Remaining Useful Life Estimation of Bearings: Meta-analysis of Experimental Procedure

Hugo M. Ferreira [1], Alexandre C. de Sousa[2]

[1,2] *INESC TEC, FEUP Campus, Rua Dr. Roberto Frias, Porto, Portugal*
hmf@inesctec.pt
alexandre.c.sousa@inesctec.pt

Abstract

In the domain of predictive maintenance, when trying to replicate and compare research in remaining useful life estimation (RUL), several inconsistencies and errors were identified in the experimental methodology used by various researchers. This makes the replication and the comparison of results difficult, thus severely hindering both progress in this research domain and its practical application to industry. We survey the literature to evaluate the experimental procedures that were used, and identify the most common errors and omission in both experimental procedures and reporting.

A total of 70 papers on RUL were audited. From this meta-analysis we estimate that approximately 11% of the papers present work that will allow for replication and comparison. Surprisingly, only about 24.3% (17 of the 70 articles) compared their results with previous work. Of the remaining work, 41.4% generated and compared several models of their own and, somewhat unsettling, 31.4% of the researchers made no comparison whatsoever. The remaining 2.9% did not use the same data set for comparisons. The results of this study were also aggregated into 3 categories: problem class selection, model fitting best practices and evaluation best practices. We conclude that model evaluation is the most problematic one.

The main contribution of the article is a proposal of an experimental protocol and several recommendations that specifically target model evaluation. Adherence to this protocol should substantially facilitate the research and application of RUL prediction models. The goals are to promote the collaboration between scholars and practitioners alike and advance the research in this domain.

1. Introduction

When trying to replicate and compare research in remaining useful life (RUL) estimation in the domain of predictive maintenance, we found that the experimental protocol used by various researchers varied significantly. Many details were not provided making replication difficult or even impossible. Most importantly, several problems in the experimental methodology were detected, making valid comparisons with existing work difficult if not impossible. Our goal is to establish a common baseline that will facilitate researchers' work in the future, allowing for consistent and reproducible comparison of existing work. This is of paramount importance because without the ability to replicate and compare results, it is not possible to make progress (Munafò et al., 2017).

We review the literature on RUL estimation of bearing failure due to wear and tear. The aim is to identify the main domain specific characteristics that are relevant to generating machine learning (ML) models. We study the general literature on ML in order to identify common pitfalls and establish a correct procedure for data analysis and model comparison. We also delve into the domain specific issues that determine how model selection, evaluation and comparison should be done. This includes details on how the signals are segmented, if and how these signal segments are categorized into degradation phases and how these signals are used for training and evaluation.

2. Background

Researchers have concluded that many experimental results are unreliable (Shepperd et al., 2019; Ioannidis, 2005). Errors have been found, some of which may be attributed to simple transcription errors. This is in part due to the complex and chaotic process that includes data pre-processing, feature generation and selection, hyper-parameter tuning via cross-validation, metric selection and model performance comparison (Shepperd et al., 2019).

Several researchers concerned with this issue have done meta-analysis on related work to try and identify what errors are committed and how one may go about avoiding these. This includes studying the validity of null hypothesis significance testing (Colquhoun, 2018), checking simple integrity constraints (arithmetical and statistical errors) (Shepperd et al., 2019) and examining the correctness of the experimental design and analysis (Ioannidis, 2005). In this work we will look at the machine learning experimental design in RUL estimation with an emphasis on the comparative analysis of the generated prediction models.

Prior work has already identified a lack of standard approaches to comparing prognostics models (Saxena, Celaya, Saha, Saha, & Goebel, 2010). The main goal was to establish a way of rigorously evaluating the performance of prognostics systems so that they can be certified for use in critical applications. References to research are also provided that showed a lack of standardized methodologies of model comparison or even absence of model evaluation (Saxena et al., 2010). The focus was on the use of metrics for off-line RUL estimation when run to failure data exists (Saxena, Celaya, Saha, Saha, & Goebe, 2009).

We make the following observations. The first is that in our work, we focus on the experimental methodology that includes problem selection, model fitting and model comparison. Second, we propose a method for performance evaluation. Any formal metric that allows for rigorous comparisons can be used use within our proposed framework. Third, our goal is to facilitate the comparison of researchers work and so do not delve into issues of prognostics specifications such as performance requirements (cost) and risk management (safety, reliability) (Saxena et al., 2010). Finally, we only consider RUL estimation based on run to failure data. Our literature review of the more recent work shows that errors and omissions still plague the research in this domain and these problems are not limited to the use of metrics.

3. REVIEW AND ANALYSIS OF RUL LITERATURE

A systematic search of the literature in the domain of RUL estimation was made. Specifically we selected articles that analyze and compare the performance of various RUL models. Additional criteria were used in the initial selection of the article (see Table 1).

The goal is to evaluate the level of adherence to proper procedure, identify the most common problems and estimate how prevalent these issues are. We have split the analysis into two main group: general protocol issues that are applicable to any domain (section 6) and protocol issues that are specific to the RUL estimation domain in predictive maintenance (section 7). Next we describe the methodology use in this meta-analysis.

3.1. Methodology

The initial sample of articles were identified, in May of 2020, using the Scopus, Engineering Village and Web of Science knowledge databases. The query included the key phrases “RUL”, “remaining useful life”, “bearing”, “prognostics”, “predictive maintenance” and “condition based maintenance”. Additional criteria included the selection of articles written in English published after the year 2000 (inclusive). This resulted in a total of 585 article. Full details on the selection criteria and the articles can be found in the supplementary material[1].

During a second phase, the digital object identifier (DOI, when available) and the titles (which were converted to lower case), were used to remove all duplicates. This resulted in 328 unique articles. A total of 177 of these articles were randomly sampled and a cursory analysis removed any articles that: did not have a DOI; were not applied research papers on prognostics for bearings (excluded reviews and books); did not make the data publicly available (excluded references claiming that data is provided upon request) and the publisher was not on Beall’s list. This produced a list of 70 articles that were analyzed in detail (full details found in the supplementary material).

A set of 21 indicators (metrics) were used to record the availability of data, the type of RUL estimate performed (section 8.1), and the model fitting (section 8.2) and model evaluation (section 8.3) procedures that were performed. The data is available as a spreadsheet included in the supplementary material. There are cases when certain features cannot be clearly determined. When in doubt, we assumed that procedures did not follow proper protocol. We therefore cautiously err on the side of underestimating compliance. Nevertheless, the conservative estimates still allow us to evaluate the current state of research.

4. RUL ESTIMATION OF BEARINGS

In this section, we briefly describe several concepts and ideas that are required to understand and fully appreciate the issues that we discuss the in following sections. First is the use of ML regression models for RUL estimation, which includes data acquisition and feature engineering. The second is the notion that bearings go through several stages of degradation and how these are related to the RUL estimate. We refer to these phases as health stages. Last, we consider the problem of RUL estimation of equipment that operate in multiple regimes and the challenge this presents.

4.1. ML Models for RUL Estimation

Rotating machinery have shafts or axles that are supported by rolling bearings. The bearing elements and the enclosing

[1] https://zenodo.org/record/3972767

Table 1. Selection Criteria of relevant articles.

Criterion	Description
Language	English
Topic	Data driven RUL estimation for predictive maintenance.
Availability	All articles and their data sets must be available to the public.
Date	January 2000 – May 2020
Reviewed	All papers must go through a peer review process (at the very least they must not be in Beall's List (Machacek & Srholec, 2019) of potential predatory journals and publishers).
Duplicates	Use only the latest peer reviewed version.
Reporting	Sufficient detail to analyze the degree of adherence to proper computational experiment procedure in ML.

(inner and outer) race suffer wear and tear through use and corrosion resulting in cracks, brinelling, spalling and fretting.

Generally, it is not possible to assess the damage of these components by directly observing the degree of wear and tear (Lei et al., 2018). This may be because: machines cannot be shutdown for inspection, it is difficult to detect and measure micro-scale defects at the incipient stage, or, to access and observe the damaged parts, may require that the full assembly be destroyed (e.g. sealed roller bearings). As a result, one or more sensors are used to indirectly measure the health status of the components. Usually accelerometers are used to measure the vibration of rotating machinery. However temperature, audio and ultrasound signals may also be used. All the articles we have analysed use the accelerometer data, however the protocol issues we discuss are independent of the type and number of sensors used.

The sensor readings may be converted to an intermediate set of features. Several features may be combined into a single generic feature that reflects the health status of the component. It is usually referred to as the health indicator. These may be features from the time domain, frequency domain (Xia et al., 2019; Li, Zhang, & Ding, 2019) or a combination of both (Benkedjouh, Medjaher, Zerhouni, & Rechak, 2013; Sutrisno, Oh, Vasan, & Pecht, 2012). Some researchers forego feature engineering altogether and use the raw signal directly (Khelif et al., 2017; C. Liu, Zhang, & Wu, 2019; Verstraete, Droguett, & Modarres, 2019; Jiang, Lee, & Zeng, 2019; Zhang, Hutchinson, Lieven, & Nunez-Yanez, 2020; B. Wang, Lei, Yan, Li, & Guo, 2020). Whether the signal is converted or not, they are then used by a ML model to detect or estimate the level of degradation. In the case of RUL estimation we use a regression model that predicts how long the components may still function properly. As with the type of sensors, the protocol issues we study are independent of the number and type of features used.

4.2. Bearing Health Stages

We assume that the degradation process is stochastic, noisy and irreversible (Lei et al., 2018). One would therefore expect that the ML model generate a monotonically decreasing RUL estimate. However, bearing wear and tear exhibit complex patterns of degradation (Lei et al., 2018) due to the *self-healing* phenomena (Duong et al., 2018). More concretely, an incipient fault caused by cracks will result in vibrations of small amplitudes. With time, the surface defect worsens resulting in an increase in the vibration. However, the continued friction of the bearing may end up smoothing the sharp edges of the crack, thereby temporarily reducing the vibration. With continued use, damage will eventually spread over a broader area, and the vibration amplitude will rise again (Williams, Ribadeneira, Billington, & Kurfess, 2001). We also refer to these degradation phases as health stages.

A single health indicator may be used to represent several health stages of several components (Lei et al., 2018). First we note that, in general and depending on the type of faults and features used, the number of health stages may differ. Second, in many cases it is difficult to identify these stages and determine exactly when they start and end (Sutrisno et al., 2012). Consequently, a variety of solutions have been proposed. Some authors suggest dividing the signal into a fixed and constant number of stages. For example two stages (T. Wang, 2012; Li et al., 2019; B. Wang, Lei, Li, & Li, 2020; Mao, He, Tang, & Li, 2018) and three stages (Z. Liu, Zuo, & Qin, 2015; Soualhi, Medjaher, & Zerhouni, 2015). Others attempt to determine these automatically using, for example, heuristics (Sutrisno et al., 2012), classification (Xia et al., 2019) or the Minimum Description Length (MDL) principle (Peng, Cheng, Liu, Li, & Peng, 2018).

This has proven to be a difficult issue when designing an appropriate protocol. We must decide whether or not signals should be divided into stages and if so how this should it be done. Stage division has several important consequences. The first issue is the incompatibility of model comparisons. RUL estimations made only for a specific degradation stage will naturally outperform a more general model trained on the full length of the signals. In section 8.1 we will delve more into this issue.

The second is model performance comparison. RUL estimates vary significantly depending on the degradation stage of the bearings. This means that models with the same average performance may incur very different errors at different stages of the bearings' life cycle. It is important that the models' performance be compared at the various phases of the components' life-cycle. Model accuracy is especially critical toward the end-of-life (EoL) stage of the component. In fact we question the usefulness of providing a RUL estimate before an initial fault event has been detected. These issues are discussed further in section 7.1.

Lastly, we must consider the practical applicability of stage division. We defend in later sections that stage division should exist, should be done automatically and should only activate the RUL estimation model after the initial failure event has been detected (see section 8.3).

4.3. Multiple Operating Regimes

Rotating machinery may operate under different loads and speeds. Ideally, the RUL estimation models should work for all operating regimes. However, this may not be viable because intrinsic signal characteristics (such as noise), vary according to the operating mode (T. Wang, 2012). Additionally, changes in the frequency and amplitude of the signal, caused by higher speeds or higher loads, may not be easily distinguished from anomalies (Heng, Zhang, Tan, & Mathew, 2009; Lei et al., 2018).

One simplification is to use a single model for each operating mode (T. Wang, 2012). However this may not be feasible in practice, especially if the number of operating regimes is large or continuous. An alternate solution is to provide information on the operating mode to the RUL model. But this information may not always be available (Kan, Tan, & Mathew, 2015). These issues pose problems for both the practical application of theses solution in real-life settings, as well as the general evaluation and comparison of RUL models.

In section 7.3 we have a more general discussion related to these issue. In section 8.1 we present a set of RUL model classes and detail how these may be compared. Ultimately we leave it to the researcher to decide what type of model to use, but impose certain restriction on how models may be compared.

5. BENCHMARK DATA SETS

In this section we review the data sets that are used to compare the ML models that make the RUL estimate of bearings. As per the selection criteria, we only analyze research that use public data sets that contain real sensor data (no synthetic data). This ensures that in the future, researchers are free to replicate prior work and compare results. Obviously we want real sensor data only, because the goal is to produce results that are applicable to real world scenarios.

To better understand how the data sets are used as benchmarks, the following was done: for each article, we identified which data sets were used. We count the number of times a data set is used by itself and the number of times it is used in conjunction with another data set. The sum of these 2 values allow us to estimate the distribution of these data set. This was done for the initial sample of 177 articles (referred to as *sampled*) and then again for the final 70 articles that respected all selection criteria described in Section 3.1 (referred to as *filtered*). The results are shown in Table 2 (data available in the supplementary material). Note that the category *others* includes all articles that did not pass all filtering criteria.

Table 2. Distribution of Benchmarks.

Data set	**Sampled (%)**	**Filtered (%)**
Pronostia	27.7	67.1
IMS	10.7	24.3
GPMS	1.7	4.3
XJTU	0.6	1.4
IMS+Pronostia	2.3	1.4
XJTU+Pronostia	0.6	1.4
Private	21.5	0
Others	35.0	0

We found the bearing data sets are limited to the following 4 public benchmarks: Pronostia (Nectoux et al., 2012), IMS (Qiu, Lee, Lin, & Yu, 2006), GPMS (Ben Ali, Saidi, Harrath, Bechhoefer, & Benbouzid, 2018) and XJTU-SY (Shan et al., 2019). More importantly the Pronostia and IMS data sets are used in the majority of the work. This is true for both the sampled (41.2%) and filtered (94.3%) cases. The reduced number of data sets is an issue because it does not allow researchers to thoroughly test the applicability of their RUL models to many real world scenarios. This lack of diversity is specifically problematic if we consider that the Pronostia, IMS and XJTU-SY data sets were generated in laboratory, under ideal accelerated degradation conditions.

We also looked at the characteristics of the Pronostia, IMS, GPMS and XJTU-SY data sets in order to identify any issues that need to be considered when performing RUL estimation. All of these data sets contain accelerometer data that are collected from bearings until a failure occurs. However, we found that only the GPMS data set, was acquired in a real world setting - from shaft bearings installed in a wind turbine generator. Unfortunately this data set consists of a single failure instance - so a test set is not available for model evaluation.

Another characteristic, is the rate of degradation (see Table 3). The GPMS, with 50 days, reflects the expected rates in a real, albeit harsh, world setting. However, in all other cases the laboratory experiments induced accelerated degradation. In particular, Pronostia, which is the most popular data set, has signals with a duration that vary from as little as 38 minutes to 7 hours and 45 minutes. The XJTU also exhibits very short run-to-failure experiments in the same order of magnitude (see Table 3). This raises issues of whether or not the models will still perform well when applied to very different degradation profiles. It also makes model evaluation problematic. The reason is that the RUL estimation error decreases towards the EoL event (see for example (Li et al.,

Table 3. Duration of Benchmark Signals.

Data set	Minimum	Maximum	Median
Pronostia	38 min	7 hours 47 min	3 hours 10 min
XJTU	42 min	42 hours 18 min	2 hours 41 min
IMS	6 days 9 hours	44 days 17 hours	34 days 12 hours
GPMS	50 days	50 days	50 days

Table 4. Number of Instances in Benchmarks.

Data (#Modes)	$\sum$ Train	$\sum$ Test	Train Med	Test Med
Pronostia (3)	6	11	2	5
XJTU (3)	15	0	5	0
IMS (1)	3	0	1	0
GPMS (∞)	1	1	0	0

2019; Y. Wang, Peng, Zi, Jin, & Tsui, 2015; Li, Zhang, Ma, Luo, & Li, 2020; Xia et al., 2019)). So longer signals usually result in larger errors. Depending on the time-series that are used for the testing data set, significant differences in error will occur. To avoid this problem, in this work we defend that the signal should be divided into health stages. Model performance should be evaluated separately in each stage using a fixed number of residuals. Additionally, we question whether or not it makes sense to perform RUL estimation in the initial stage when no degradation has occurred.

The size of the data set - the number of time-series in the data set - is an important feature. For the best results one would have enough data for the training, cross-validation and test data sets. Table 4 shows the number of time-series in the training and test data sets ($\sum$) and the corresponding medians (Med) per operating mode (different load and rotation speed). The Pronostia and XJTU data sets have 3 operating modes, the IMS data set has 1 and the GPMS's load and rotation speed vary continuously. We see that in all cases we have a limited number of examples (a maximum total of 15 irrespective of operating mode). Due to the lack of data, none of it has been reserved for model fine-tuning, which makes cross-validation impossible (see Section 6.1). We can also expect that, because of the very low number of instances per operating mode (maximum 5), training and evaluating models for each operating mode is susceptible to over-fitting. Finally, we note the Pronostia test data sets are truncated (because it was used for a competition, nevertheless the RUL times are available). This makes model evaluation at the later life-cycle stage impossible.

It is imperative that much larger real-world data sets, with several hundreds of complete run-to-failure instances, be made available for future research efforts. We also believe that such data sets should have a much larger range of operating modes (similar to GPMS). And finally, we should have enough data sets to cover several classes of machinery (lathes, drilling, generators, motors) in order to determine the effectiveness of RUL estimation in different scenarios.

In regards to the reduced number of operating modes (#Modes in Table 4), care must also be taken when evaluating the models. Models that are trained for a single or multiple operating modes should not be compared indiscriminately. Details on what constitute valid comparisons are detailed in section 7.

6. GENERAL PROTOCOL ISSUES

In ML, the goal is to achieve good generalization performance, i.e. incur low prediction errors on previously unseen data (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009; Russell & Norvig, 2010; P. Murphy, 2012; Goodfellow, Bengio, & Courville, 2016). In the following subsections, we discuss some methodological errors in experiment execution and performance evaluation that lead to overly optimistic generalization errors being reported (errors are much higher on real-world unseen data than on the test data). In addition to describing the most common errors in ML experimental procedures, we also provide a number of recommendations that may help avoid or at least mitigate these problems.

6.1. Underfitting and Overfitting

The performance of a machine learning algorithm is measured by its ability to simultaneously minimize the training and test errors. We say that there is underfitting when the model is not able to yield a low training error, and there is overfitting when the training error is low but the test error is large (Goodfellow et al., 2016). We control the propensity to either overfit or underfit by adjusting the model capacity or complexity (Hastie et al., 2009; Goodfellow et al., 2016). This is known as the bias-variance trade-off and is the central challenge of ML (Goodfellow et al., 2016).

Hyper-parameters control model complexity. They are optimized using a validation set (Bishop, 2006; Russell & Norvig, 2010; Goodfellow et al., 2016), on which we aim to find a good trade-off between model complexity and goodness of fit (Russell & Norvig, 2010). If we were to optimize the hyper-parameters using the training set, it would always result in maximum possible model capacity, i.e. overfitting (Goodfellow et al., 2016). As such, we use the validation set to estimate the generalization error according to the selected hyper-parameter values (Goodfellow et al., 2016). It provides a reasonable estimate of the expected test error (Hastie et al., 2009). Because the test data set can only be used in the final model evaluation, the validation set is built by partitioning the training set (Bishop, 2006; Russell & Norvig, 2010; P. Murphy, 2012; Goodfellow et al., 2016). Note that if the model design requires many iterations to tune the hyper-parameters, it may also overfit to the validation set (Bishop, 2006).

Recommendations:

1. Use cross-validation as a way of making a reasonable trade-off between model bias and complexity (Bishop, 2006; Hastie et al., 2009; Russell & Norvig, 2010; Good-

fellow et al., 2016).

2. Create a validation set by partitioning the training dataset only (Bishop, 2006; Russell & Norvig, 2010; P. Murphy, 2012; Goodfellow et al., 2016). Make sure the data labels are balanced by using for example *stratified data partitioning* (see (Moreno-Torres, Saez, & Herrera, 2012) for details).
3. If overfitting is a problem, increase the size of the dataset in order to promote model generalization and robustness (Bishop, 2006).

6.2. Data leaking

Data leaking occurs when information contained in the test or validation sets are used to train the model, thereby invalidating the results (Russell & Norvig, 2010). In these cases the model will overfit, and as such, the best model will underestimate the true prediction error, sometimes by a substantial amount (Hastie et al., 2009; P. Murphy, 2012)

In the worst case scenario, part or all of the test and validation data records may be directly introduced into the training set. However, information may also be inadvertently leaked into the training set during the initial phases of data preparation (Kaufman, Rosset, & Perlich, 2011). We next describe how information may be indirectly leaked in the various phases of model fitting and evaluation (Kaufman et al., 2011).

During data collection, data may be leaked by providing time-dependent information during model training and evaluation that will not be available during the prediction time in a real system. Another issue is when seemingly unrelated features may be used to infer the dependent variable's outcome. For example, a feature may consistently encode or re-scale values of the dependent variable.

Feature engineering may also result in information leaked if not applied correctly. For example scaling must be done independently on the validation and test sets using the same mean and standard deviation obtained from the training data set. All calculated features must be recomputed for the validation and test data sets.

Partitioning data must be done with care. When data is sampled or augmented prior to data splitting data leakage may occur. This will result in closely related data being shared among training and test data sets. This is particularly tricky when dealing with time-dependent data. For example, signals of the same entity at different states (instances) may be placed in both training and test data sets.

Recommendations:

1. Split the data sets (train, validation and test) before model training and evaluation. Make sure no data is shared among the partitioned data sets. (Hastie et al., 2009; Russell & Norvig, 2010; P. Murphy, 2012; Goodfellow et al., 2016).
2. Check if any co-linearity exists between features and the dependent variable. Make sure no feature can be calculated using one or more of the other features.
3. Make sure that all features in the training data set are available at prediction time only.
4. Do not use any external data to indirectly infer the dependent variable.
5. Determine all pre-processing parameters using only the training data set. Apply those when pre-processing the validation and test data sets.

6.3. Data snooping

Data snooping (data dredging) is the deliberate selection (cherry-picking) of samples to produce the expected results. It results in assigning meaning to spurious patterns (Giles & Lawrence, 1997), or letting the test set directly influence the training process (England & Cheng, 2019). It may also indirectly bias the hypothesis (hypothesizing after results) (Murphy, 2017).

This process involves testing multiple hypothesis using the same data by means of exhaustive search - what some call "oversearching" (Jensen, 2000). Over-searching has a parallel to p-hacking in statistical significance testing, i.e. run a sufficiently large amount of experiments and report only the best results (Giles & Lawrence, 1997; Russell & Norvig, 2010; England & Cheng, 2019). Due to the large search space, finding a model with good performance may be attributed solely to chance (Giles & Lawrence, 1997). The use of the expression *"we optimized the learning rate"*, used to summarize research results, may be indicative of data snooping (Giles & Lawrence, 1997).

By the same token, the effects of data snooping might result from the collective efforts of a research community at large. By using the same test data repeatedly to evaluate the performance of different techniques over many experiments (e.g. the use of common benchmark datasets), we end up with overly optimistic evaluations. Collectively "over-searching" makes benchmarks stale, and the reported results do not reflect the true performance in the field of study (Goodfellow et al., 2016).

Recommendations:

1. Provide information on the experimental process (Giles & Lawrence, 1997; England & Cheng, 2019), namely the hypothesis and model selection criteria should be made explicit (England & Cheng, 2019). Be aware of potential data snooping biases when formulating the experimental procedures (Giles & Lawrence, 1997).
2. If model fine-tuning is required or desired, optimization must be performed by using the validation set only (sec-

tion 6.1) (England & Cheng, 2019). Use cross-validation to *algorithmically* identify the "best" model (Jensen, 2000).

3. Consider the size of the model space searched to support the research hypothesis, namely by correcting for the statistical effects of the search (Jensen, 2000).
4. Use the same model on different data sets and report all results. Good performance on a variety of data sets is indicative of a robust model that is not overfitting.

6.4. Model Performance Averaging

Data averaging is the presentation of the results of the multiple experiments using descriptive statistics instead of presenting the results individually. This practice, in itself, is not incorrect. However, usually defaulting to reporting a single mean and standard deviation value can be misleading or lead to incorrect conclusions (Giles & Lawrence, 1997).

When comparing models, the use of descriptive statistics may not be enough. More concretely a single aggregate may hide important information regarding the model's performance under various conditions. For example, it is difficult to convey information of a RUL model's performance in the various health stages using a single value. In these cases performance values should be provided for the various health stages.

Recommendations:

1. In addition to the mean and standard deviation, report more informative statistics regarding the distribution - e.g. median, inter-quartile range, minimum and maximum (Giles & Lawrence, 1997).
2. Plot the results. For example, the box-whisker plot is suitable for comparing models and identifying outliers. Plots may be used to show how robust predictions are. Plots can also be used to show time or state dependent performance.
3. If experiments are executed in significantly different conditions, report all the results. Use aggregates, statistics (such as Skewnes and Kurtosis) and plots to facilitate comparisons.

7. DOMAIN PROTOCOL ISSUES

In this section we analyse some of the methodological issues that are specific to the RUL estimation domain. We look at the potential problems that may occur during model fitting, model evaluation and with the use of data generated under different experimental conditions.

7.1. Model Fitting and Prediction

Many research papers report on the use of novel models for RUL estimation. In their work, not only do the models differ, but so do the conditions under which these models are trained and evaluated. Concretely, several researchers use a model to estimate the RUL throughout the components full lifetime. Others, only fit and estimate the RUL during the last phase of the bearings' life-cycle - usually after an initial fault event has occurred (Z. Liu et al., 2015; Lei et al., 2018). The time at which this event occurs is defined as the First Predicting Time (FPT) (Lei et al., 2018; Li et al., 2019) and marks the start of second stage of bearings life-cycle (section 4.2). In addition to this, the identification of the fault event may be done manually, based on some criteria (such as a sensor reading threshold) or even automatically using anomaly detection methods (Lei et al., 2018; Li et al., 2019).

Initial RUL estimates will incur larger errors (section 4.3, (Saxena et al., 2010)); so comparing against models that only make predictions at later FPTs, is unfair. Researchers must provide details on how models are trained and evaluated. When comparing results, the conditions under which model fitting and evaluation is done, must be the same. Alternatively automatic FPT detection can be combined with metrics that are designed to set the comparisons on an equal footing. The advantage of this solution is that the FPT may be determined automatically, which is important when deploying RUL models in a real-world setting (section 7.2).

Recommendations:

1. Researchers should clearly state if the training and evaluation is performed for the full length of the signal. If signal evaluation is done after the FPT, then for each train, evaluation and test signal, the FPT timestamp must be provided.
2. If the FPT is selected manually, then the comparison must only be with models that are trained and evaluated after that FPT.
3. If an anomaly detection task is used to identify the FPT, then for the comparison, an anomaly detection model (not necessarily the same one used by the work that is being compared to), must first identify and flag the FPT. The RUL model must only be trained and then evaluated after that FPT (section 7.2).

7.2. Evaluation Metrics

There is no standardized methodology for the evaluation of prognostic performance (Saxena et al., 2010). This significantly hinders comparison efforts among researchers. RUL estimation is specifically challenging because the accuracy of the predictions become critical as the system nears its EoL (Saxena et al., 2010). Additionally, RUL estimates are inaccurate at the start of life of a component and tend to converge towards the EoL. These issues must be taken into account when comparing RUL performance. This includes penalizing models that initiate predictions late (for example weighing the RUL estimates according to the health stages or FPT), reporting on the model's performance for each health stage and providing aggregate performance values using additional statistics such a deviation and plots (section 6.4).

Recommendations:

1. Penalize RUL predictions differently according to the actual true EoL time (Nectoux et al., 2012). Weigh underestimates less than overestimates relative to the EoL. Use the same weights in all comparisons.
2. Measure the error relative to the EoL (Saxena et al., 2010) because the life-span of the same type of ball bearings vary widely (Vlcek, Hendricks, & Zaretsky, 2003).
3. Measure performance for a fixed number of timestamps or at fixed intervals of time before the EoL(Saxena et al., 2010).

7.3. Dealing with Different Experimental Set-ups

Experiments have different set-ups, which include: type and geometry of component to be analyzed, the means of accelerating the degradation, the operating modes of the equipment under test, the sensor location, sampling rates used and data collection methodology. Additionally, sensors and data equipment with different characteristics are used (sensor range, sensor sensitivity, analog to digital conversion resolution, accuracy and sampling rate).

Assuming that experiments have been set-up correctly and that their are no issues with low sampling rates (Taylor, 1997; Measurement Computing Corporation, 2004) or spurious noise (Bozchalooi & Liang, 2008; Dong & Chen, 2012; Soualhi et al., 2015) (which complicates the early detection of incipient faults (Bozchalooi & Liang, 2008; Dong & Chen, 2012)), the differences referred to above must be taken into account when comparing models. More specifically, different RUL estimation methods have various advantages depending on the signal characteristics and data acquisition conditions. So models should be trained and evaluated separately for each data set.

We have seen that several data sets include sensor data for one or more operating regimes (section 4.3). Models trained and evaluated on a single operating mode will usually outperform a more general (biased) model. As such, one must only compare the more general (multi-mode) models against their more specific counterparts.

Recommendations:

1. Train, evaluate and compare the proposed technique on each experiment's data separately. If different failure modes (for example inner race outer race and ball bearing damage in bearings) and/or multiple operation modes (load and velocity) are available, evaluate these separately also. We refer to these as single-mode models.
2. Additionally, if different failure modes and/or multiple operation modes are available, one can also train and evaluate the RUL estimation models using all multiple failure and operation modes. We refer to these as multi-mode models.
3. Aggregate the results to provide a summary of the performance so that one can determine if the method has, on average, better performance for a specific single-mode or multi-mode problem. Use formal methods to compare the result (Shepperd et al., 2019) (for example the Bonferroni correction, Benjamini-Hochberg false discovery rate estimate and Nemenyi post hoc procedure).

8. PROPOSED PROTOCOL

According to the discussion in the previous sections, we propose a set of protocols that should allow researchers to make valid comparison with prior work. We select a subset of the recommendations above and present them as a single consistent process. We also consider issues such as reproducibility and evaluation fairness. However, we do not look into the more general issues of methods, reporting, dissemination and incentives (Munafò et al., 2017).

8.1. Problem classes

Different prediction problems present very different challenges (section 7.3). As we have seen, it is easier to generate models that estimate the RUL for a single operating regime than it is for the case of multiple operating conditions. It is therefore imperative that we only compare against models that work under the same or more restricted conditions. Under no circumstances can a set of single-mode RUL models be favourably compared against a multi-mode model.

The same rational applies to models that are trained and evaluated after a FPT (section 7.1). The more specific models that estimated the RUL for the last phase of the components' lifespan cannot be favourably compared against a model that estimates the RUL for the full duration of the experiment.

Several researchers establish the FPT manually according to a subjective criterion (Zheng, 2019; Cheng, 2017). Determining the first fault event manually has several important disadvantages. First, it makes comparisons with new benchmarks difficult and arbitrary. Second, it makes the use of the model impractical in a real world scenario. Finally this form of establishing a FPT is prone to data snooping. We believe the FPT must always be identified automatically. The proposed protocol ensures that such an FPT is used correctly for model training and evaluation.

According to the review and the analysis above, a list of compatible experimental types has been identified according to the RUL estimation time-span (full signal or after an FPT), whether the FPT is detected automatically and if the RUL predictions are made for single or multiple operating regimes (see Table 5):

Experiments that use manual FPT (class ID 3 and 5) should

Table 5. RUL model classes.

ID	Lifetime	Lifetime	Op. Regimes	Comparable To
1	full	-	all	2,4,6
2	full	-	single	6
3	partial	manual	all	-
4	partial	automatic	all	6
5	partial	manual	single	-
6	partial	automatic	single	-

never be used. Model classes 1, 2, 4 and 6 are valid. The most general model class 1 can be compared against all others. Model classes 2 and 4 can only be compared to the more specific model classes 6. Finally, models classes that are trained and predict only after a FPT, must establish this event automatically. For example, using anomaly detection techniques. Evaluation will require conditionally performing the RUL estimation only after the FPT event has been automatically detected.

It is important to note that we have adhered to the standard practice in this domain of generating and evaluating one model per benchmark (section 8.2) irrespective of the failure types (bearing, inner or outer race).

8.2. Model fitting

Correct protocol requires that the problem class be selected and clearly stated (see section 8.1). Each model must be trained under the exact same conditions. A model may only be trained on data sets that were generated under the same conditions (sampling rate, sensor characteristics, degrading component). If one model is trained for each benchmark, then the models and corresponding training and test benchmarks must be identified. Same applies if a model is generated for each of the operational condition - the model's training and test data instances must also be indicated. To avoid leakage, test data set instances must not, under any circumstance, be used for training (see sections 8.3 and 6.2). If many models are generated during hyper-parameter tweaking, then an appropriate method must be used to select the best model (Shepperd et al., 2019) (section 7.3), or alternately, report on the robustness of the models (using for example sensitivity analysis (Cortez & Embrechts, 2013) or at the very least simple aggregates such as median, inter-quartile range, minimum and maximum (Giles & Lawrence, 1997). This avoids a form of data snooping where the selected model has a particularly high performance that is due to chance alone (6.3).

The conditions above apply to any machine learning modelling effort. In the case of the RUL estimation domain however, additional care must be taken when generating the models. The sensor signals are divided into segments of shorter duration (for example windows of 0.1 seconds). First, researcher must ensure that different segments of the same signal never appear in both training and test data sets, thus avoiding data leakage (section 6.2). Second, because we are dealing with time-series data, we cannot assume that the segments are independent and identically distributed (i.i.d). If during model fitting i.i.d is assumed, then the signal segments must be randomly shuffled. Several models should be trained and evaluated with different random orderings. Finally, many models do not require nor use all of the measurements. For example, for tri-axial accelerometers, only data from a single axis may be used. In these cases, to avoid snooping (section 6.3), researchers must either justify the selection of a specific axis/sensor (ex.: automated selection based on some criteria) or show that any axis can be used arbitrarily (ex.: repeat the experiment for all measurements).

8.3. Model Evaluation

Each model must be evaluated under the exact same conditions it was trained. Models should be evaluated with the same error metric that was used for model fitting and cross-validation. Researchers have to use an independent test data set to avoid data leaking (section 6.2). If hyper-parameter tuning is performed, then use cross-validation with a validation data set. All results must be reported, not just a select few (avoid data snooping, section 6.3). Aggregates of results must provide additional statistics (see data averaging 6.4 such as median, inter-quartile range, minimum, maximum (Giles & Lawrence, 1997; Saxena et al., 2010). Formal methods (Shepperd et al., 2019) (Bonferroni correction, Benjamini-Hochberg false discovery rate and Nemenyi post hoc procedure) should be used to compare model performance. Cross-validation (see Section 6.1) has proven to be one of the greater challenges using current benchmarks due to the lack of data.

The most important issue is using an error metric that can fairly and correctly reflect the performance of a model for the full lifetime of a component. Components' life-cycle, and ball bearings in particular, go through various stages of degradation (section 4.2). It is highly non-linear (Kan et al., 2015) and has very high variability (Vlcek et al., 2003). Model accuracy also varies significantly during a component's lifespan (Saxena et al., 2010) - it is low at the start and increases towards the EoL (see for example (Li et al., 2019; Y. Wang et al., 2015; Li et al., 2020; Xia et al., 2019)). Finally, the error of an early or late RUL should reflect the higher cost associated with delayed maintenance interventions.

Work has already been done in this area (Saxena et al., 2010). Here we will present and justify additional recommendations of our own. We first look at the general case where the RUL is estimated during the components' full lifetime (section 8.1). This procedure is then slightly modified for the case of the partial lifetime RUL estimates (only after an FPT).

Note that lifespans differ significantly (in time and number of samples) and the magnitude of the error is much higher at the start of the lifespan (higher uncertainty). This has 2 consequences. The first is that final mean error estimate using all samples depends on the data-series' lengths, and may ultimately, be under or over estimated depending on the lengths of the time-series used in the test data set. The second is that the more conservative models may end up with better overall performance even at a cost of larger errors at the EoL stage. To avoid these issues the error should be calculated as follows:

- Include time-series with large differences in lifetime in the test data set.
- Divide each time-series into 3 sections: the time-series should be divided in the following (arbitrary) proportions: $1/2$, $1/3$ and $1/6$. The goal here is to evaluate how well the model performs during the various stages of degradation.
- For each section select a fixed number of equidistant samples (Saxena et al., 2010) starting from the first segment. We suggest 100 segments to facilitate calculations, but any value will do.
- Calculate the residuals or any other compatible error measurement for each of these 3 sections of the each time-series. The positive error measurements should multiplied by 2 and then used to calculate the error (ex.: square the residual). The constant is an arbitrary value that penalizes late RUL estimates, as is the norm in the domain of predictive maintenance (akin to a much simpler weighing scheme that was used in (Nectoux et al., 2012)). The error must also be weighted by increasing values as a function of the section (ex.: $\{1.0, 2.0, 3.0\}$). These values are arbitrary and reflect the importance of accurate predictions near the EoL.
- Calculate and report the aggregates per section (mean, median, deviations, inter-quartile range, etc.). Provide plots such as the error distribution per section.
- Calculate and report the aggregates for all the benchmark, irrespective of the time-series sections.

The protocol above should provide us with an understanding of the expected behaviour of the model during the degradation phases. The smaller proportions of the later sections reflect the rapid degradation that occurs in the later stages of the component's life-cycle. The expected error (total or per section) will then be independent of the time-series length. As per the recommendation of (Saxena et al., 2010), weighing and ranking of the performance of the algorithm in each section, can be used to compare algorithms. Any of the standard ML metrics (such as mean squared error and standard deviation) or domain specific metrics (such as relative accuracy and convergence in (Saxena et al., 2010)), can be used within this framework.

We have not delved into issues regarding the optimal value of penalty weights, number of sections, proportions of each section, nor the number of samples to take in each section. We assume that this will be set on a per benchmark basis and will most probably depend on domain knowledge and experimental conditions. What is important is that the researchers report on model performance in a more consistent and transparent way.

For the case of the partial lifetime RUL estimates, we assume that FPT will be identified automatically (section 8.1). Following the protocol and evaluation reporting described above, the RUL model will only be *triggered* (learned and evaluated) after the FPT. So the question is, what RUL estimate should be assigned to the segments that occur before the FPT? Note that delaying the FPT as much as possible will facilitate RUL estimation and thereby reduce the error. So a large constant RUL estimate should be assigned to all samples prior to the FPT to penalize late FPTs (equivalent to (Li et al., 2019)). This value must be pretty high - for example using a RUL equal the time duration of the longest time-series. Several other proposals have also been made in the literature (for example (T. Wang, 2012)). The efficacy of the identification of the FPT will indirectly influence the RUL model's error rate. However, with such a large cost, early FPT's (longer prediction periods) will be rewarded.

9. RESULTS

A total of 21 metrics were identified and 19 were used to quantify adherence to proper protocol (section 3.1). For each article, these indicators were evaluated and conformity to best practices was recorded. For each metric a total was tallied and simple statistics were calculated. The data and calculations are available in a spreadsheet included with the supplementary material.

It is important to note that evaluating the correctness of the methodology is difficult because not all of the information is available and no attempt was made to replicate the experiments to check if anything was missing. Replicating experiments is infeasible but, in a few cases, attempts were made to contact authors and obtain additional information. Consequently, when in doubt, we assume proper procedures are not being followed and err on the side of underestimating compliance. Nevertheless, we believe that even so these conservative estimates still allow us to gauge the current status of research in the RUL estimation domain.

In this section we present a summary of the results and highlight several interesting outcomes that supports our hypothesis - that research in the RUL estimation domain do not always follow best practices and that significant improvements are possible.

9.1. General ML best practices

In regards to data leakage, we endeavoured to determine if the researchers performed data splitting correctly using the test set for the final evaluation only (online algorithms must use prequential error estimation (Gama, 2010)). The estimate of around 64% (metric "mf_sep_train_test" in the supplementary spreadsheet) shows that even the most basic requirement, such as having an independent test set, may not always be satisfied. The research results in the RUL domain are therefore most likely overestimating model performance, making future comparative work invalid.

Our analysis points to problems in data snooping, both at the individual and research community level. We found that only around 26% (metric "mf_experiment" in the supplementary spreadsheet) of the references used all the available data for training and testing. Although we cannot claim this for the almost 3/4 of the remaining work, it is a clue that data sets were potentially cherry-picked for reporting.

We also see that a significant proportion of research effort (over 91%) use the same 2 public data sets (see table 2). In other words, we may be witnessing "oversearching" at the community level. Other researchers have already pointed out the need for more data (for example (T. Wang, 2012)).

9.2. Domain Specific best practices

On a more technical note, we have found that of those research articles that use data sets with multiple accelerometer measurements (measured for different directions), about 70% (metric "mf_accel" in the supplementary spreadsheet) of those only use one of the accelerometer readings. In these cases no justification is given for the selection. It is important this be done in future research, so that data snooping is ruled out.

To evaluate the effectiveness of a proposed model, researchers must compare results to prior work using the same data sets. We counted the number of valid comparisons that were made per article (see comparisons variables in the worksheet in the supplementary material). Surprisingly, only 24.3% of the sampled articles made such comparisons. On average only 1.82 articles were used for comparison (maximum was 4 articles).

We further analyzed the 17 cases that did compare results to prior work. First, we counted how many of these cases used models that performed RUL estimation for the same lifetime (see Table 5). We found that 88.2% made equivalent comparisons (2 of the 17 were possibly incorrect). Second, we counted how many of the 17 cases used models that performed RUL estimation for the same operating regimes (see Table 5). Only 47.1% made equivalent comparisons (9 of the 17 were possibly incorrect). This means that at best only 11.4% (8 of the total 70) of all the research efforts made equivalent comparisons.

As described in section 8.1 on problem classes, it is still correct to compare the performance of general models with their more specific counterparts (see Table 5). We therefore extended the previous analysis to include these cases. For compatible lifetime comparisons, we only found 1 more article resulting in a total of 94.1% valid comparisons (1 of the 17 was incorrect). For compatible operating regimes we found an additional 7 articles resulting in a total of 88.2% valid comparisons (2 of the 17 were incorrect). This means that at best only 21.4% (15 of the total 70) of all the research efforts made valid comparisons.

We then investigated what type of analysis was made by the majority of the articles that did not compare to prior work. We found that a total 41.4% of the researchers developed and compared several models of their own. Surprisingly, 31.4% made no comparisons whatsoever. The remaining 2.9% compare to research that used different data sets (we assume they replicated the research with the correct data set).

9.3. Best practices by category

For an overview of the results, we estimated the % of adherence to proper protocol according to 3 categories of best practices. We used the median of the metrics that were grouped into those categories to aggregate the values (see the "by category" table in the supplementary worksheet). We obtained the following results:

- Problem class selection: 26%
- Model Fitting best practices: 30%
- Evaluation best practices: 14%

The number of variables in each category varies from 3 to 7 (see supplementary material). We opted to use the median because it is more robust to the extreme values found in categories with a small number of variables. Some caution is therefore required when interpreting these results. Notwithstanding the limitation, we believe the results are informative, and several conclusions may be drawn.

First, the "evaluation best practices" category has the poorest performance of all. Both the evaluation procedure and metrics are neither standardized nor used consistently. This issue was already identified several years ago (Saxena et al., 2010), but is still observed in the most recent research. We see for example that only a little more than 18% (metric "ev_weighted" in the supplementary spreadsheet) of researchers weigh differently early and late predictions. And even in these cases, the weighing is not always fully specified (for example (Duong et al., 2018)). More importantly only 11% (metric "ev_nstep" in the supplementary spreadsheet) of the researchers use a fixed number of samples for error estimation in order to avoid issue related to the time-series length. Finally, we have yet to find reports that show how the model performs in the various stages of degradation (metric "ev_per_hs" in the supplementary spreadsheet) - only a single metric is provided for the full signal duration. In the cases where aggregates on model performance are provided (test data sets with multiple instances), only 17% (metric "ev_aggregates" in the supplementary spreadsheet) provide additional statistical information, such as a confidence interval.

Second, all categories show that adherence to good practices is very low (all well below 50%). This indicates that replication and comparison of the results is severely impaired. With the already mentioned caveat that these are only rough estimates, we may conclude that no more than 11% (ev_nstep) of the research in RUL estimation can be replicated without additional work on the original models.

10. CONCLUSION

The difficulty of replicating and comparing research in the domain of RUL estimation of ball bearing failures, has motivated this meta-analysis on experimental protocol. We have found that the procedures vary significantly among researchers and in many cases are incomplete or incorrect. As a result we have made a general review on machine learning literature to determine which are the most common pitfalls. We then use that information to establish metrics that measure adherence to correct experimental procedures for data analysis and model comparison. We also studied a number of research article on RUL estimation with the aim of identifying the best practices in this specific domain. With this information we proposed a general protocol that should be used when comparing RUL estimation models.

A total of 70 papers on RUL were audited in order to identify the extent to which errors and omission in both experimental procedures and reporting inhibit progress in this domain. The survey showed that roughly 11% of the papers present work that will allow for replication and comparison. Of these, no article follows correct procedure without some error. The most common issue that was identified is in the category of the "evaluation best practices".

We believe the sample is large enough that it represents the true extent of these issues. However, it is a challenge to ascertain whether or not we have all of the information required to replicate experiments and compare models without attempting to do so. Quite possibly the estimates of adherence to correct protocol we have are underestimated.

We think that making a platform or workframe available that automates and systematizes the procedure described in our protocol, will go a long way in avoiding many of the problems that were identified. The emphasis should be on enforcing the domain specific procedures that take into account the detection of the FPT and evaluating compatible RUL models (problem classes). It should allow researchers to quickly add new data sets and easily compare results with existing work. This has the potential of advancing research in this domain. Such a platform could also contribute to other related areas of RUL estimation.

Although establishing correct protocol is of paramount importance, another serious issue is the lack of data. Almost all of the research effort (over 91%) use the same 2 public data sets (see table 2). These data sets also present additional problems. First, all but the GPMS (Ben Ali et al., 2018) data set, were obtained under laboratory conditions, making model evaluation sub-optimal. Second, they consist of a very limited number of instances (Table 4), which potentially introduces problem with data snooping. And finally, the duration of the time-series are very different (Table 3), which make model evaluation difficult. The models' good performance on these benchmarks are therefore overly optimistic. These results may inadvertently create high expectations of good performance. However, try as they might, practitioners training these same models on new unseen data, will be unable to reach the same level of performance.

Although it is an expensive and time-consuming endeavour (Saxena et al., 2010), it is imperative that an extensive set of benchmarks be made available to researchers. More concretely, data sets obtained from factory settings are urgently needed.

In addition to this we have found that a total of 21.5% (Table 2) of the articles do not make any of their data available and none of the articles we audited made their source code available. We urge researchers to adopt the principles of *Open Science* as a means of disseminating their results. Reproducible experiments provide opportunities for identifying and correcting errors, thereby facilitating progress (Shepperd et al., 2019).

We have, by no means, exhausted the subject of RUL estimation best practices. For example, issues regarding the applicability of model fitting and evaluation procedures using infinite time-series, imposes additional restrictions. However, we hope the details and recommendations herein will facilitate researchers' work in the future.

ACKNOWLEDGMENT

This work was supported by the European Regional Development Fund (FEDER) through a grant to the Operational Program for the Competitiveness and Internationalization of Portugal 2020 via the PRODUTECH-SIF (Ref. n^{o} POCI-01-0247-FEDER-024541) Partnership Agreement. We thank the anonymous reviewers whose suggestions helped improve and clarify this work.

REFERENCES

Ben Ali, J., Saidi, L., Harrath, S., Bechhoefer, E., & Benbouzid, M. (2018). Online automatic diagnosis of wind turbine bearings progressive degradations under real experimental conditions based on unsupervised machine learning. *Applied Acoustics*, *132*(November 2017), 167–181. doi: 10.1016/j.apacoust.2017.11.021

Benkedjouh, T., Medjaher, K., Zerhouni, N., & Rechak, S. (2013, aug). Remaining useful life estimation based on nonlinear feature reduction and support vector regression. *Engineering Applications of Artificial Intelligence*, *26*(7), 1751–1760. Retrieved from `https://www.sciencedirect.com/science/article/abs/pii/S0952197613000365?via{\%}3Dihub` doi: 10.1016/j.engappai.2013.02.006

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Bozchalooi, I. S., & Liang, M. (2008). A joint resonance frequency estimation and in-band noise reduction method for enhancing the detectability of bearing fault signals. *Mechanical Systems and Signal Processing*, *22*(4), 915–933. doi: 10.1016/j.ymssp.2007.10.006

Cheng, Z. (2017, aug). Residual useful life prediction for rolling element bearings based on multi-feature fusion regression. In *Proceedings - 2017 international conference on sensing, diagnostics, prognostics, and control, sdpc 2017* (Vol. 2017-Decem, pp. 246–250). IEEE. Retrieved from `http://ieeexplore.ieee.org/document/8186510/` doi: 10.1109/SDPC.2017.54

Colquhoun, D. (2018). Correction to 'the reproducibility of research and the misinterpretation of p-values'. *Royal Society Open Science*, *5*.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1 - 17. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0020025512007098` doi: https://doi.org/10.1016/j.ins.2012.10.039

Dong, G., & Chen, J. (2012). Noise resistant time frequency analysis and application in fault diagnosis of rolling element bearings. *Mechanical Systems and Signal Processing*, *33*, 212–236. Retrieved from `http://dx.doi.org/10.1016/j.ymssp.2012.06.008` doi: 10.1016/j.ymssp.2012.06.008

Duong, B. P., Khan, S. A., Shon, D., Im, K., Park, J., Lim, D. S., ... Kim, J. M. (2018, nov). A reliable health indicator for fault prognosis of bearings. *Sensors (Switzerland)*, *18*(11), 3740. Retrieved from `http://www.mdpi.com/1424-8220/18/11/3740` doi: 10.3390/s18113740

England, J. R., & Cheng, P. M. (2019, mar). *Artificial intelligence for medical image analysis: A guide for authors and reviewers* (Vol. 212) (No. 3). American Roentgen Ray Society. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/30557049` doi: 10.2214/AJR.18.20490

Gama, J. (2010). *Knowledge discovery from data streams* (1st ed.). Chapman & Hall/CRC. doi: 10.1201/EBK1439826119

Giles, C. L., & Lawrence, S. (1997). Presenting and analyzing the results of AI experiments: data averaging and data snooping. *Proceedings of the National Conference on Artificial Intelligence*, 362–367.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Retrieved from `http://www.deeplearningbook.org`

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer-Verlag New York. doi: 10.1007/978-0-387-84858-7

Heng, A., Zhang, S., Tan, A. C., & Mathew, J. (2009). Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, *23*(3), 724–739. doi: 10.1016/j.ymssp.2008.06.009

Ioannidis, J. P. A. (2005, 08). Why most published research findings are false. *PLoS Med*, *2*(8), e124. Retrieved from `http://dx.doi.org/10.1371%2Fjournal.pmed.0020124` doi: 10.1371/journal.pmed.0020124

Jensen, D. (2000, jan). Data snooping, dredging and fishing. *ACM SIGKDD Explorations Newsletter*, *1*(2), 52. Retrieved from `http://portal.acm.org/citation.cfm?doid=846183.846195` doi: 10.1145/846183.846195

Jiang, J.-R., Lee, J.-E., & Zeng, Y.-M. (2019, dec). Time Series Multiple Channel Convolutional Neural Network with Attention-Based Long Short-Term Memory for Predicting Bearing Remaining Useful Life. *Sensors*, *20*(1), 166. Retrieved from `https://www.mdpi.com/1424-8220/20/1/166` doi: 10.3390/s20010166

Kan, M. S., Tan, A. C., & Mathew, J. (2015). A review on prognostic techniques for non-stationary and non-linear rotating systems. *Mechanical Systems and Signal Processing*, *62*, 1–20. Retrieved from `http://dx.doi.org/10.1016/j.ymssp.2015.02.016` doi: 10.1016/j.ymssp.2015.02.016

Kaufman, S., Rosset, S., & Perlich, C. (2011). Leakage in data mining: Formulation, detection, and avoidance. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining*

(p. 556–563). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2020408.2020496 doi: 10.1145/2020408.2020496

Khelif, R., Chebel-Morello, B., Malinowski, S., Laajili, E., Fnaiech, F., & Zerhouni, N. (2017). Direct Remaining Useful Life Estimation Based on Support Vector Regression. *IEEE Transactions on Industrial Electronics*. doi: 10.1109/TIE.2016.2623260

Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mechanical Systems and Signal Processing*, *104*, 799–834. Retrieved from https://doi.org/10.1016/j.ymssp.2017.11.016 doi: 10.1016/j.ymssp.2017.11.016

Li, X., Zhang, W., & Ding, Q. (2019). Deep learning-based remaining useful life estimation of bearings using multi-scale feature extraction. *Reliability Engineering and System Safety*, *182*, 208–218. doi: 10.1016/j.ress.2018.11.011

Li, X., Zhang, W., Ma, H., Luo, Z., & Li, X. (2020, jun). Data alignments in machinery remaining useful life prediction using deep adversarial neural networks. *Knowledge-Based Systems*, *197*, 105843. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0950705120302124?via{\%}3Dihub doi: 10.1016/J.KNOSYS.2020.105843

Liu, C., Zhang, L., & Wu, C. (2019, dec). Direct Remaining Useful Life Prediction for Rolling Bearing Using Temporal Convolutional Networks. In *2019 ieee symposium series on computational intelligence, ssci 2019* (pp. 2965–2971). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/SSCI44817.2019.9003163

Liu, Z., Zuo, M. J., & Qin, Y. (2015, jun). Remaining useful life prediction of rolling element bearings based on health state assessment. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *230*(2), 314–330. Retrieved from https://doi.org/10.1177/0954406215590167 doi: 10.1177/0954406215590167

Machacek, V., & Srholec, M. (2019). *Predatory publications in scopus: Evidence on cross-country differences* (Working Papers IES No. 2019/20). Charles University Prague, Faculty of Social Sciences, Institute of Economic Studies. Retrieved from https://EconPapers.repec.org/RePEc:fau:wpaper:wp2019_20

Mao, W., He, J., Tang, J., & Li, Y. (2018, dec). Predicting remaining useful life of rolling bearings based on deep feature representation and long short-term memory neural network. *Advances in Mechanical Engineering*, *10*(12), 168781401881718. Retrieved from http://journals.sagepub.com/doi/10.1177/1687814018817184 doi: 10.1177/1687814018817184

Measurement Computing Corporation. (2004). *Data Acquisition Handbook. A Reference for DAQ And Analog & Digital Signal conditioning* (3rd ed.). MA: Author. Retrieved from https://www.mccdaq.com/pdfs/anpdf/Data-Acquisition-Handbook.pdf

Moreno-Torres, J. G., Saez, J. A., & Herrera, F. (2012). Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(8), 1304-1312.

Munafò, M., Nosek, B., Bishop, D., Button, K., Chambers, C., Percie Du Sert, N., ... Ioannidis, J. (2017, 1 10). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1). doi: 10.1038/s41562-016-0021

Murphy, K. (2017, 12). Harking: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*. doi: 10.1007/s10869-017-9524-7

Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-morello, B., Zerhouni, N., ... Varnier, C. (2012). PRONOSTIA : An experimental platform for bearings accelerated degradation tests. *IEEE International Conference on Prognostics and Health Management*, 1–8.

P. Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press.

Peng, Y., Cheng, J., Liu, Y., Li, X., & Peng, Z. (2018, jun). An adaptive data-driven method for accurate prediction of remaining useful life of rolling bearings. *Frontiers of Mechanical Engineering*, *13*(2), 301–310. doi: 10.1007/s11465-017-0449-7

Qiu, H., Lee, J., Lin, J., & Yu, G. (2006). Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of Sound and Vibration*, *289*(4), 1066 - 1090. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022460X0500221X doi: https://doi.org/10.1016/j.jsv.2005.03.007

Russell, S., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach* (3rd ed.). Pearson. Retrieved from http://aima.cs.berkeley.edu/

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebe, K. (2009). On applying the prognostic performance metrics. *Annual Conference of the Prognostics and Health Management Society, PHM 2009*, 1–16.

Saxena, A., Celaya, J., Saha, B., Saha, S., & Goebel, K. (2010). Metrics for offline evaluation of prognostic performance. *International Journal of Prognostics and Health Management*, *1*(1), 1–20.

Shan, P., Hou, P., Ge, H., Yu, L., Li, Y., & Gu, L. (2019, oct). Image Feature-Based for Bearing Health Monitoring with Deep-Learning Method. In *2019 prognostics and system health management conference (phm-qingdao)* (pp. 1–6). IEEE. Retrieved from `https://ieeexplore.ieee.org/document/8942972/` doi: 10.1109/PHM-Qingdao46334.2019.8942972

Shepperd, M. J., Guo, Y., Li, N., Arzoky, M., Capiluppi, A., Counsell, S., ... Yousefi, L. (2019). The prevalence of errors in machine learning experiments. In H. Yin, D. Camacho, P. Tiño, A. J. Tallón-Ballesteros, R. Menezes, & R. Allmendinger (Eds.), *Intelligent data engineering and automated learning - IDEAL 2019 - 20th international conference, manchester, uk, november 14-16, 2019, proceedings, part I* (Vol. 11871, pp. 102–109). Springer. Retrieved from `https://doi.org/10.1007/978-3-030-33607-3\_12` doi: 10.1007/978-3-030-33607-3_12

Soualhi, A., Medjaher, K., & Zerhouni, N. (2015, jan). Bearing Health Monitoring Based on Hilbert–Huang Transform, Support Vector Machine, and Regression. *IEEE Transactions on Instrumentation and Measurement*, *64*(1), 52–62. Retrieved from `http://ieeexplore.ieee.org/document/6847199/` doi: 10.1109/TIM.2014.2330494

Sutrisno, E., Oh, H., Vasan, A. S. S., & Pecht, M. (2012). Estimation of remaining useful life of ball bearings using data driven methodologies. In *2012 ieee conference on prognostics and health management* (pp. 1–7). doi: 10.1109/ICPHM.2012.6299548

Taylor, H. R. (1997). *Data Acquisition for Sensor Systems* (1st ed.). Springer US. doi: 10.1007/978-1-4757-4905-2

Verstraete, D., Droguett, E., & Modarres, M. (2019, dec). A Deep Adversarial Approach Based on Multi-Sensor Fusion for Semi-Supervised Remaining Useful Life Prognostics. *Sensors*, *20*(1), 176. Retrieved from `https://www.mdpi.com/1424-8220/20/1/176` doi: 10.3390/s20010176

Vlcek, B. L., Hendricks, R. C., & Zaretsky, E. V. (2003, jan). Determination of Rolling-Element Fatigue Life From Computer Generated Bearing Tests. *Tribology Transactions*, *46*(4), 479–493. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/10402000308982654` doi: 10.1080/10402000308982654

Wang, B., Lei, Y., Li, N., & Li, N. (2020). A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings. *IEEE Transactions on Reliability*, *69*(1), 401–412. doi: 10.1109/TR.2018.2882682

Wang, B., Lei, Y., Yan, T., Li, N., & Guo, L. (2020, feb). Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery. *Neurocomputing*, *379*, 117–129. doi: 10.1016/j.neucom.2019.10.064

Wang, T. (2012, jun). Bearing life prediction based on vibration signals: A case study and lessons learned. In *Phm 2012 - 2012 ieee int. conf.on prognostics and health management: Enhancing safety, efficiency, availability, and effectiveness of systems through phm technology and application, conference program* (pp. 1–7). IEEE. Retrieved from `http://ieeexplore.ieee.org/document/6299547/` doi: 10.1109/ICPHM.2012.6299547

Wang, Y., Peng, Y., Zi, Y., Jin, X., & Tsui, K.-L. (2015, jul). An integrated Bayesian approach to prognositics of the remaining useful life and its application on bearing degradation problem. In *2015 ieee 13th international conference on industrial informatics (indin)* (pp. 1090–1095). IEEE. Retrieved from `http://ieeexplore.ieee.org/document/7281887/` doi: 10.1109/INDIN.2015.7281887

Williams, T., Ribadeneira, X., Billington, S., & Kurfess, T. (2001). Rolling element bearing diagnostics in run-to-failure lifetime testing. *Mechanical Systems and Signal Processing*, *15*(5), 979 - 993. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0888327001914189` doi: https://doi.org/10.1006/mssp.2001.1418

Xia, M., Li, T., Shu, T., Wan, J., De Silva, C. W., & Wang, Z. (2019). A Two-Stage Approach for the Remaining Useful Life Prediction of Bearings Using Deep Neural Networks. *IEEE Transactions on Industrial Informatics*, *15*(6), 3703–3711. doi: 10.1109/TII.2018.2868687

Zhang, Y., Hutchinson, P., Lieven, N. A., & Nunez-Yanez, J. (2020). Remaining useful life estimation using long short-term memory neural networks and deep fusion. *IEEE Access*, *8*, 19033–19045. doi: 10.1109/ACCESS.2020.2966827

Zheng, Y. (2019, dec). Predicting Remaining Useful Life Using Continuous Wavelet Transform Integrated Discrete Teager Energy Operator with Degradation Model. In *2019 ieee 5th international conference on computer and communications, iccc 2019* (pp. 240–244). IEEE. Retrieved from `https://ieeexplore.ieee.org/document/9064232/` doi: 10.1109/ICCC47050.2019.9064232