Proceedings of the 8th European Conference of the Prognostics and Health Management Society 2024 - ISBN – 978-1-936263-40-0

Towards a Hybrid Framework for Prognostics with Limited Run-to-Failure Data

Luc S. Keizers[1,2], Richard Loendersloot[1], Tiedo Tinga[1,2]

[1] *University of Twente, Enschede, 7522NB, the Netherlands*
l.s.keizers@utwente.nl,
r.loendersloot@utwente.nl,
t.tinga@utwente.nl

[2] *Netherlands Defence Academy, Den Helder, 1781AC, the Netherlands*

Abstract

The introduction of cyber-physical systems with increased availability of sensor data creates a lot of research interest in prognostic algorithms for predictive maintenance. Although a lot of algorithms are successfully applied to benchmark case studies based on simulated data and experimental set-ups, deployment in industry lags behind. From a comparison between three benchmark case studies with two real-world case studies based on prognostic metrics (monotonicity, prognosability and trendability), two main issues are observed: 1) the lack of run-to-failures and 2) low prognostic metrics due to a low signal-to-noise ratio of degradation trends, as a result of unexplained physical phenomena. To make prognostics feasible, a hybrid framework is proposed that focuses on improving system knowledge. The framework consists of a quantitative diagnostic assessments, guided by (modular) system models in which damage is induced. This quantitative damage assessment provides input for prognostics based on Bayesian filtering, enabling prognostics for assets in varying operational conditions. Implementation and validation of the framework requires investments, but modularity within the framework can accelerate development for new systems.

1. Introduction

During the fourth industrial revolution, cyber-physical systems are being introduced where sensor data communicates between machinery and with operators (Pinciroli et al., 2023). This sensor data can be used to find characteristics of failures within the systems which can be used to develop models that predict the remaining useful life (RUL) (Yan et al., 2017), giving new opportunities for implementation of *predictive maintenance*. When failures can be predicted, catastrophic accidents are prevented, unexpected downtime can be reduced, components are used until the end of their actual lifetime and maintenance logistics can be optimized (Fernandes et al., 2022).

The recent increasing interest in *predictive maintenance* is clearly visible by observing the amount of published scientific papers in this field. Only the number of review papers is already growing significantly, as the number of counts in the Scopus database on article titles with (*survey* or *review*) and (*predictive maintenance*) grew from a total of 20 published documents up to 2020 to a total of 84 published documents up to 2023.

The increasing number of sensors and data availability, specifically increase the interest in *data-driven* prognostic approaches (Pinciroli et al., 2023). This type of approach requires sufficient historical run-to-failure data. However, in safety-critical systems (Chao et al., 2021) or when availability of assets is more important than costs (Tinga et al., 2021), failures are sparse and as a consequence the required historical run-to-failure data are rarely being collected. Also for new types of machinery, no historical data are available (Calabrese et al., 2021). If historical data are available, they are often unlabeled and unorganized, lacking context such as operating conditions and maintenance recordings (Calabrese et al., 2021; Lukens et al., 2022).

In contrast to data-driven approaches, physics-based approaches considering Physics-of-Failure (PoF) models have less strict data requirements. They provide a relation between usage and degradation rates (Tinga, 2013b). This yields benefits compared to purely data-driven approaches, specifically when failures are rare and when future operating conditions (and consequently degradation rates) are different from historical operating conditions (Tiddens et al., 2023). However, these models are expensive to develop and are component or system specific (Elattar et al., 2016). Also, the relation between usage and degradation rates should be known and must not be too complex.

To overcome issues with purely data-driven or purely physics-based prognostics, combinations of them (i.e. hybrid approaches) are often proposed as a solution (Elattar et al., 2016; Guo et al., 2020). Still, many hybrid methods use data-driven models to estimate the degradation behavior (Pugalenthi et al., 2021; Borutzky, 2020) or only use physics to improve input parameters of a data-driven prognostic algorithm (Gálvez et al., 2021; Chao et al., 2022). This yields improved results compared to purely data-driven approaches, but the relation between usage and degradation rates are generally not considered, still limiting applicability in cases of rare historical failures and varying operating conditions.

Fernandes et al. (2022) described that in many studies, real-world challenges are often overlooked and additional research to address these real-world challenges is needed. This suggests a gap between science and industry. This suggestion is strengthened by a survey in 280 companies in Belgium, Germany and the Netherlands, showing that only 11% are actually implementing predictive maintenance techniques in 2017 (Mulders & Haarman, 2017). Although a similar survey showed that this number increased to 17% in 2023, mainly attributed to original equipment manufacturers (OEMs) and sectors with large numbers of the same assets (van der Velde et al., 2023), a large majority of developed methods is still only being applied to experimental or simulated data sets (Ferreira & Gonçalves, 2022) such as C-MAPSS (Saxena & Goebel, 2008).

In this paper, issues with implementation of prognostics in two real-world cases are pinpointed. The case studies are performed by two organizations within the Netherlands who try to implement predictive maintenance for military assets. The case studies are challenging, as availability of these assets is more important than costs (Tinga et al., 2021), fleets are relatively small and the assets operate in varying operational and environmental conditions. Prognostic metrics defined by Coble (2010) (monotonicity, prognosability and trendability) are calculated for the two real-world case studies, and also for three well-known benchmark cases to compare the potential for application of prognostic algorithms. Based on observed issues in the real-world cases, a possible solution is proposed in the form of a hybrid framework.

The remainder of the paper is organized as follows. Section 2 starts with calculating prognostic metrics of features from three benchmark cases: the Virkler crack growth data set, the NASA milling data set and the C-MAPSS data set. Then, metrics are calculated for features from the two real-world case studies, concerning condition monitoring of Apache helicopter engines and a naval main diesel engine respectively. Following from the issues observed in the case studies, section 3 proposes a hybrid framework for prognostics. Lastly, section 4 discusses the results and concludes the paper.

2. Prognostic Potential of Case Studies

2.1. Prognostic Metrics

Coble (2010) developed three metrics to assess the suitability of features as input for a (data-driven) prognostic algorithm. The suitability is evaluated based on monotonicity (M), prognosability (P) and trendability (T). The range of the scores is from 0 (unsuitable) to 1 (perfectly suitable). The weighted sum of the three metrics gives the prognostic score, and features with the highest score are most suitable to be used as the input for a prognostic algorithm. So, data sets from which features with high scores can be derived, have high potential for prognostics.

The first metric is monotonicity, assessing the extent to which run-to-failure trajectories are purely increasing or decreasing. It is calculated as follows:

$$M = \text{mean}\left(\left|\frac{N^+ - N^-}{n-1}\right|\right) \tag{1}$$

with N^+ the number of increments in the run-to-failure trajectory of the feature (i.e. $n_{i+1} - n_i > 0$), N^- the number of decrements in the trajectory (i.e. $n_{i+1} - n_i < 0$) and n the number of data points in the trajectory. The absolute mean monotonicity of all considered run-to-failure trends yields the final monotonicity.

Prognosability estimates how similar the start values and the values at failure are for the features when comparing different run-to-failure trajectories. It is calculated as follows:

$$P = \exp\left(-\frac{\text{std}(\mathbf{f}_{end})}{\text{mean}\left(|\mathbf{f}_{end} - \mathbf{f}_0|\right)}\right) \tag{2}$$

with $\mathbf{f}_{end}$ a vector with all values of the features at failures and $\mathbf{f}_0$ a vector with all values of the features at the start of the run-to-failure trajectories. Std refers to the standard deviation.

Trendability describes similarity between the shapes of run-to-failure trajectories. It is calculated as follows:

$$T = \min\left(|\boldsymbol{\rho}_{ij}|\right) \tag{3}$$

with $\boldsymbol{\rho}$ a vector with the correlation coefficients between each run-fo-failure trajectory i and j of the feature. For trajectories with different lengths, linear interpolation is applied such that the lengths of the correlated trajectories match.

The final score S is calculated by:

$$S = W_m \cdot M + W_p \cdot P + W_t \cdot T \tag{4}$$

with W_M, W_P and W_T the weight factors for monotonicity, prognosability and trendability respectively. In many applications they can be set identically, but in some applications some metrics may be less relevant (Coble, 2010). In the case studies discussed in the next subsection, the weight factors are all set to $\frac{1}{3}$, yielding prognostic scores ranging from 0 to 1.

2.2. Benchmark Data Sets

2.2.1. Virkler Crack Growth

The first benchmark data set considered is the Virkler crack growth data set. Virkler et al. (1979) performed 68 run-to-failure fatigue tests of 2024-T3 aluminium and measured the crack length directly. This yields the run-to-failure trajectories as shown in Fig. 1.

The data set contains crack lengths at specific numbers of stress cycles. As crack lengths already provide a direct indicator of the damage severity, no additional features need to be calculated and the prognostic metrics are directly calculated on the crack length measurements. As the crack length always increases monotonically, the monotonicity is 1. All trajectories have the same end value (50mm) and starting value (9mm), yielding a prognosability of 1. As all trajectories are monotonically increasing, all trajectories have a perfect positive correlation, yielding a trendability of 1. Consequently, the total prognostic score is 1.

A direct measurement of degradation can be considered as a perfect prognostic metric, as the nature of degradation (an irreversible process) makes it monotonic, a proper threshold can be defined based on system knowledge and the monotonicity yields also perfect trendability. As the underlying model is well understood, the Virkler data set is perfectly suitable for prognostic methods based on Bayesian updating (Sun et al., 2014; Baral et al., 2023), but also data-driven methods are well applicable (Eker & Jennions, 2012).

2.2.2. Milling Tool Wear

The Milling Data Set (Agogino & Goebel, 2007) contains data collected from an experimental setup for tool wear estimation. There are two operational settings for the Depth of Cut (DOC), feed rate and material. Two experiments are performed for each combination of operational settings, yielding a total of 16 experiments. Each experiment consists of a number of runs, lasting 72s. After each run, the actual tool wear (i.e. *VB*: the measured distance from the cutting edge of the tool to the end of abrasive wear on the flank (Agogino & Goebel, 2007)) is measured with a microscope. Experiments were terminated at a certain (not further specified) tool wear limit (and some beyond). As measuring tool wear with a microscope after each operation is not feasible in practical applications, measurements are also collected from sensors that can provide an online estimation of tool wear. A total of six sensor are installed which measure at a sampling rate of 250Hz: AC and DC motor current sensors of the spindle, and vibration and acoustic emission sensors at both the spindle and the table.

The data set is shown in Fig. 2. Fig. 2a shows the offline tool wear depth (*VB*) measurements. As most experiments run until $\approx 0.50 - 0.80$mm as shown by the gray band in the figure, it is expected that the tool wear limit is around this band. The gaps in the trajectories in Fig. 2a are due to missing measurements in between some of the runs. Note that only 14 out of the 16 experiments are displayed: for one of the experiments, only one run is available, making it unsuitable to calculate prognostic metrics. Both experiments in the corresponding

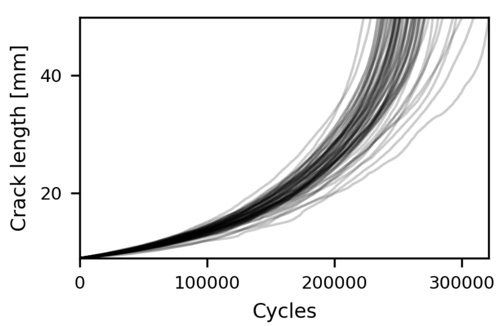
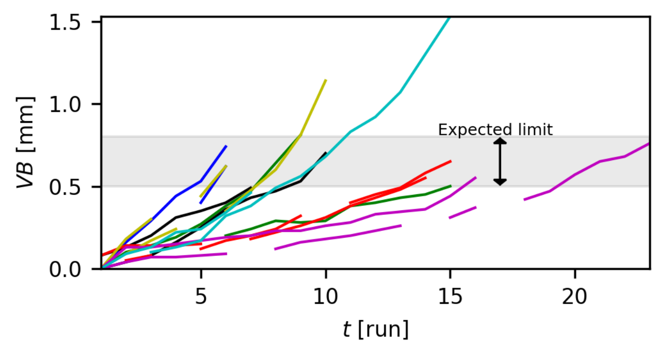
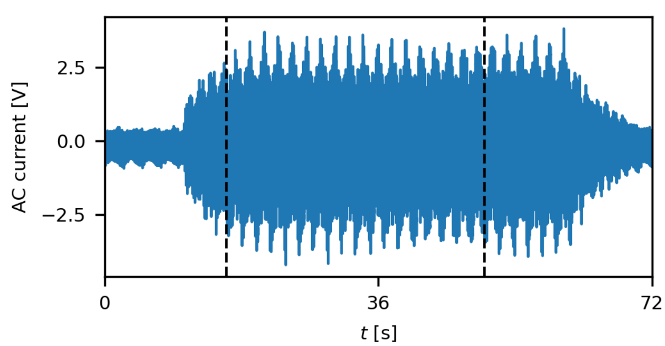


Figure 1. Virkler Crack Growth Data Set Virkler et al. (1979) (M = 1, P = 1, T = 1, S = 1)



(a) All tool wear trajectories in milling data set. Each color corresponds to a set of operational settings (M = 0.95, P = 0.66, T = 0.78, S = 0.80)



(b) Example (AC motor current) measurements for one run in milling data set

Figure 2. Visualization of the milling data set

operational conditions are removed from the data set, such that seven pairs of experiments remain.

The trajectories for the *VB* measurements yield a monotonicity M of 0.95, prognosability P of 0.66, trendability T of 0.78 and S of 0.80. Although perfect metrics of 1 are expected for direct degradation condition measurements, as explained in subsection 2.2.1, some non-monotonic behavior can be observed in Fig. 2a (e.g. the dark green line before $t = 10$), reducing M and T. As wear is irreversible, this non-monotonic behavior is likely to be due to measurement errors. P is affected by the fact that some experiments were performed beyond the tool wear limit. Still, the prognostic metrics are high, and can be improved by reducing measurement error and running experiments until a fixed failure threshold.

A challenge is to estimate tool wear from the real-time sensor data. Fig. 2b shows an example of data from the AC motor current sensor for one run of an experiment. No clear trend can be observed in the raw sensor data, so features need to be calculated. As the start and run of an experiment yield no stable signal (i.e. magnitude increase and decrease as seen in Fig. 2b), only the stable period (defined to be 16-50s, indicated by the vertical dashed bars in Fig. 2b) is considered for feature calculation. For each sensor and each run, the mean, standard deviation, maximum, minimum, absolute maximum, absolute minimum, root mean squared and sum of values are calculated.

The standard deviation from the AC motor current measurements is found to have the highest prognostic score and its trajectories are shown in Fig. 3. M is 0.74, P is 0.57 and T is 0.84, yielding a score of 0.72. The different operating conditions clearly yield different feature values, as only the start and end values of curves with the same operating condition have approximately the same start- and end values (i.e. the dark green, black, red and pink pairs). Prognostic scores can be further improved by compensating for operating conditions, by calculating the P and T for two experiments with the same

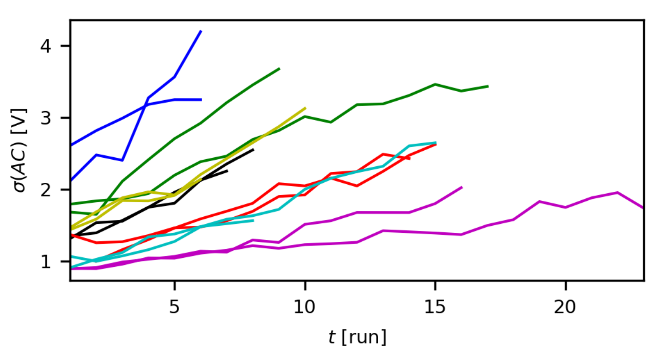


Figure 3. Feature with highest score from milling data set, color-coded by operating conditions
(M = 0.74, P = 0.57, T = 0.84, S = 0.72)

operating conditions and taking the mean of the separately calculated P and T (note that M is unaffected). It is found that in this case, P increased to 0.78 and T increased to 0.86, yielding a final score of 0.79, which is almost the same score as for direct *VB* measurements.

This case shows that by calculating only a simple set of features, features with high prognostic potential can already be obtained. These characteristics make the data set applicable for prognostics. In literature, mainly quantitative diagnostic methods are applied, based on e.g. nearest neighbor-based approaches (Sheng & Zhu, 2020), Recurrent Neural Networks (Lu et al., 2022), Kernel Extreme Learning Machines (Zhou & Sun, 2020), particle filters (P. Wang & Gao, 2016) and Long Short-Term Memory Networks (Kumar et al., 2022). Subsequently, prognostics can be performed (J. Wang et al., 2015).

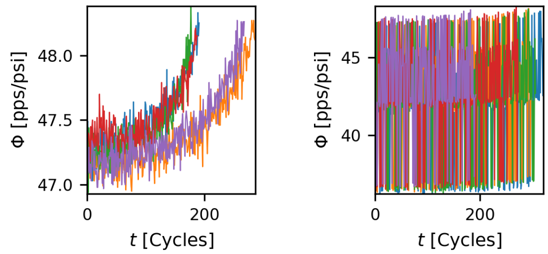
2.2.3. C-MAPSS

The Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) (Saxena & Goebel, 2008) is a very popular benchmark data set due to the inclusion of sensor noise, different operating conditions and multiple fault modes (Ramasso & Saxena, 2014). Already in 2014, Ramasso & Saxena (2014) published a review paper on 70 different prognostic methods utilizing the data set. The data set is still being used and more recently a new version of this data set (N-CMAPSS) has been released (Chao et al., 2021). The data is generated with a simulator built in Matlab and Simulink. The operational settings are defined by three parameters and 21 (virtual) sensors are available.

In this paper, the FD001 and FD004 train data sets are evaluated. The FD001 set contains 100 degradation trajectories in one operating condition and one fault mode (High Pressure Compressor (HPC) degradation). The FD004 set contains 248 degradation trajectories in six operating conditions and two degradation modes (HPC degradation and fan degradation). As an example, Fig. 4 shows raw sensor data from one of the sensors (Φ, a fuel flow ratio) for five run-to-failure trajectories of FD001 and FD004. In Fig. 4a the degradation trends can be clearly observed, which is more difficult in Fig. 4b due to the effect of changing operating conditions on the data.

To reveal the degradation trend for the FD004 data set, a K-Nearest Neighbors regressor is trained on the first 40 data points to learn the (nominal) relation between the three operational settings and measurements. The regressor is built using the $sklearn$ Python package and uses two neighbors. Fig. 5 shows that the residuals (i.e. difference between measured and expected measurements) reveal similar degradation trends as was observed for the FD001 data set.

Although the raw sensor data of FD001, or the residuals for FD004, already reveal a strong degradation trend, better prog-



(a) Five run-to-failure trajectores of Φ in constant operating conditions (FD001)
(b) Five run-to-failure trajectores of Φ in varying operating conditions (FD004)

Figure 4. Five run-to-failure trajectories of Φ in the C-MAPSS data set

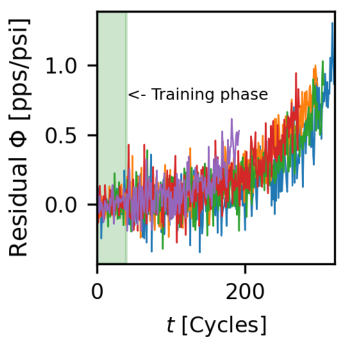
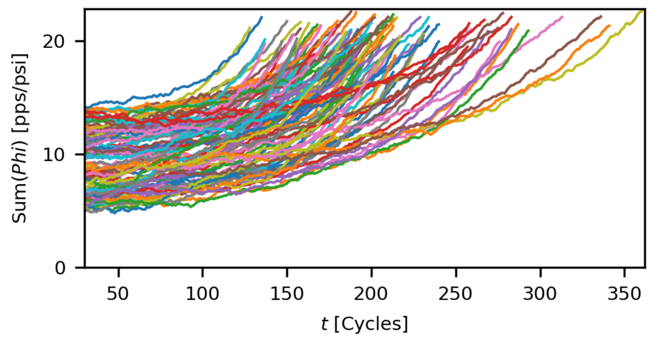
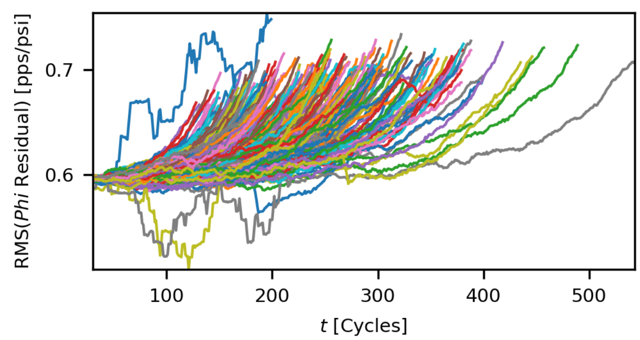


Figure 5. Five run-to-failure residual trajectories of Φ sensor in varying operating conditions (FD004)



(a) Top feature of C-MAPSS FD001 data set (directly derived from sensor data) (M = 0.69, P = 0.88, T = 0.94, S = 0.84)



(b) Top feature of C-MAPSS FD004 data set (derived from residuals) (M = 0.55, P = 0.80, T=0.06, S =0.51)

Figure 6. Best features on CMAPSS data set

nostic metrics can be found by extracting features. This is done for both the FD001 raw measurements and the FD004 residuals using the Python package *tsfresh* (Christ et al., 2018) for automatic feature extraction. It first creates a rolling window over the data set and for each window, features are calculated. The window size is a trade-off between noise reduction and response time to signal changes, and is set to 30 by trial-and-error. To reduce noise in the calculated features, the prognostic metrics are calculated for each 5th data point.

The best performing feature for the FD001 and FD004 data set are shown in 6. For FD001 the sum of Φ measurements in the rolling window obtained the highest metrics, as shown in Fig. 6a: M=0.69,P=0.94 and T=0.84, yielding a score of 0.84. For FD004 the root mean square of the Φ residuals gave the highest metrics, which are significantly lower (i.e. M=0.55,P=0.80 and T=0.06, yielding a score of 0.47) compared to FD001. This is mainly caused by some outlying trajectories (e.g. the blue trajectory on the left in Fig. 6b) due to faulty residual calculations (e.g. because not all operating conditions are observed in the training phase of the KN-regressor). It is found that the mean of the correlation coefficients of all trajectories is 0.97, but the fact that T is determined by the lowest correlation coefficient between all trajectories yields the low trendability. Therefore, higher scores can be obtained when removing outliers and improving residual generation.

To conclude, it is found to be straightforward to retrieve features with high prognostic metrics, mainly for the FD001 data set. For the FD004 data set it is more challenging due to the varying operating conditions. However, features could be extracted which show similar run-to-failures trajectories, although additional effort is required to remove outliers and improve prognostic metrics further. The general characteristics of the data set make it feasible for data-driven prognostics, and methods such as Nearest-Neighbors, Random Forests, Extreme Gradient Boosting and Multilayer Perceptrons (Alomari et al., 2023), Convolutional Neural Networks-based approaches, Long-Short-Term-Memory-based approaches (de Pater et al., 2022) and others are widely found in literature.

2.3. Real-world Case Studies

This subsection introduces two real-world case studies. Despite an extensive search, only two cases were found to have sufficient measurements and meta data available to calculate metrics of run-to-failure trajectories. Organizations often do not have, cannot or do not want to disclose the data necessary for a proper analysis. The authors thank NLR and the Royal Netherlands Navy gratefully for making these case studies available for evaluation.

It should be noted that the case studies concern the most complex cases for prognostics: they concern monitoring of individual assets in varying operating conditions, where future operating conditions can be different from historical operating conditions. This corresponds to the highest ambition level an organization can have, with high requirements on data availability or system knowledge (Tiddens et al., 2023).

2.3.1. Apache ETF Monitoring

The power level of helicopter turboshaft engines decreases over the lifetime due to wear of seals, vanes and blades or due to faults in other components (Vos, 2019). Engine performance is measured by the Engine Torque Factor (ETF), which is the ratio between the actual engine power and the rated engine power. If the ETF drops below 0.85, or if the combination of the ETFs of both tail engines of an Apache helicopter drops below 0.90, the Apache is not allowed to be used. The Netherlands Aerospace Centre (NLR) developed an algorithm to calculate the ETF from in-flight parameters, rather than from time-consuming manual Max Power Checks (MPCs) (Vos, 2019).

Vos (2019) selected the turbine gas temperature (TGT) as health indicator, and fitted a polynomial model using data from the Health and Usage Monitoring System (HUMS) to translate the in-flight TGT to the TGT at the reference condition (which in turn allowed to calculate ETF). The considered operational parameters are gas inlet temperature, outside air temperature, pressure altitude, speed, and engine torque.

For the development of a prognostic algorithm, run-to-failure trajectories are required. Although the engine needs an overhaul when the ETF reaches 85%, the data set does not contain any trajectory running till this threshold. Therefore, to calculate the prognostic metrics of trajectories, periods between (documented) engine replacements are selected. This does not fully represent run-to-failure, but no better option is available.

The rolling mean of these ETF trajectories is calculated over five ETF measurements, yielding the trajectories in Fig. 7. The calculated metrics over these trajectories are: M=0.09, P=0.19 and T=0, yielding the extremely low score of 0.12. This can be expected from Fig. 7, as the data are covered in low-frequency noise which make the actual degradation trend barely visible. This low-frequent wobbling behavior is caused by physical phenomena not explained by the polynomial model (Vos, 2019), i.e. by confounding factors. This noise yields a low signal-to-noise ratio (i.e. degradation to other external influences), and therefore low prognostic metrics.

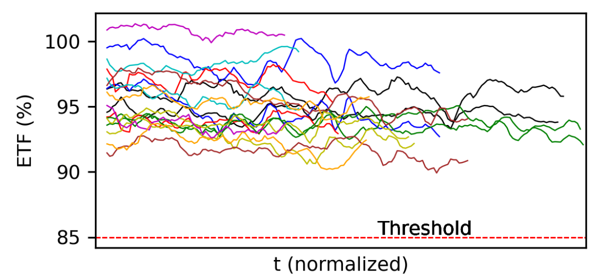


Figure 7. All ETF trajectories considered. Each color corresponds to the engines of a specific Apache. (M = 0.09, P = 0.19, T = 0, S = 0.12)

Real-time estimation of the ETF offers potential to replace expensive MPC by real-time ETF monitoring (i.e. additional inspections can be performed when the ETF drops below the threshold). However, the lack of actual run-to-failures and the low signal-to-noise ratio make the step towards prognostics extremely complicated.

2.3.2. Marine Diesel Engine Bearing Monitoring

The main diesel engine (MDE) of a naval vessel contains seven journal bearings that support the crankshaft. Failure of one of the bearings yields failure of the MDE and is therefore critical for availability of the vessel. Heek (2021) developed a monitoring method based on bearing temperature to detect failures timely. The underlying idea is that damage increases friction in the bearing and subsequently increases the bearing temperature.

Because the bearing temperature is also affected by operating conditions, Heek (2021) fitted data from the Integrated Platform Management System (IPMS) in a multiple linear regression model. This model estimates the bearing temperature in nominal conditions based on the RPM of the engine, RPM of the turbocharger, and the lube oil temperature at the outlet, which were found to have the highest predictive performance. Subsequently, the residual between the measured and the predicted bearing temperature is selected as a health indicator and is continuously monitored. Alarms can be raised when the measured temperature is higher than expected. The complete procedure of data selection and residual generation can be found in Heek (2021).

Heek (2021) evaluated three case studies, from which two concerned an actual failure. Cases 1 and 2, which concern failure cases, are visualized in Fig 8. The regression model is trained within the shaded areas (until $t = 993$ and $t = 1748$ respectively) and the residual monitoring is deployed after (note that case 2 runs longer than case 1). To evaluate prognostic metrics of these trajectories, every 10th data point is selected and the rolling mean over 10 data points is calculated. The calculated metrics are: M=0.06, P=0.92 and T=0.57,

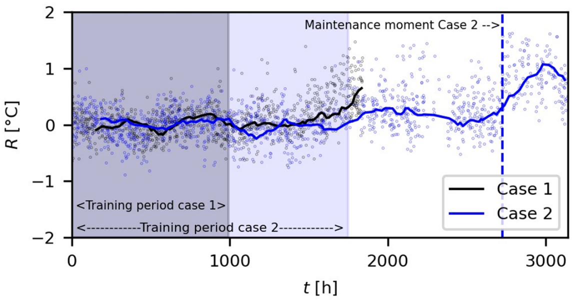


Figure 8. Trajectories of diesel engine temperature residuals with failure at the end (M = 0.06, P = 0.92, T = 0.57, S = 0.52)

yielding a score of 0.52.

The monotonicity of the trends is extremely low due to low-frequency oscillations over time. According to Heek (2021), this is caused by unobserved effects in the engines, probably caused by "*minor maintenance actions*", which again can be considered as confounding factors. P is relatively high due to the small difference between end values (i.e. 0.5 and 0.6 respectively), and a weak trendability is observed due to the increase in data points near the end.

However, the actual meaning of these high metrics is disputable. Heek (2021) described that in case 2, maintenance was performed around $t = 2700$h, as indicated by the vertical bar in Fig. 8. After this moment, the residuals make a jump of 1°C. This behavior is also visible when evaluating the third case study in which no failure was observed, shown in Fig. 9. Similar to case 2, maintenance was performed after which residuals make a jump (of approximately 0.6°C) as indicated by the vertical bar around $t = 1200$h.

Again, this yields a gradually increasing trend, but it is unrelated to degradation. Heek (2021) proposed to retrain the regression model after maintenance is performed to reduce the number of false positives, which may provide a solution for

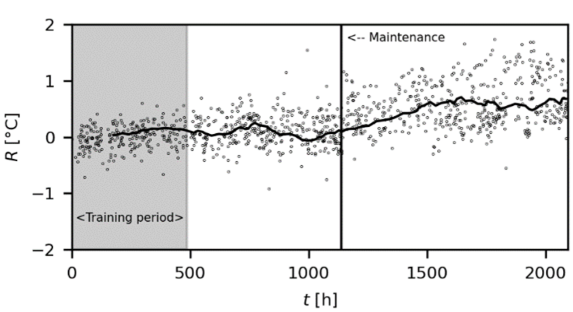


Figure 9. Trajectories of diesel engine temperature residuals without failures

anomaly detection. However, this way the physical relations and meaning of the residuals is inconsistent, i.e. there is no clear link between the residuals and the quantitative amount of damage. This makes it difficult to isolate maintenance actions (or other external confounding factors) from degradation, i.e. the signal-to-noise of the residuals can be considered to be low.

Real-time monitoring the residuals offers potential for early fault detection. However, similar to the ETF case study, it is still extremely complicated to implement prognostics due to the limited number of run-to-failure trajectories and a low signal-to-noise ratio due to limited physical understanding of the data. An additional problem in this case, is that the lack of physical understanding of the residuals complicate threshold definition, and without a threshold it is impossible to estimate the RUL.

2.4. Discussion and Conclusion Case Studies

A first observation of the case studies is that a good direct measurement of the degradation severity (e.g. crack length) can be considered as a perfect prognostic feature. As degradation is in general an irreversible process, it is monotonic and yields perfect trendability. Also thresholds can be defined based on phyical knowledge, yielding perfect prognosability. However, in practice it is complicated or impossible to obtain direct measurements of the actual degradation severity and real-time sensors should provide monotonic, prognosable and trendable run-to-failure trajectories. It is found to be relatively easy to extract such trajectories from well-defined benchmark data sets with labeled data (i.e. milling data set (Agogino & Goebel, 2007)) or many historical run-to-failures (i.e. C-MAPSS (Saxena & Goebel, 2008)).

However, both real-world case studies suffered from two main issues: 1) the number of failures is extremely low, or even non-existent, and 2) low prognostic metrics due to a low signal-to-noise ratio between the health indicator (i.e. the signal) and other confounding factors such as maintenance actions and environments (the noise). The low number of failures, with low prognostic potential in the derived feature, make training of data-driven prognostic algorithms infeasible. Furthermore, the low signal-to-noise ratio makes it complicated to distinguish nominal operating conditions from faulty behavior, such that estimation of the onset of degradation, as well as identifying the degradation trends are difficult.

Note that the latter issue is mainly contributed to confounding factors affecting system behavior, originating from varying operating conditions. This shows the main difference between the benchmark and real-world data sets: in benchmark data sets, either an experimental setup or a simulation is used to generate data, in which external factors influencing system behavior can easily be excluded (lab experiment) or by definition do not exist (simulation). However, in real-world cases such factors are not always well understood or measured such that they are not included in the developed models. Note that available prognostic algorithms can work if data requirements are met (e.g. as found by van der Velde et al. (2023), chances are higher for assets in large fleets, and requirements are less strict in constant operating conditions (Tiddens et al., 2023)), but a solution needs to be found for these complex real-world cases in varying conditions.

3. Proposed Framework

The lack of run-to-failures and the lack of physical understanding of the monitored signal complicated prognostics in the real-world case studies. Following the decision framework proposed by Tiddens et al. (2023), two solutions are possible: 1) improving the data set or 2) improving the system knowledge. Collecting more run-to-failures is not a realistic option as the failures are critical and therefore prevented by performing preventive maintenance actions. Worldwide data sharing may help (Coelho et al., 2022), especially for similar assets existing in large fleets, such as industrial machinery (Peng et al., 2022) and wind turbines (Li et al., 2021). However, data sharing is often complicated due to standardization and different data structures (Coelho et al., 2022). Furthermore, the military applications bring additional challenges regarding confidentiality of data, but also companies may not be keen on sharing data as they can consider it as intellectual property.

Therefore, the focus in development of the framework (Fig. 10) is on improving system knowledge. System knowledge can be improved in two ways: 1) learning the relation between usage and the degradation rates (i.e. with PoF-models) and 2) and quantifying the relation between measured signals and damage severities. The first part (part I in Fig. 10) is already explored by previous work of the authors (Tinga, 2013a; Keizers et al., 2021, 2022). The second part (part II in 10) focuses on improved quantitative diagnostics to obtain features with higher prognostic metrics, as found to be required for the real-world case studies discussed in section 2.

Degradation can vary heavily between assets used in varying operational conditions (Tiddens et al., 2023), and in absence of historical run-to-failures, the quantitative relation between usage and degradation is essential for accurate prognostics. Therefore, PoF-models are used in part I of the framework. Such models are often tuned for prognostics of specific assets with Bayesian filters (Jouin et al., 2016). However, in literature the effect of actual loading conditions is often simplified, e.g. by substituting loads with a constant model parameter (Zio & Peloni, 2011). This takes away one of the main strengths of a PoF-model, as handling the loads as a model parameter only makes extrapolation of the latest trend possible, yielding wrong RUL for changing future usage profiles, as was shown in Keizers et al. (2021).

To achieve PoF-based prognostics (i.e. part I of the framework) it is proposed to use the method described in Keizers et al. (2021), taking loads as separate input for a Bayesian filter and for prognostics. Loads are first monitored (for $t \leq t_p$, with t_p the time of prediction) to update the PoF-model. Then, expected future loads (for $t > t_p$) are substituted in the updated PoF-model for prognostics (see the lower input of the PoF-model in Fig. 10). This enables RUL prediction based on expected future operating conditions, or adaptation of system usage to extend the RUL.

A conceptual example can be given in the form of the Apache ETF case: it is observed that the ETF decreases faster in sandy environments Vos (2019), which can be explained by increased wear of vanes, blades and seals due to the increased number of sliding sand particles over the components. Erosive wear is already studied for decades (Sundararajan, 1991) and can be modeled by Archard law (Archard, 1953) or by more detailed empirical models that also take characteristics (e.g. hardness, size) of sand particles into account (Gülich, 2020). Such models can be used to estimate degradation rates in specific (and varying) operational and environmental conditions.

The update of the PoF-model requires corresponding degradation measurements. In the studies described in Keizers et al. (2021) and Keizers et al. (2022) direct condition measurements (i.e. measurement of the parameter calculated by the PoF-model) were assumed, which can be the available in some practical applications. For example, in case of fatigue crack growth, DC Potential Drop Methods can measure crack lengths directly (Bär, 2020) and in case of corrosion, electrochemical measurements can measure corrosion rates directly (Homborg et al., 2014). Such direct condition measurements are preferred, as they can be considered to be a perfect prognostic metric as discussed in section 2. However, in many practical applications, such as the real-world case studies of section 2, these types of direct condition measurements are expensive or impossible to obtain. Therefore, monitoring options are used that measure (indirect) consequences of the actual degradation, e.g. vibrations or temperatures.

Here, part II of the framework is introduced. It considers a quantitative diagnostic block, linking indirect condition measurements in specific operating conditions (the right-side input of the block) to the direct condition (i.e. damage severity). However, as illustrated by the case studies in section 2 the features derived from the real-world data sets have low prognostic metrics, and are unlabeled. Therefore, the relation between measured data and degradation severity is unknown, and quantitative diagnostic algorithms cannot be trained. Experimental set-ups could help to gather training data to learn this relation, but it is economically infeasible to collect data for all fault types and locations in all possible operating conditions (Sawalhi & Randall, 2008). Here, the second way of including system knowledge is relevant, which is positioned below the quantitative diagnostic block in Fig. 10. By introducing faults

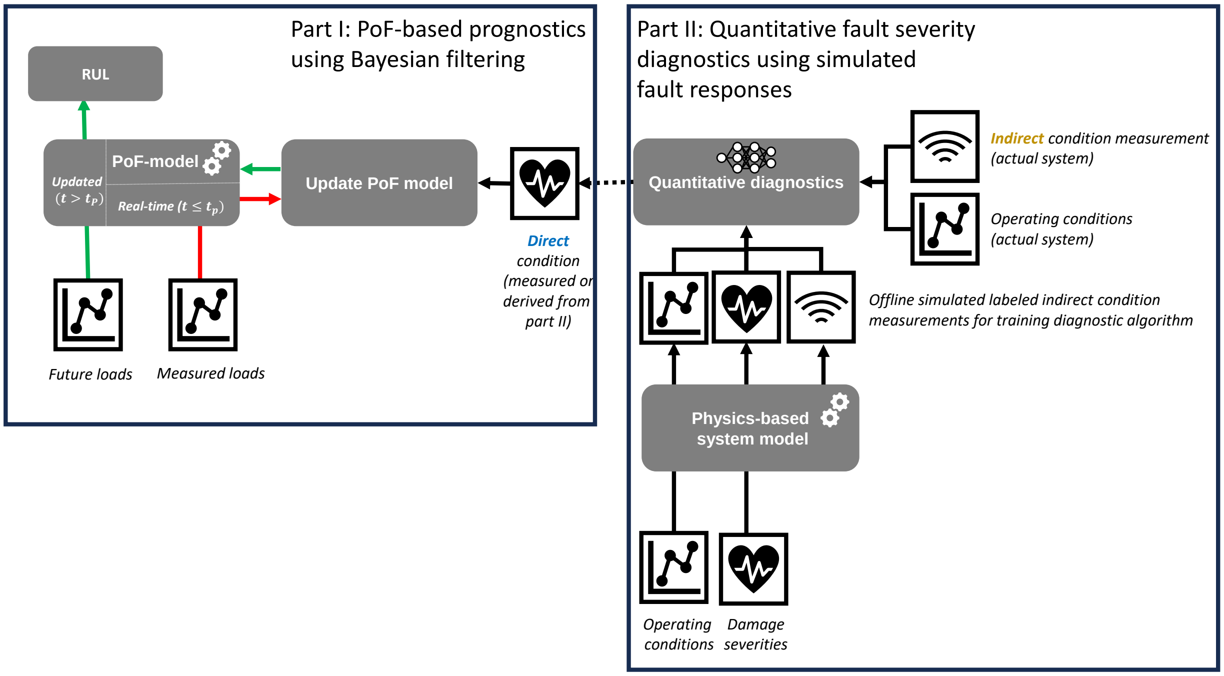


Figure 10. Proposed framework

in physics-based system models and (virtual) measuring the system response of degradation, a better understanding of the effect of faults with varying sizes on measured signals in different operating conditions can be obtained. In the conceptual example of erosive wear in the Apache components, a system model should reveal the efficiency loss for different amounts of material removal. Subsequently, the required additional power, and forthcoming turbine gas temperature, need to be modeled in different operational conditions.

Indeed, developing such a model brings major challenges. First, building an extensive physics-based model for each component or system is time-consuming and expensive. Second, damage induces responses in different physical domains (i.e. mechanical wear yields increased temperature). For these reasons, a modular modeling method that is relatively easy to reuse, adapt and scale, and which is applicable in multiple domains is proposed. Bond graphs are such models (Mkadara et al., 2021). It will be profitable to develop models of standard equipment (e.g. bearings) that can be easily reused in other system models, accelerating development for new machinery.

As an example, Nakhaeinejad & Bryant (2011) showed that different types of bearing faults and their vibration response can be modeled using these types of models. Note that bond graphs are proposed more often in hybrid prognostic frameworks for determining residuals from a nominal system model (e.g. (Medjaher & Zerhouni, 2013)) or for tracking faulty parameters (e.g. Borutzky (2020)), but the link with an actual PoF-model and its corresponding direct condition measurement is missing, limiting prognostic capabilities in cases of varying operating conditions.

To conclude, part II of the framework can help to create a quantitative damage assessment, acting as an input for part I in the likely scenario where direct condition measurements are unavailable. Subsequently, prognostics can be performed, and the RUL can be predicted based on assumed usage profiles. This not only improves prognostic performance in cases of varying operating conditions, but also offers the possibility to adapt usage profiles to extend the RUL.

4. Discussion and Conclusion

The paper showed that prognostic metrics of data from real-world data sets are extremely low compared to benchmark data sets. The main issues observed are unavailability of run-to-failure trajectories and low signal-to-noise ratios of the available trajectories due to always present confounding factors. As a consequence, developed data-driven prognostic methods are often not applicable in practical cases.

The criticality of discussed real-world cases make it unlikely that much run-to-failure data will be collected in the future, so the proposed solution is defined in a framework based on enhancing and utilizing system knowledge. The limited understanding of measured signals in the real-world cases make trending complicated, so system models with induced damage are proposed to increase understanding of the effects of damage on measured signals. A quantitative diagnostic algorithm can improve the signal-to-noise ratio of measured signals, providing the input for a Bayesian filter that quantifies the relation between usage and degradation rates. Subsequently, prognostics can be performed based on expected system usage.

The framework has strict requirements on system knowledge (i.e. PoF, load monitoring, system models with induced damage), but low data requirements as no historical run-to-failures are needed. To accelerate development for application to new systems, usage of modular system models such as bond graphs is proposed. The framework still needs to be implemented and validated for a real application and the strict requirements on system knowledge requires investments. However, it enables better RUL predictions (or RUL extension by usage adaptation) which can yield great benefits. Before moving to complex cases such as the Apache ETF or the marine diesel engine, it is proposed to start with relatively simple standard equipment such as bearings to validate the benefits of the method. This will be presented in a future publication.

ACKNOWLEDGEMENTS

This work is carried out as a part of the PrimaVera Project, funded by the NWO under grant agreement NWA.1160.18.238. The authors thank the NLR and Royal Netherlands Navy gratefully for making their case studies available for evaluation.

REFERENCES

Agogino, A., & Goebel, K. (2007). *Milling data set.* Retrieved from `https://data.nasa.gov/Raw-Data/Milling-Wear/vjv9-9f3x/data`

Alomari, Y., Andó, M., & Baptista, M. (2023). Advancing aircraft engine rul predictions: an interpretable integrated approach of feature engineering and aggregated feature importance. *Scientific Reports*, *13*(1), 13:13466. doi: https://doi.org/10.1038/s41598-023-40315-1

Archard, J. (1953). Contact and rubbing of flat surfaces. *Journal of Applied Physics*, *24*(8), 981 - 988. doi: https://doi.org/10.1063/1.1721448

Baral, T., Saraygord Afshari, S., & Liang, X. (2023). Residual life prediction of aluminum alloy plates under cyclic loading using an integrated prognosis method. *Transactions of the Canadian Society for Mechanical Engineering*, *47*(5), 1-12. doi: https://doi.org/10.1139/tcsme-2023-0010

Borutzky, W. (2020, Jan). A hybrid bond graph model-based - data driven method for failure prognostic. *Procedia Manufacturing*, *42*, 188-196. (International Conference on Industry 4.0 and Smart Manufacturing (ISM 2019)) doi: https://doi.org/10.1016/j.promfg.2020.02.069

Bär, J. (2020). Crack detection and crack length measurement with the dc potential drop method–possibilities, challenges and new developments. *Applied Sciences*, *10*(23), 8559. doi: https://doi.org/10.3390/app10238559

Calabrese, F., Regattieri, A., Bortolini, M., Gamberi, M., & Pilati, F. (2021). Predictive maintenance: A novel framework for a data-driven, semi-supervised, and partially online prognostic health management application in industries. *Applied Sciences*, *11*(8), 3380. doi: https://doi.org/10.3390/app11083380

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, *6*(1), 5. doi: https://doi.org/10.3390/data6010005

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2022, Jan). Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, *217*, 107961. doi: https://doi.org/10.1016/j.ress.2021.107961

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018, Sept). Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, *307*, 72-77. doi: https://doi.org/10.1016/j.neucom.2018.03.067

Coble, J. B. (2010). *Merging data sources to predict remaining useful life – an automated method to identify prognostic parameters* (Unpublished doctoral dissertation). University of Tennessee.

Coelho, L. B., Zhang, D., Ingelgem, Y. V., Steckelmacher, D., Nowé, A., & Terryn, H. (2022, Jan). Reviewing machine learning of corrosion prediction in a data-oriented perspective. *Materials Degradation*, *6*, 8. doi: https://doi.org/10.1038/s41529-022-00218-4

de Pater, I., Reijns, A., & Mitici, M. (2022, May). Alarm-based predictive maintenance scheduling for aircraft engines with imperfect remaining useful life prognostics. *Reliability Engineering & System Safety*, *221*, 108341. doi: https://doi.org/10.1016/j.ress.2022.108341

Eker, C. F., O.F., & Jennions, I. (2012). Major challenges in prognostics: Study on benchmarking prognostics datasets. In *Proceedings of the european conference of the phm society 2012* (Vol. 1, p. 1-8). doi: https://doi.org/10.36001/phme.2012.v1i1.1409

Elattar, H., Elminir, H., & Riad, A. e.-d. (2016, June). Prognostics: a literature review. *Complex Intelligent Systems*, *2*, 125-154.

Fernandes, M., Chorchaco, J. M., & Marreiros, G. (2022, Mar). Machine learning techniques applied to mechanical fault diagnosis and fault prognosis in the context of real industrial manufacturing use-cases: a systematic literature review. *Applied Intelligence*, *52*, 14246-14280. doi: https://doi.org/10.1007/s10489-022-03344-3

Ferreira, C., & Gonçalves, G. (2022, April). Remaining useful life prediction and challenges: A literature review on the use of machine learning methods. *Journal of Manufacturing Systems*, *63*, 550-562. doi: https://doi.org/10.1016/j.jmsy.2022.05.010

Guo, J., Li, Z., & Li, M. (2020). A review on prognostics methods for engineering systems. *IEEE Transactions on Reliability*, *69*(3), 1110-1129. doi: https://doi.org/10.1109/TR.2019.2957965

Gálvez, A., Diez-Olivan, A., Seneviratne, D., & Galar, D. (2021). Fault detection and rul estimation for railway hvac systems using a hybrid model-based approach. *Sustainability*, *13*(12), 6828. doi: https://doi.org/10.3390/su13126828

Gülich, J. F. (2020). *Centrifugal pumps*. Springer.

Heek, D. (2021). *A data-driven condition monitoring approach for the main bearings of a marine diesel engine.* https://research.tue.nl/en/studentTheses/a-data-driven-condition-monitoring-approach-for-the-main-bearings. (MSc thesis, Eindhoven University of Technology)

Homborg, A., Leon Morales, C., Tinga, T., de Wit, J., & Mol, J. (2014, Aug). Detection of microbiologically influenced corrosion by electrochemical noise transients. *Electrochimica Acta*, *136*, 223-232. doi: https://doi.org/10.1016/j.electacta.2014.05.102

Jouin, M., Gouriveau, R., Hissel, D., Péra, M.-C., & Zerhouni, N. (2016, May). Particle filter-based prognostics: Review, discussion and perspectives. *Mechanical Systems and Signal Processing*, *72-73*, 2-31. doi: https://doi.org/10.1016/j.ymssp.2015.11.008

Keizers, L. S., Loendersloot, R., & Tinga, T. (2021). Unscented kalman filtering for prognostics under varying operational and environmental conditions. *International Journal of Prognostics and Health Management*, *12*(2), 1-20. doi: https://doi.org/10.36001/ijphm.2021.v12i2.2943

Keizers, L. S., Loendersloot, R., & Tinga, T. (2022). Atmospheric corrosion prognostics using a particle filter. In *Book of extended abstracts for the 32nd european safety and reliability conference.* doi: https://doi.org/10.3850/978-981-18-5183-4_r22-08-170-cd

Kumar, S., Kolekar, T., Kotecha, K., Patil, S., & Bongale, A. (2022, Jan). Performance evaluation for tool wear prediction based on bi-directional, encoder–decoder and hybrid long short-term memory models. *International Journal of Quality Reliability Management*, *ahead-of-print*, 1551-1576. doi: https://doi.org/10.1108/IJQRM-08-2021-0291

Li, Y., Jiang, W., Zhang, G., & Shu, L. (2021, June). Wind turbine fault diagnosis based on transfer learning and convolutional autoencoder with small-scale data. *Renewable Energy*, *171*, 103-115. doi: https://doi.org/10.1016/j.renene.2021.01.143

Lu, S., Zhu, Y., Liu, S., & She, J. (2022). A tool wear prediction model based on attention mechanism and indrnn. In *2022 international joint conference on neural networks (ijcnn)* (p. 1-7). doi: https://doi.org/10.1109/IJCNN55064.2022.9889794

Lukens, S., Rousis, D., Baer, T., Lujan, M., & Smith, M. (2022). Data quality scorecard for assessing the suitability of asset condition data for prognostics modeling. In *Proceedings of the annual conference of the phm society* (Vol. 14, p. 1-15). doi: https://doi.org/10.36001/phmconf.2022.v14i1.3188

Medjaher, K., & Zerhouni, N. (2013, Jan). Framework for a hybrid prognostics. In (Vol. 33, p. 91-96). doi: https://doi.org/10.3303/CET1333016

Mkadara, G., Maré, J.-C., & Paulmann, G. (2021). Methodology for model architecting and failure simulation supported by bond-graphs—application to helicopter axial piston pump. *Sustainability*, *13*(4), 1863. doi: https://doi.org/10.3390/su13041863

Mulders, M., & Haarman, M. (2017). *Predictive maintenance 4.0 predict the unpredictable* (Tech. Rep.). https://www.pwc.nl/nl/assets/documents/pwc-predictive-maintenance-4-0.pdf. (Retrieved on 21-02-2024)

Nakhaeinejad, M., & Bryant, M. (2011). Dynamic modeling of rolling element bearings with surface contact defects using bond graphs. *Journal of Tribology*, *133*(1), 011102. doi: https://doi.org/10.1115/1.4003088

Peng, D., Liu, C., & Gryllias, K. (2022). A transfer learning-based rolling bearing fault diagnosis across machines. In *Annual conference of the phm society* (Vol. 14, p. 1-9). doi: https://doi.org/10.36001/phmconf.2022.v14i1.3257

Pinciroli, L., Baraldi, P., & Zio, E. (2023, June). Maintenance optimization in industry 4.0. *Reliability Engineering & System Safety*, *234*, 109204. doi: https://doi.org/10.1016/j.ress.2023.109204

Pugalenthi, K., Park, H., Hussain, S., & Raghavan, N. (2021, sept). Hybrid particle filter trained neural network for prognosis of lithium-ion batteries. *IEEE Access*, *9*, 135132-135143. doi: https://doi.org/10.1109/ACCESS.2021.3116264

Ramasso, E., & Saxena, A. (2014). Performance benchmarking and analysis of prognostic methods for cmapss datasets. *International Journal of Prognostics and Health Management*, *5*(2), 1-15. doi: https://doi.org/10.36001/ijphm.2014.v5i2.2236

Sawalhi, N., & Randall, R. (2008). Simulating gear and bearing interactions in the presence of faults: Part i. the combined gear bearing dynamic model and the simulation of localised bearing faults. *Mechanical Systems and Signal Processing*, *22*(8), 1924-1951. doi: https://doi.org/10.1016/j.ymssp.2007.12.001

Saxena, A., & Goebel, K. (2008). *C-mapss data set.*

Sheng, R., & Zhu, X. (2020, Dec). Tool wear assessment approach based on the neighborhood rough set model and nearest neighbor model. *Shock and Vibration*, *2020*, 1-15. doi: https://doi.org/10.1155/2020/8876187

Sun, J., Zuo, H., Wang, W., & Pecht, M. G. (2014). Prognostics uncertainty reduction by fusing on-line monitoring data based on a state-space-based degradation model. *Mechanical Systems and Signal Processing*, *45*(2), 396-407. doi: https://doi.org/10.1016/j.ymssp.2013.08.022

Sundararajan, G. (1991). A comprehensive model for the solid particle erosion of ductile materials. *Wear*, *149*(1), 111-127. doi: https://doi.org/10.1016/0043-1648(91)90368-5

Tiddens, W., Braaksma, J., & Tinga, T. (2023). Decision framework for predictive maintenance method selection. *Applied Sciences*, *13*(3), 2021. doi: https://doi.org/10.3390/app13032021

Tinga, T. (2013a, July). Predictive maintenance of military systems based on physical failure models. *Chemical engineering transactions*, *33*, 295-300. doi: https://doi.org/10.3303/CET1333050

Tinga, T. (2013b). *Principles of loads and failure mechanisms. applications in maintenance, reliability and design*. Springer.

Tinga, T., Wubben, F., Tiddens, W. W., Wortmann, H., & Gaalman, G. (2021). Dynamic maintenance based on functional usage profiles. , *27*(1), 21-42. doi: https://doi.org/10.1108/JQME-01-2019-0002

van der Velde, R., Moerkerken, A., Hofstraat, K., Rosier, M., Haarman, M., de Klerk, P., ... Nedelcheva, Y. (2023). *Digital trends in maintenance* (Tech. Rep.). `https://www.pwc.nl/en/evenementen/digital-trends-in-maintenance.html`. (Retrieved on 01-03-2024)

Virkler, D., Hillberry, B., & Goel, P. (1979). The statistical nature of fatigue crack propagation. *Journal of Engineering Materials and Technology*, *101*(2), 148-153.

Vos, P. (2019). *Engine condition trend monitoring for apache turboshaft engines* (Tech. Rep.). NLR. (Classified)

Wang, J., Wang, P., & Gao, R. (2015, July). Particle filter for tool wear prediction. *Journal of Manufacturing Systems*, *36*, 35-45. doi: https://doi.org/10.1016/j.jmsy.2015.03.005

Wang, P., & Gao, R. (2016). Stochastic tool wear prediction for sustainable manufacturing. *Procedia CIRP*, *48*, 236-241. doi: https://doi.org/10.1016/j.procir.2016.03.101

Yan, J., Meng, Y., Lu, L., & Guo, C. (2017). Big-data-driven based intelligent prognostics scheme in industry 4.0 environment. In *2017 prognostics and system health management conference (phm-harbin)* (p. 1-5). doi: https://doi.org/10.1109/PHM.2017.8079310

Zhou, Y., & Sun, W. (2020, May). Tool wear condition monitoring in milling process based on current sensors. *IEEE Access*, *8*, 95491-95502. doi: https://doi.org/10.1109/ACCESS.2020.2995586

Zio, E., & Peloni, G. (2011). Particle filtering prognostic estimation of the remaining useful life of nonlinear components. *Reliability Engineering & System Safety*, *96*(3), 403-409. doi: https://doi.org/10.1016/j.ress.2010.08.009

BIOGRAPHIES

Luc S. Keizers received his MSc degree in Mechanical Engineering at the University of Twente in 2020. He graduated on "Structural Fatigue Analysis using Flexible Multibody Dynamics" in the research chair of Applied Mechanics and Data Analysis. Shortly after, he started his PhD at the same university in the chair of Dynamics Based Maintenance in collaboration with the Netherlands Royal Navy. His research interest is in prognostics that combine his engineering background with data-driven models.

Richard Loendersloot received a MSc degree in Mechanical Engineering (2001) and a did his PhD research at the University of Twente, on thermoset resin flow processes through textile reinforcements during composite production process Resin Transfer Moulding and obtained his PhD degree in 2006. He worked in an engineering office on high-end FE simulations of a variety mechanical problems to return to the University of Twente in 2008 as assistant professor for Applied Mechanics. His research started to focus on vibration based structural health and condition monitoring, being addressed in both research and education. He became part of the research chair Dynamics Based Maintenance upon its initiation in 2012. His research covers a broad range of applications: from rail infra structure monitoring, to water mains condition inspection and aerospace health monitoring applications, using both structural dynamics and ultrasound methods. He is involved in a number of European and National funded research projects. He became associate professor in 2019, currently in charge of the daily lead of the Dynamics Based Maintenance group.

Tiedo Tinga is a full professor in dynamics based maintenance at the University of Twente since 2012 and full professor Life Cycle Management at the Netherlands Defence Academy since 2016. He received his PhD degree in mechanics of materials from Eindhoven University in 2009. He is chairing the smart maintenance knowledge center and leads a number of research projects on developing predictive maintenance concepts, mainly based on physics of failure models, but also following data-driven approaches.