothing + wrapper vs. filter + direct LSTM vs. indirect LSTM vs. indirect HMM*) will be studied based on different dataset characteristics with regard to their impact on selected performance metrics. In this way, it is planned to establish a better and unbiased understanding to what extent certain method components affect the results of prognostic solutions, which can guide the development and the application of such data-driven approaches in similar maintenance environments.

References

Bousdekis, A., Magoutas, B., Apostolou, D., and Mentzas, G. 2018. "Review, Analysis and Synthesis of Prognostic-Based Decision Support Methods for Condition Based Maintenance," *Journal of Intelligent Manufacturing* (29:6), pp. 1303–1316.

vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., and Cleven, A. 2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," in *European Conference on Information Systems*, Verona, Italy.

Chen, G., Chen, J., and Zi, Y. 2016. "A GKPCA-NHSMM Based Methodology for Accurate RUL Prognostics of Nonlinear Mechanical System with Multistate Deterioration," in *Prognostics and System Health Management Conference*, Chengdu, China.

Dragomir, O. E., Gouriveau, R., Dragomir, F., Minca, E., and Zerhouni, N. 2009. "Review of Prognostic Problem in Condition-Based Maintenance," in *European Control Conference*, pp. 1585–1592.

Eker, O. F., Camci, F., and Jennions, I. K. 2012. "Major Challenges in Prognostics: Study on Benchmarking Prognostics Datasets," in *European Conference of the Prognostics and Health Management Society*, Dresden, Germany, pp. 148–155.

Elattar, H. M., Elminir, H. K., and Riad, A. M. 2016. "Prognostics: A Literature Review," *Complex & Intelligent Systems* (2:2), pp. 125–154.

Elattar, H. M., Elminir, H. K., and Riad, A. M. 2018. "Conception and Implementation of a Data-Driven Prognostics Algorithm for Safety–Critical Systems," *Soft Computing*, pp. 1–18.

Gregor, S. 2006. "The Nature of Theory in Information Systems," *MIS Quarterly* (30:3), pp. 611–642.
Juesas, P., and Ramasso, E. 2016. "Ascertainment-Adjusted Parameter Estimation Approach to Improve Robustness against Misspecification of Health Monitoring Methods," *Mechanical Systems and Signal Processing* (81), pp. 387–401.
Krishnan, R., Jagannathan, S., and Samaranayake, V. A. 2017. "Deep Learning Inspired Prognostics Scheme for Applications Generating Big Data," in *International Joint Conference on Neural Networks*, pp. 3296–3302.
LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521), pp. 436–444.
Li, N., Lei, Y., Yan, T., Li, N., and Han, T. 2018. "A Wiener Process Model-Based Method for Remaining Useful Life Prediction Considering Unit-to-Unit Variability," *IEEE Transactions on Industrial Electronics* (66:3), pp. 2092–2101.
Li, X., Ding, Q., and Sun, J.-Q. 2018. "Remaining Useful Life Estimation in Prognostics Using Deep Convolution Neural Networks," *Reliability Engineering & System Safety* (172), pp. 1–11.
Lim, P., Goh, C. K., and Tan, K. C. 2016. "A Time Window Neural Network Based Framework for Remaining Useful Life Estimation," in *International Joint Conference on Neural Networks*, pp. 1746–1753.
Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., and Shroff, G. 2016. *Multi-Sensor Prognostics Using an Unsupervised Health Index Based on LSTM Encoder-Decoder*.
Mosallam, A., Medjaher, K., and Zerhouni, N. 2015. "Component Based Data-Driven Prognostics for Complex Systems: Methodology and Applications," in *International Conference on Reliability Systems Engineering*, pp. 1–7.
Muchiri, P., Pintelon, L., Gelders, L., and Martin, H. 2011. "Development of Maintenance Function Performance Measurement Framework and Indicators," *International Journal of Production Economics* (131:1), pp. 295–302.
Nickerson, R. C., Varshney, U., and Muntermann, J. 2013. "A Method for Taxonomy Development and Its Application in Information Systems," *European Journal of Information Systems* (22:3), pp. 336-359.
Peel, L. 2008. "Data Driven Prognostics Using a Kalman Filter Ensemble of Neural Network Models," in *International Conference on Prognostics and Health Management*, pp. 1–6.
Peng, Y., Dong, M., and Zuo, M. J. 2010. "Current Status of Machine Prognostics in Condition-Based Maintenance: A Review," *The International Journal of Advanced Manufacturing Technology* (50:1–4), pp. 297–313.
Ramasso, E. 2014. "Investigating Computational Geometry for Failure Prognostics," *International Journal of Prognostics and Health Management* (5:1), pp. 1–18.
Ramasso, E., and Saxena, A. 2014. "Performance Benchmarking and Analysis of Prognostic Methods for CMAPSS Datasets," *International Journal of Prognostics and Health Management* (5:2), pp. 1–15.
Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., and Schwabacher, M. 2008. "Metrics for Evaluating Performance of Prognostic Techniques," in *International Conference on Prognostics and Health Management*, Denver, USA, pp. 1–17.
Saxena, A., Goebel, K., Simon, D., and Eklund, N. 2008. "Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation," in *International Conference on Prognostics and Health Management*, Denver, USA, pp. 1–9.
Wang, T. 2010. "Trajectory Similarity Based Prediction for Remaining Useful Life Estimation," Doctoral Dissertation, University of Cincinnati.
Wang, T., Yu, J., Siegel, D., and Lee, J. 2008. "A Similarity-Based Prognostics Approach for Remaining Useful Life Estimation of Engineered Systems," in *International Conference on Prognostics and Health Management*, pp. 1–6.
Wu, Y., Yuan, M., Dong, S., Lin, L., and Liu, Y. 2018. "Remaining Useful Life Estimation of Engineered Systems Using Vanilla LSTM Neural Networks," *Neurocomputing* (275), pp. 167–179.
Xi, X., Chen, M., Zhang, H., and Zhou, D. 2018. "An Improved Non-Markovian Degradation Model with Long-Term Dependency and Item-to-Item Uncertainty," *Mechanical Systems and Signal Processing* (105), pp. 467–480.
Zhang, J., Wang, P., Yan, R., and Gao, R. X. 2018. "Deep Learning for Improved System Remaining Life Prediction," *Procedia CIRP* (72), CIRP Conference on Manufacturing Systems, pp. 1033–1038.
Zschech, P. 2018. "A Taxonomy of Recurring Data Analysis Problems in Maintenance Analytics," in *Proceedings of the European Conference on Information Systems*, Portsmouth, UK.
Zschech, P., Heinrich, K., Bink, R., and Neufeld, J. S. 2019. "Prognostic Model Development with Missing Labels: A Condition-Based Maintenance Approach Using Machine Learning," *Business & Information Systems Engineering* (61:3), pp. 327–343.

Appendix

		Dataset					Fault Modes		Operat. Condit.		Normali-zation		Noise Reduction			Feature Selection			Dim. Reduct.		Prognostic Approach			Performance Metric		
Reviewed Studies	**Prognostic Modelling Approach**	FD001	FD002	FD003	FD004	PHM08	Single	Multiple	Single	Multiple	Standard.	Rescaling	Mov. Avg.	Exp. Smo.	Poly. Smo.	Manual	Filter	Wrapper	Hierarch.	Non-Hier.	Direct	Indirect	Similarity	Accuracy	Precision	Prognostic
[1] Heimes (2008)	Recurrent Neural Network (RNN)					X	X			X											X			X		
[2] Peel (2008)	Kalman Filter (KF), Ensemble of Neural Network Models					X	X			X	X										X			X		
[3] Wang et al. (2008)	Damage Estimation with Linear Regression (LR), Similarity-based Matching					X	X			X						X							X	X		
[4] Peysson et al. (2009)	Support Vector Regression (SVR)					X	X			X						X					X			X		
[5] Siegel (2009)	Health Indicator (HI), Logistic Regression	X					X		X							X						X		X		
[6] Coble (2010)	Weibull Fitting (WF), Markov Chain, General Path Model (GPM)					X	X			X							X					X		X		
[7] Gouriveau & Ramasso (2010)	Takagi-Sugeno Neuro-Fuzzy System (TSNFS), Evidential Hidden Markov Model (EHMM)					X	X			X							X					X		X		
[8] Jianzhong et al. (2010)	Ensemble of Multi-Layer Perceptron (MLP) Networks															X					X			X		
[9] Riad et al. (2010)	MLP					X	X			X	X		X			X					X			X		
[10] Wang (2010)	HI-based Similarity Matching, Kernel Smoothing				X			X		X	X					X	X			X			X			X
[11] Zemouri & Gouriveau (2010)	Recurrent Radial Basis Function Network, Autoregressive Model (ARM)					X	X			X											X			X	X	
[12] Coble & Hines (2011)	GPM with Dynamic Bayesian Updating					X	X			X							X					X		X		
[13] El-Koujok et al. (2011)	Takagi-Sugeno NFS					X	X			X												X		X		
[14] Gauvain et al. (2011)	Connexionist-based Multi-Step Prediction					X	X			X												X		X		
[15] Giantomassi et al. (2011)	MLP, HMM					X	X			X												X		X		
[16] Hu et al. (2012)	Ensemble of SVM/Relevance VM (RVM)/Exponential Fitting (EF)/Quadratic Fitting (QF)/RNN					X	X			X	X					X					X	X	X	X		X
[17] Le Son et al. (2012)	Non-homogeneous Gamma Process with Gaussian Noise Model, Gibbs Sampling (GS)					X	X			X												X		X		
[18] Peng, Wang, et al. (2012)	Echo State Network (ESN), KF	X					X		X												X			X		
[19] Peng, Xu, et al. (2012)	Instance-Based Learning (IBL)		X				X			X						X	X						X			X
[20] Sun et al. (2012)	State Space Model (SSM), Sequential Monte Carlo (SMC) Method															X						X		X		
[21] Wang et al. (2012)	Sparse Bayes Learning with RVM					X	X			X						X						X		X		
[22] Ishibashi & Júnior (2013)	Genetic Fuzzy Rule-based System, Decision Tree (DT)					X	X			X		X			X	X		X			X					
[23] Le Son et al. (2013)	Wiener Process Model (WPM)					X	X			X						X				X		X		X		
[24] Li et al. (2013)	Mixture of Gaussian HMM, Fixed Size Least Squares SVR	X	X	X	X		X	X	X	X		X										X		X			X
[25] Louen et al. (2013)	SVM, WF	X					X		X													X		X		
[26] Ramasso et al. (2013)	Case-based Reasoning, K-Nearest Neighbours (KNN), Belief Functions	X					X		X							X							X	X		
[27] Bluvband et al. (2014)	SVM, SVR	X					X		X					X	X						X			X		
[28] Fagogenis et al. (2014)	AR Locally Weighted Projection Regression, Random Undersampling Boosting Classifier	X					X		X												X					
[29] Khelif et al. (2014)	IBL	X					X		X														X	X		
[30] Liu et al. (2014)	Superstatistics Theory, Information Fusion, SSM, KF																					X		X		
[31] Mosallam et al. (2014)	Discrete Bayesian Filter (DBF), KNN, Gaussian Process Regression (GPR)	X					X		X								X			X		X		X		
[32] Ramasso & Denoeux (2014)	Evidential Expectation-Maximization (EM), Partially HMM (PHMM)	X					X		X													X		X		
[33] Ramasso & Gouriveau (2014)	TSNFS, Evidential HMM					X	X			X							X					X		X		
[34] Ramasso (2014)	Imprecise Health Indicator, Similarity-based Reasoning, Polygon Clipping	X	X	X	X	X	X	X	X	X						X		X					X	X		
[35] Wang & Gao (2014)	Regularized Particle Filtering (PF)																					X		X		
[36] Xi et al. (2014)	Copula-based Sampling Method					X	X			X						X						X		X		
[37] Xu et al. (2014)	Comentropy-based Fusion Model, Dempster-Shafer Regression (DSR), SVM, RNN	X					X		X							X						X		X		
[38] García Nieto et al. (2015)	SVM, Particle Swarm Optimization (PSO)	X					X		X							X						X				
[39] Javed et al. (2015)	Extreme Learning Machine (ELM), Subtractive Maximum Entropy Fuzzy Clustering	X					X		X							X		X				X		X		
[40] Lasheras et al. (2015)	Regression Trees, Multivariate Adaptive Regression Splines (MARS)	X					X		X								X			X		X		X		
[41] Le et al. (2015)	Multibranch Hidden Semi-Markov Model (HSMM)					X	X			X												X		X		
[42] Malinowski et al. (2015)	RUL Shapelet Extraction	X			X		X	X	X	X					X	X							X	X		X
[43] Mosallam et al. (2015)	DBF, GPR	X					X		X								X			X		X		X		
[44] Peng et al. (2015)	IBL	X					X		X							X		X					X			X
[45] Xinxin et al. (2015)	Exponential Damage Modelling		X				X			X						X							X			
[46] Al-Dahidi et al. (2016)	Homogeneous Discrete-Time Finite-State SMM, WF, Monte Carlo Simulation (MCS)		X				X			X	X			X		X						X		X	X	X
[47] Babu, Li, et al. (2016)	Meta-Cognitive Regression Neural Network (MCRNN)					X	X			X	X										X			X		
[48] Babu, Zhao, et al. (2016)	Deep Convolutional Neural Network (CNN)	X	X	X	X	X	X	X	X	X	X										X			X		
[49] Chen et al. (2016)	Non-homogeneous HSMM (NHSMM)	X					X		X											X		X		X		
[50] Juesas & Ramasso (2016)	Discrete HMM, EM with Weighted Distribution Theory (WDT)	X	X	X	X		X	X	X	X												X		X		
[51] Juesas et al. (2016)	Autoregressive PHMM, WDT	X					X		X													X		X		
[52] Le Son et al. (2016)	Non-homogeneous Gamma Process with Gaussian Noise Model, GS, Stochastic EM					X	X			X												X		X		
[53] Li et al. (2016)	Discrete-Time Markov Chain					X	X			X						X		X				X		X		
[54] Lim et al. (2016)	Time-Window Neural Network	X	X	X	X		X	X	X	X		X						X		X	X			X		
[55] Malhotra et al. (2016)	Long Short Term Memory based Encoder-Decoder (LSTM-ED), Curve Matching	X					X		X		X									X			X	X		
[56] Porotsky & Bluvband (2016)	SVM, SVR	X					X		X					X	X						X			X		
[57] Ramasso (2016)	EHMM, EM	X					X		X													X				
[58] Shi et al. (2016)	Hierarchical Failure State Clustering, RVM	X					X		X			X				X							X	X		
[59] Tao et al. (2016)	Dynamic Neural Network (DNN)	X					X		X			X				X						X		X		
[60] Xin et al. (2016)	Similarity-based Method based on Kernel Density Estimation					X	X			X	X									X			X			
[61] Xinxin et al. (2016)	Average and Comentropy Fusion Model, HSMM, Similarity-based Approach, SVM	X					X			X						X					X	X	X	X		
[62] Yan et al. (2016)	Extended GPM, Data-Level Fusion via Composite HI		X				X			X						X						X		X		
[63] Yang et al. (2016)	Back Propagation Neural Network (BPNN), ELM		X				X			X	X			X		X	X				X			X		X
[64] Yongxiang et al. (2016)	IBL, Weighted Euclid Distance (WED)	X					X		X		X					X				X			X			
[65] Yuan et al. (2016)	SVM Sample Labeling, LSTM/Gated Recurrent Unit (GRU)-LSTM/AdaBoost-LSTM	X	X	X	X		X	X	X	X	X										X			X		
[66] Aydin & Guldamlasioglu (2017)	LSTM															X	X				X			X		
[67] Bektas et al. (2017)	Multiple LR, SSM	X	X	X	X		X	X	X	X												X		X		
[68] Bouzidi et al. (2017)	Non-linear Autoregressive Neural Network, Adaptative-Network-based Fuzzy Inference System					X	X			X											X			X	X	
[69] Chehade et al. (2017)	Degradation Estimation with Convex Quadratic Formulation, Bayesian Updating																					X		X		
[70] Chen et al. (2017)	Three Layer Non-linear Multistate Deterioration Model, NHSMM	X					X		X											X		X		X	X	
[71] Dong et al. (2017)	LSTM					X	X			X											X			X		
[72] Fang et al. (2017)	Adaptive Penalized (Log)-Location-Scale Regression	X					X		X									X		X	X			X	X	
[73] Khelif et al. (2017)	SVR	X					X		X						X			X			X			X		
[74] Krishnan et al. (2017)	Deep Neural Network (DNN)					X	X			X									X		X			X		
[75] Li et al. (2017)	Locally Weighted LR, Degradation-Dependent Weights, Ensemble of RVM/SVM/EF/QF/RNN					X	X			X						X						X		X		
[76] Lim et al. (2017)	Switching KF, Ensemble of MLP Networks	X	X	X	X	X	X	X	X	X		X										X		X		
[77] K. Liu et al. (2017)	Signal-to-Noise-Ratio (SNR)-based Data Fusion Model, Bayesian Updating	X					X		X							X						X		X		
[78] L. Liu et al. (2017)	GPR	X	X	X	X		X	X	X	X							X							X	X	
[79] Mathew et al. (2017)	LR, DT, SVM, Random Forest (RF), KNN, K-Means, Gradient Boosting, AdaBoost, DNN, Anova	X	X	X	X		X	X	X	X											X			X		
[80] Salah et al. (2017)	ESN, PSO	X					X		X							X					X			X		
[81] Singh et al. (2017)	Stacking Ensemble of Gradient Boosted Trees/Feed-Forward NN, NN as Meta Learner	X	X	X	X		X	X	X	X	X										X			X		
[82] Wang et al. (2017)	Trajectory Matching with Information Fusion	X					X		X							X							X	X		
[83] Xi et al. (2017)	Fractional Brownian Motion (FBM) Degradation Model, MCS	X					X		X				X									X		X		
[84] Zhang et al. (2017)	Ensemble of Multi-Objective Deep Belief Networks	X	X	X	X		X	X	X	X		X				X					X			X		
[85] Zhao et al. (2017)	Adjacent Difference Neural Network (ADNN)	X				X	X			X	X											X		X		
[86] Zheng et al. (2017)	LSTM	X	X	X	X	X	X	X	X	X	X										X			X		
[87] Zhou et al. (2017)	Reduced Kernel Recursive Least Squares Algorithm, HMM	X					X		X			X	X			X						X		X		
[88] Baraldi et al. (2018)	ELM		X				X			X	X			X			X					X				X
[89] Elattar et al. (2018)	KF, MLP Combination for Inference and Projection					X	X			X						X						X		X		
[90] Hsu & Jiang (2018)	LSTM	X	X	X	X		X	X	X	X	X										X			X		
[91] N. Li et al. (2018)	WPM, PF with Fuzzy Resampling	X					X		X							X						X				X
[92] X. Li et al. (2018)	CNN	X	X	X	X		X	X	X	X		X				X					X			X		
[93] Lim et al. (2018)	MLP	X	X	X	X		X	X	X	X		X	X					X		X	X			X		
[94] Lin et al. (2018)	Integrated Hierarchical Learning via Stacked ELM-based Auto-Encoder	X					X		X												X			X		
[95] Rigamonti et al. (2018)	Ensemble of ESN, Mean Variance Estimation					X	X			X	X			X			X				X					X
[96] Schlegel et al. (2018)	Distribution-based Similarity Estimation, Bucketized RUL Regression via RF/SVR		X		X	X	X	X		X							X	X					X	X		
[97] Song et al. (2018)	SNR- and Kernel-Based Data Fusion Model, Bayesian Updating	X					X		X							X						X			X	
[98] Q. Wu et al. (2018)	Combination of LSTM Networks and Gaussian Mixture Model	X					X		X			X	X				X				X			X		
[99] Y. Wu et al. (2018)	Vanilla LSTM	X	X	X	X		X	X	X	X	X										X			X		
[100] Z. Wu et al. (2018)	Cluster-based HMM	X					X		X													X				
[101] Xi et al. (2018)	FBM Degradation Model, Weak Convergence Theorem for Approximation			X				X	X				X			X						X		X		
[102] Zeming et al. (2018a)	Operational Reliability Fused Index, Multi-factor Similarity Measurement	X					X		X								X						X			
[103] Zeming et al. (2018b)	Degradation Degree Weighted Similarity Measurement	X					X		X								X						X			
[104] Zhang et al. (2018a)	Single-Layer Perceptron, LSTM	X		X			X	X	X		X					X						X		X		
[105] Zhang et al. (2018b)	Single-Layer Perceptron, Bi-directional LSTM	X					X		X		X					X						X		X		
[106] Zhou et al. (2018)	Echo State Kernel Recursive Least Squares Algorithm, Bayesian Updating											X				X						X		X		
Total Number of Coverage		**60**	**24**	**19**	**20**	**36**	**97**	**22**	**60**	**57**	**20**	**12**	**6**	**6**	**5**	**44**	**18**	**10**	**1**	**13**	**36**	**54**	**19**	**93**	**6**	**11**

Table 2. Application of the Taxonomy on Reviewed Studies