Hybrid Fault Prognostics for Nuclear Applications: Addressing Rotating Plant Model Uncertainty

J Blair[1], B Stephen [2], B Brown [3], A Forbes [4], and S McArthur[5]

[1,2,3,5] *University of Strathclyde, Glasgow, United Kingdom*
j.blair@strath.ac.uk
bruce.stephen@strath.ac.uk
blair.brown@strath.ac.uk
s.mcarthur@strath.ac.uk

[1,4] *National Physical Laboratory, Teddington, United Kingdom*
alistair.forbes@npl.co.uk

Abstract

Nuclear plant operators are required to understand the uncertainties associated with the deployment of prognostics tools in order to justify their inclusion in operational decision - making processes and satisfy regulatory requirements. Operational uncertainty can cause underlying prognostics models to underperform on assets that are subject to evolving impacts of age, manufacturing tolerances, operating conditions, and operating environment effects, of which may be captured through a condition monitoring (CM) system that itself may be degraded. Sources of uncertainty in the data acquisition pipeline can impact the health of CM data used to estimate the remaining useful life (RUL) of assets. These uncertainties can disguise or misrepresent developing faults, where (for example) the fault identification is not achieved until it has progressed to an unmanageable state. This leaves little flexibility for the operator's maintenance decisions and generally undermines model confidence.

One method to quantify and account for operational uncertainty is calibrated hybrid models, employing physics, knowledge or data driven methods to improve model accuracy and robustness. Hybrid models allow known physical relations to offset full reliance on potentially untrustworthy data, whilst reducing the need for an abundance of representative historical data to reliably identify the monitored asset's underlying behavioural trends. Calibration of the model then ensures the model is updated and representative of the real monitored asset by accounting for differences between the physics or knowledge model and CM data.

In this paper, an open-source bearing knowledge informed machine learning (ML) model and CM datasets are utilized in an illustrative bearing prognostic application. The uncertainty incurred by the decisions made at key stages in the development of the model's data acquisition and processing pipeline are assessed and demonstrated by the resultant impact on RUL prediction performance. It was shown that design decisions could result in multiple valid pipeline designs which generated different predicted RUL trajectories, increasing the uncertainty in the model output.

Index Terms— bearing prognostics, condition monitoring, hybrid systems, model calibration, uncertainty capture.

1. Introduction

Most often, asset maintenance is conducted reactively, whereby corrective maintenance is conducted once a failure has occurred (Canada Nuclear Safety Commission, 2012). In a nuclear power plant (NPP), an unexpected outage of an asset can be expensive due to lost revenue from interrupted generation with downtimes being potentially lengthened by the requirement to: retrospectively identify the root of the fault, source required components and perform the maintenance action. With many NPP's coming to the end of their designed lifetime, many operators are utilising CM and condition based maintenance (CBM) techniques to justify and manage NPP lifetime extensions and to avoid unplanned outages (Coble, Ramuhalli, Bond, Hines, & Upadhyaya, 2015). This requires aging assets to be closely monitored to estimate asset health and ensure extension plans are affordable.

A common asset in NPP's are rotating plant (e.g. motors, turbines, centrifugal pumps, fans), which are prone to bearing failure (Yung & Bonnett, 2004). These could be turbine or motor driven pumps which form part of a larger generation or cooling system. Despite being relatively simple components, bearings are largely responsible for the reliable operation of rotating plant by supporting huge loads to reduce friction on downstream components (Jammu & Kankar, 2011). As such, if bearing faults are left untreated, damage could propagate through the drivetrain and create wider system complications in more expensive components, such as the gear box (Rexnord Industries, LLC, Gear Group, n.d.). Cascading failures would lead to expensive and lengthy maintenance intervention which would cause disruption to plant generation and incur additional regulatory reporting overhead.

CBM and data based analytics can be used to estimate the RUL of rotating plant bearings which, if effective, can provide sufficient warning of an impending failure and an indication of the type of failure developing. An operator can incorporate this into their resource scheduling and budgeting actions to ensure the asset is taken offline and serviced while minimising disruption to NPP operation. However, developing and applying this approach requires access to data, and with specific regards to NPP's, these systems were designed before modern digital sensing and monitoring techniques were available for the hostile environments where they operate, which is an additional consideration that can impact upon the associated data acquisition components. This can result in operators making decisions on unhealthy data collected from NPP's which are not ideally designed for modern sensing systems, adding additional uncertainty to maintenance plans.

Sources of uncertainty can impact the data acquisition pipeline at every stage, including: the choice of sensor type and placement; the chosen sampling rate; data pre-processing steps to present the data in a specific format; and, the metric(s) used by the analytics to convey information to an operator. Design choices at each of these stages offer a trade off, which will incur uncertainty in the output of the pipeline at each stage and can be compounded by the interaction between upstream and downstream pipeline stages. In addition to this, data based analytics generally do not attribute a measure of confidence in their output, making it difficult to determine if the analytics are performing poorly in a sub-optimal pipeline. This makes ML outputs difficult to trust for inclusion in risk and cost assessments. Also they do not provide the operator with relevant information that could allow future improvements to the pipeline to be made.

2. Contribution

The contribution of this work is not the creation of a novel RUL technique, but to demonstrate and quantify the confidence associated with the application of existing hybrid RUL approaches with the associated data acquisition pipeline decisions. Confidence can be undermined by these choices, which impact the performance of the underpinning model and can reduce the operators trust in the whole decision support system. Without sufficient trust, especially in the heavily regulated nuclear engineering environment, decision support tools will not be utilised to support maintenance scheduling activities. As such, the methodology presented in this paper is concerned with investigating the uncertainty in analytic design and deployment by capturing the sources of uncertainty and demonstrating how these impact on an uncertainty budget for the whole data to decision pipeline rather than just the output of the ML model. In this work, the uncertainty in the model performance due to the whole pipeline deisgn is captured by analysing the quantiles of the model outputs under different data acquisition pipeline designs. Evidence is presented in the form of case-studies using open-source, curated test rig data to reduce the impact of excessive operational noise, that were performed to evaluated data pipeline uncertainty.

3. Literature

The literature review covers research trends in bearing prognostics applications, with a focus on data-based methods as these require access to healthy CM data. This is supported by a section on hybrid modelling where knowledge- and data-based methods are combined, and how diverse approaches may be combined for prognostic applications. Finally, uncertainty capture methods with particular focus on computer modelling prognostic methods is presented.

3.1. Data-Based Bearing Prognostics

Bearings are subject to high stress operating conditions which makes failures common. These can manifest due to overloading or imbalanced loading, lubrication issues due to insufficient lubrication, contamination or sealing failures. Bearings are mechanical faults and mechanical failures are most commonly monitored via vibration monitoring, although have been approached using temperature, oil analysis and accoustic emission approaches (Kumar et al., 2019). Vibration monitoring, while subjected to the robustness and cost of the sensor system, allows changes in bearing health to be observed immediately and has been proven as a reliable method for bearing fault prognosis. Temperature based schemes are most useful for end of life where the fault has progressed significantly, oil analysis methods require the bearings to have a dedicated supply system and acoustic emission requires access to high quality measurements (Jammu & Kankar, 2011).

A survey of 274 prognostic approaches by (Lei et al., 2018) separated works into statistical-, AI-, physics- and hybrid-based approaches, with 56% contribution from statistical based methods, and 26% from AI based approaches which both rely heavily on available CM data. ML or Deep Learning (DL) approaches are gaining increasing popularity as they can handle complex prognosis problems which may be traditionally difficult to create reliable physics or statistical models for, however due to their black-box nature it is difficult to justify their usage in safety critical applications. The approaches which gained the most attention for machine prognosis in (Lei et al., 2018) review were Artificial Neural Networks, Neuro-Fuzzy systems (both DL methods), Support Vector Machines (SVM), K-nearest neighbour (*k*NN) and Gaussian Process Regression. DL approaches require access to large quantities of high quality, representative data which can be unobtainable in some industrial settings, however can produce excellent RUL predictions in return. ML models such as the SVM and *k*NN methods can provide better performance in cases with limited access to representative data, however are subject to appropriate kernel and parameter selection (Nisbet, Elder, & Miner, 2009). Gaussian Process Regression are computationally expensive when utilising large number of samples due to a required matrix inversion, but is a flexible method that can be updated with new data, adapt to limited data and incorporate uncertainties (Hart, 2018).

3.2. Hybrid Models

A single knowledge-, physics- or data- based approach is unlikely to provide effective system coverage for multiple failure modes and fault types. Utilising a combination of approaches aims to leverage the relative advantages of each individual method while limiting the impact of their respective weaknesses (Baur, Albertelli, & Monno, 2020). (Goebel, Eklund, & Bonanni, 2006) found that combining a bearing physics of failure model with an empirical method based on measured data (Dempster-Shafer Regression) produced more accurate RUL prediction results than either method independently.

The method of combining two or more of these methods in a hybrid approach varies and tends to be application specific due to the relatively early development stage of the research field as shown by the small (8 %) contribution to the canvassed literature in (Lei et al., 2018). As such, many methods of creating hybrid models are being explored, such as utilising one model to estimate the asset health state and another for RUL estimation; combining the RUL estimates from multiple methods; or utilising one method for short-term forecasting and another method for long-term forecasting (Ramuhalli, Walker, Agarwal, & Lybeck, 2020). Of particular interest in this work is the combination of knowledge- and data-based approaches. Incorporating domain knowledge into data-driven approaches allows known trends and rules that govern the degradation patterns to be encoded to support the prognostic tool in identifying and predicting the failure dynamics of well understood failure modes. The data-driven component can provide the needed flexibility to apply and extrapolate these rules into an RUL estimate tailored to the monitored asset, while providing capability to identify new failure modes not included in the encoded expert knowledge (Liao & Köttig, 2014).

3.3. Uncertainty Capture in computer models

ML models tend to produce point estimates which does not provide information about the likely distribution of potential predictions. Attributing confidence intervals to output predictions (typically corresponding to a confidence level of 95 % (JCGM Working Group 1, 2008)) provides more appropriate information about the anticipated range of outputs and expected value of the prediction, allowing operators more agency to utilise the results. Bayesian methods are usually incorporated into prognostic approaches to handle uncertainty capture and propagation, such as in Bayesian Networks, Bayesian Neural Networks and Kalman/particle filtering algorithms. Non-Bayesian methods that have been used include Monte Carlo based, bootstrapping or closed-form mathematical solutions. A major drawback of these approaches are the lengthy and difficult process of collecting and formalising prior knowledge and assumptions which may be impractical for some complex system applications (Zhao et al., 2021).

Additionally, computer models themselves introduce sources of uncertainty. This was demonstrated by (Kennedy & O'Hagan, 2001) who utilised a Bayesian approach to computer model calibration that incorporated all forms of uncertainty previously discussed in the research space. These included: parameter uncertainty, where the value of context specific features are unknown but assumed; random effects, where the real process may experience random fluctuations given the same experienced conditions; model inaccuracy, where the complexity of the model is unable to truly reflect the real process; data collection errors, where there are sources of uncertainty in the CM data; and uncertainty of the unseen code output, where there exists unprocessed and potentially more optimal configurations of the model.

4. Uncertainty in Hybrid Model Data Pipelines

The proposed methodology to assess the uncertainty is presented in the form of a case study that investigates the impact of decisions made in the data acquisition and processing pipeline through the resulting uncertainty in the RUL prediction for motor bearing prognostics.

4.1. Condition Monitoring Datasets

Two open source bearing prognostics datasets are used in this work: NASA IMS (Lee, J. , Qiu, H. , Lin, J. and Rexnord Technical Services, 2007) and NASA FEMTO (NASA Ames Prognostics Data Repository, 2012). Both datasets observe run to failure experiments for bearings with no initial defects. Each data set has visibility of the bearings failures by vertically and horizontally mounted accelerometers (termed 'x-axis' and 'y-axis' respectively), with limited access to the vertical data for the IMS dataset. Four distinct bearing failures are observed in the IMS dataset, with two occurring concurrently, while the FEMTO dataset contains 17 run to failure examples. The IMS failures were accelerated due to intensive, but in specification, bearing loading conditions, while the FEMTO dataset was created using the PRONOSTIA test rig which artificially overloaded the bearings to further accelerate wear.

4.2. Existing Hybrid Model

4.2.1. Combining Knowledge- and Data-driven Components

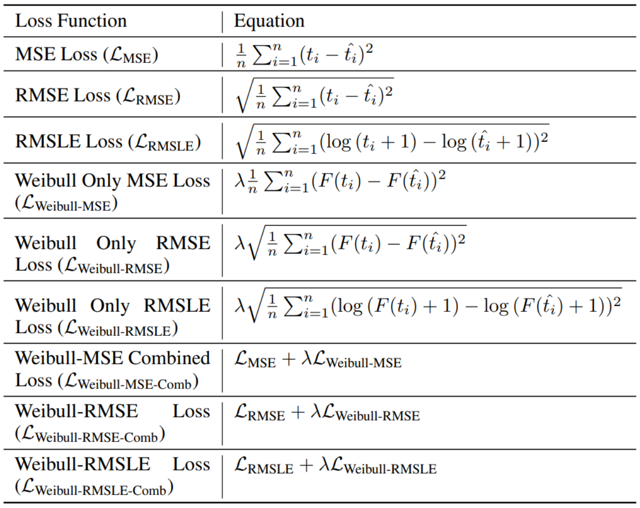
An open source hybrid RUL model consisting of a novel Weibull-based loss function for Neural Networks (NN) by (Hahn & Mechefske, 2022) was chosen as the basis for this study. Utilising a Weibull distribution to capture domain knowledge from the field of reliability engineering, the authors create 9 NN loss functions to evaluate the success of their knowledge informed ML model for bearing prognosis on the IMS and FEMTO datasets. The knowledge component of the hybrid model is captured by using the data to calibrate the Weibayes equation (Abernethy, 2004) shown in equation 1. The one parameter Weibayes has been shown to produce accurate results for a small number of failures (<20) where the estimated value of shape parameter, β, is representative of the true system behaviour (Abernethy, 2004). The value of β was fixed at a value of 2 in (Hahn & Mechefske, 2022) due to model stability concerns, and this value being deemed a reasonable shape estimate for ball bearing failures (Abernethy, 2004). The values of η and β are used to calculate the Weibull cumulative distribution function (CDF) in equation 2. The 9 loss functions are shown in figure 1 and are incorporated into the model as the loss function to be minimised by the NN in the back-propagation step.

$$\eta = \left[\sum_{i=1}^{N} \frac{t_i^{\beta}}{r} \right]^{\frac{1}{\beta}} \tag{1}$$

$$F(t) = 1 - e^{-\left(\frac{t}{\eta}\right)^{\beta}} \tag{2}$$

Where

- t = time or cycles,
- r = number of failed units,
- N = total number of failures plus currently running units (incomplete failures)
- η = maximum likelihood estimate of the unit characteristic life (63.2 distribution percentile)
- β = Weibull shape parameter, and
- F(t) is the Weibull CDF



Loss Function	Equation
MSE Loss ($\mathcal{L}_{\text{MSE}}$)	$\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t_i})^2$
RMSE Loss ($\mathcal{L}_{\text{RMSE}}$)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t_i})^2}$
RMSLE Loss ($\mathcal{L}_{\text{RMSLE}}$)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(t_i + 1) - \log(\hat{t_i} + 1))^2}$
Weibull Only MSE Loss ($\mathcal{L}_{\text{Weibull-MSE}}$)	$\lambda\frac{1}{n}\sum_{i=1}^{n}(F(t_i) - F(\hat{t_i}))^2$
Weibull Only RMSE Loss ($\mathcal{L}_{\text{Weibull-RMSE}}$)	$\lambda\sqrt{\frac{1}{n}\sum_{i=1}^{n}(F(t_i) - F(\hat{t_i}))^2}$
Weibull Only RMSLE Loss ($\mathcal{L}_{\text{Weibull-RMSLE}}$)	$\lambda\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(F(t_i) + 1) - \log(F(\hat{t_i}) + 1))^2}$
Weibull-MSE Combined Loss ($\mathcal{L}_{\text{Weibull-MSE-Comb}}$)	$\mathcal{L}_{\text{MSE}} + \lambda\mathcal{L}_{\text{Weibull-MSE}}$
Weibull-RMSE Loss ($\mathcal{L}_{\text{Weibull-RMSE-Comb}}$)	$\mathcal{L}_{\text{RMSE}} + \lambda\mathcal{L}_{\text{Weibull-RMSE}}$
Weibull-RMSLE Loss ($\mathcal{L}_{\text{Weibull-RMSLE-Comb}}$)	$\mathcal{L}_{\text{RMSLE}} + \lambda\mathcal{L}_{\text{Weibull-RMSLE}}$

Figure 1. Loss functions from (Hahn & Mechefske, 2022)

Dataset	Train.	Val.	Test.
IMS	Run 2 (B 1) Run 3 (B 3)	Run 1 (B 3)	Run 1 (B 4)
FEMTO	Bearing1_1 Bearing2_1 Bearing3_1	Bearing1_2 Bearing2_2 Bearing3_2	Bearing1_3 Bearing2_3 Bearing3_3

Table 1. Data split between training, validation and testing

4.2.2. RUL Estimation Procedure

(Hahn & Mechefske, 2022) conducted the following process to generate RUL estimates for both the IMS and FEMTO datasets. First, the input data from the horizontal sensors was processed into spectrograms to obtain the frequency representation of the vibration data. The number of input features was reduced by 'binning' the spectrogram into 20 bins, where the maximum value of the frequency bands included in each bin is taken as the value for that bin, repeated for each timestep. The response variable was the lifetime percentile status of the bearing, with 0 % being healthy bearing at the start of the experiment, to 100 % signifying the failure of the bearing at the end of the experiment. The training, validation and testing split of the datasets are shown in table 1.

The Weibayes equation was calibrated with the training data to be incorporated into the loss functions. To initialise and optimise the NN architecture, a random search was conducted to select from the hyper parameters shown in table 2 for each of the loss functions, which the authors set to 1000 in their study. The coefficient of determination (R^2) and Root Mean Squared Error (RMSE) were used to discard models that performed poorly, with models with a $R^2 > 0.2$ and $RMSE < 0.35$ progressed to the testing stage. After testing, the models were filtered again by the R^2 and $RMSE$ bounds before selecting a subset of the top performing models based on the R^2 metric. The authors found that the top performing loss function for the IMS dataset was the Weibull-RMSLE combined, and the Weibull-MSE combined for the FEMTO dataset, both containing the knowledge informed loss function.

Parameter	Selection Choice
Batch size	32, 64, 128, 256, 512
Learning rate	0.1, 0.01, 0.001, 0.0001
Lambda	Floating point number 0-3
Number of layers	Integer between 2 and 7
Number of units per layer	16, 32, 64, 128, 256
Probability of dropout	0.1, 0.2, 0.25, 0.4, 0.5, 0.6

Table 2. NN Architecture Hyperparameter Options Table from (Hahn & Mechefske, 2022)

4.3. Pipeline Design Uncertainty

For this sensitivity study, the data acquisition pipeline design was varied, considering the following stages and settings, also summarised in table 3.

4.3.1. Dataset

The FEMTO and IMS datasets were chosen due to initial bearing states with no faults; their curated, open source nature; but also their differences in aging methods, timescales and number of recorded failures. In the IMS dataset, the bearings are operated under their maximum specified operating condition limits and failed after their design lifetime (in number of revolutions). This represents scenarios where the bearings are operated in an unhealthy but within technical specification manner. However, as it took weeks to months to observe these failures, only 4 failures over 3 runs were observed, severely limiting the analytic's scope to learn from a diverse sample of run-to-failure trajectories. This issue is reversed for the FEMTO dataset, where 17 distinct failures were observed due to the run-to-failure process taking several hours. However, the conditions the bearings were operated in would not be practical in an industrial setting. Data pipeline choices at this stage investigate the impact on the analytics RUL performance due to the amount and nature of the failures observed, and how the analytics perform on the different methods of accelerated lifetime testing.

4.3.2. Sensor Channel

Both datasets have access to vertically and horizontally aligned vibration sensors (noting limited availability for the IMS dataset). Depending on the nature of the fault, ML models may be more successful in identifying failure signatures in one axis over another, leading to more reliable RUL estimates *if* measurement data is available for this orientation. However, it is not always feasible or maintainable to retrofit assets with extensive sensor coverage, meaning the developing failure may not be measured from the most suitable angle. With no prior knowledge of the bearing failure, data pipeline choices at this stage investigate the consequences on the RUL estimate of having limited, and potentially inadequate, sensor coverage of an impending failure.

4.3.3. Data Sampling

In an ideal scenario, condition monitoring would consist of high resolution, continuous measurement to ensure that as much data is available to the prognostic algorithms as possible. In practice, this would generate enormous volumes of data that would be impractical to transmit, process and store, while potentially providing diminishing returns on the useful information contained in the data streams. Communications and storage infrastructure is limited in an industrial setting where fleets of assets are expected to be monitored simultaneously. At this stage of data pipeline uncertainty assessment, comparisons are made for RUL estimates where 1/8, 1/4, 1/2 and no data is lost due to these constraints.

4.3.4. Spectrogram Bin Count

The spectrogram binning process from (Hahn & Mechefske, 2022) allows the frequency domain information from the full spectrogram to be used while condensing this information into a more manageable number of input features to the ML stage. This forms a trade off between the amount of information lost in the binning process, and the dimensionality. The spectrogram bin count is chosen to be 10, 20 (as original author) and 40, to compare how the RUL is impacted by this trade off.

4.3.5. Hyperparameter Optimisation

NNs are computationally expensive to train, and it may be infeasible to evaluate a large selection of models in order to optimise the selected hyperparamters. Selecting a sub-optimal model will impact the quality of the RUL estimate. The original author runs a parameter search by selecting n combinations of model hyperparameters (table 2), then filtering out models with unsatisfactory performance. Computational limitations may make training many models to allow the most optimal hyperparameters to be chosen an unfeasible action to take. This stage of the pipeline design process investigates the impact on the RUL estimate when the best 10 models are selected from a random search of 10 (90 unique models based on 10 random hyperparameter initialisations for each of the original authors 9 loss functions) and a random search of 100 (900 unique models),

4.3.6. Model Choice

The original author utilises NNs in their study which are black box and computationally expensive. This can undermine the operators trust in the chosen analytic as outputs can not be explained by the model, increasing the risk associated with incorporating model suggestions into decision making processes. Linear Regression (LR) models reside at the other end of the model spectrum as they are cheap to train and simple to understand. However, NNs are able to tackle complex data problems with complicated underlying relationships

Pipeline Stage	Parameter Settings
Dataset	IMS or FEMTO dataset
Sensor channel	Horizontal or Vertical aligned
Subsampling	Lose 1/8, 1/4, 1/2 or no data
Spectrogram Bins	10, 20 or 40 bins
Hyperparam. Opt.	Random search of 10, or 100
Model Choice	NNs or Linear Regression (LR)

Table 3. Summary of pipeline stages and parameters

which cannot be captured by the LR model. In this stage of the pipeline design process, the chosen models are NNs and LR models to compare the RUL prediction between computationally expensive, sophisticated models and interpretable, low computation models.

4.3.7. Evaluating Uncertainty

The original data for each dataset was processed to remove every 8th, 4th or 2nd data point for every datafile in the dataset and resaved; and this process was repeated for each sensor channel. This ensured all combinations of dataset, data sampling and sensor channel were available to train the models. Each model type was trained on all combinations of dataset, data sampling and sensor channel, with the data preprocessed for each selected bin count. For each of these combinations, the NN model hyperparameters were chosen with a random search of 10 or 100, with the model and metrics saved for later processing. The metrics chosen to validate the models were R^2, mean squared error (MSE), RMSE, mean squared log error (MSLE), root mean squared log error (RMSLE), in line with those chosen by (Hahn & Mechefske, 2022). The conditions for successful models to be progressed to the testing stage were a training (and for NN models, validation) performance of $R^2 > 0.2$ and $RMSE < 0.35$, which was applied again after the testing stage to shortlist the top models. To obtain the quantiles, the testing data was run through each of the top models to obtain their RUL predictions, where the 5 %, 25 %, mean, 75 % and 95 % percentiles were calculated for each timestep. The choice of testing data was Run 1, Bearing 4 for IMS and Bearing 1_3 for FEMTO, as the original authors method performed well on these and was decided to be a good point of comparison. This process generated results for all combinations of the 2 datasets, 2 sensor channels, 4 data sampling regimes, 3 spectrogram bin counts, 2 hyperparameter optimisation searches and 2 model choices, resulting in 192 distinct pipeline designs. For each pipeline, the maximum number of models to analyse is the top 10 NNs and a LR model, however not all combinations produced this amount of models that successfully passed the metric bounding criteria.

5. Results

As mentioned in section 4.3.7, the case for comparison between (Hahn & Mechefske, 2022) and this work was Run 1, Bearing 4 testing data from IMS and Bearing 1_3 testing data for FEMTO.

5.1. IMS Results

The RUL prediction shown in figure 2 shows (Hahn & Mechefske, 2022) results for their best performing model on the IMS dataset. This NN model has a Weibull-RMSE Combined loss function, 4 layers with 32 units per layer, 0 % dropout probability, lambda of 0.53, Weibull shape parameter (β) of 2 and characteristic lifetime (η) of 63.9 days. In figure 2, the bearing lifetime extends from 0 % to 100 %, where the jumps are due to the gaps in data collection from the original IMS experiment. The NN predictions are smoothed using a 2 hour rolling average to more clearly demonstrate the trends in the prediction. As shown, the model fits this data well, with a low RMSE score of 0.146, and a high R^2 score of 0.735.

Figure 3 shows the quantiles and mean RUL estimate from the top NN models across all IMS pipelines which met the training and validation metric bounding criteria. The quantiles are calculated on the models performance on Run 1 Bearing 4 testing data from the IMS dataset and the mean of these predictions result in a R^2 of 0.355 and RMSE of 0.228. From approximately 50 % bearing lifetime the quantiles bound the actual lifetime percentage until failure, with the mean fitting the true lifetime percentage well from 60 % lifetime onwards. As shown, the models do not predict early-mid life with any success, which may mislead an operator incorporating the model into a maintenance decision as the model cannot distinguish between any states $< 50\%$ lifetime. Some of this deviation may be explained by the large jumps in lifetime % within the first 10 days of the experiment, compared to the much smoother data collection from day 15 to failure, regardless, this still undermines confidence in the predictions.

The results for the IMS LR models are shown in figure 4. While the quantiles bound the true lifetime from experiment start to end, the lack of incorporated knowledge allows the models to expand to many multiples of bearing lifetime and into negative values. This results in a mean R^2 score of -0.223 and RMSE of 0.314, despite all of the models successfully meeting the R^2 and RMSE bounds in the training stage. Additionally, the RMSE for the testing results is still within (Hahn & Mechefske, 2022) 0.35 boundary while producing unreliable predictions, suggesting other forms of validation are required in tandem to discount unsuitable models. This demonstrates that applying regression models that minimise computational cost or maximise interpretability cannot always perform the required task, and further demonstrates the need for hybrid modelling approaches to incorporate known behaviour.

The pipeline design parameter summary is shown in table 4 which shows the breakdown of pipeline stage parameter counts in the final model selection. The maximum number of models is the top 10 NNs from the 46 IMS NN pipelines,

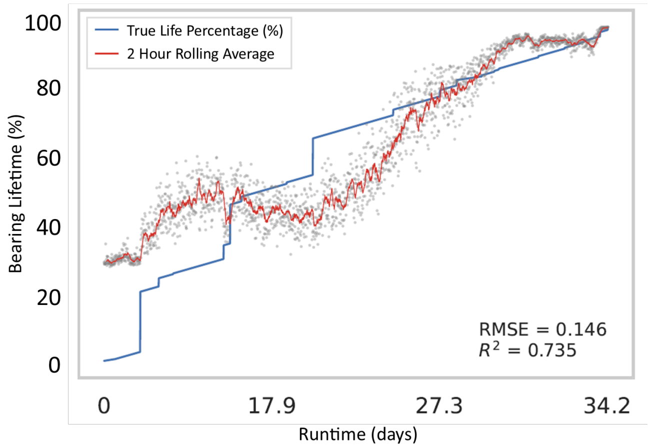


Figure 2. IMS Run 1, Bearing 4 Test Results ((Hahn & Mechefske, 2022)). R^2 = 0.735, RMSE = 0.146

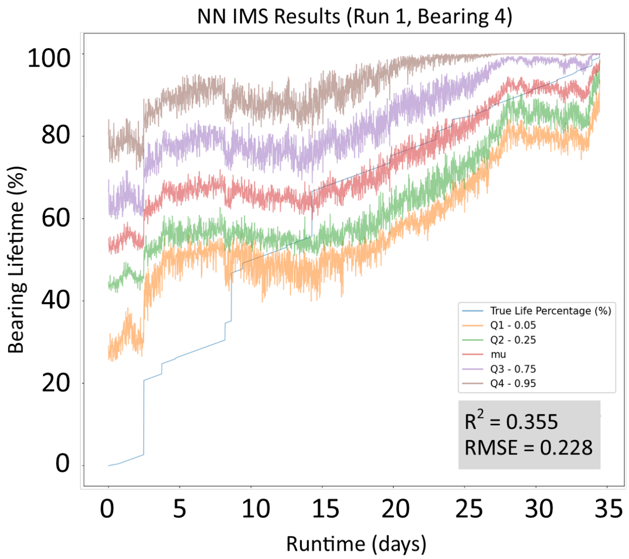


Figure 3. IMS Run 1, Bearing 4 Test Result Uncertainty (NN Model). R^2 = 0.355, RMSE = 0.228

Pipeline Stage	Value	NN	LR
Max Models	NN - 460 LR - 46	367 -	- 18
Sensor	Horizontal Vertical	214 153	9 9
Sampling	Normal - 1/8 - 1/4 - 1/2	101 95 86 85	4 4 4 6
Spec.Bins	10 20 40	126 127 114	2 8 8
HyperParam Search	10 100	137 230	- -

Table 4. Summary of IMS pipeline settings for LR and NN models (by successful model counts)

and single LR model for each of the 46 IMS LR pipelines if all of these models trained successfully. This results in an acceptance rate of 79.8 % for the NNs (367 out of potential 460 models were successful) and only 39.1 % for the LR (18 out of potential 46 models were successful), demonstrating that the NN is more likely to be successful at this prognostic task. For the 18 successful LR models, the sensor alignment choices are split evenly, implying the sensor orientation neither hindered nor helped the models performance, while the NNs tended to favour the horizontal channel as chosen by (Hahn & Mechefske, 2022). Interestingly, for the sampling regime the LR models favoured learning from the least data and fared equally amongst the other options. The NNs were also fairly evenly spread amongst the sampling options, favouring the maximum amount of data. The LR models selected the most condense spectrogram the least, implying the higher dimensional representations provided more useful degrees of freedom to the model. Conversely, the NNs were more evenly spread across the bin options, suggesting all options provided the NNs with enough information. To summarise, it appears that on the IMS dataset, the most influential design parameter was the dimensionality of the input data for the LR models as shown by the aversion to the 10 bin spectrogram, and the time available to optimise the hyperparameters for the NN as this displayed the largest diversion by model contribution in favour of larger number of searches.

5.2. FEMTO Results

The RUL prediction shown in figure 5 shows (Hahn & Mechefske, 2022) results for their best performing model on the FEMTO dataset, with a Weibull only RMSLE loss function, 2 layers with 32 units per layer, 0.25 % dropout probability, lambda of 2.28, Weibull shape parameter (β) of 2 and characteristic lifetime (η) of 4.8 hours. The trend of the predictions is shown by a 2-minute rolling average with straight line from 0 - 100 % demonstrating the bearing lifetime. This NN fits the data well as shown by the low RMSE of 0.133 and high R^2 of 0.788.

The NN FEMTO uncertainty plot is shown in figure 6, which shows the quantiles bounding the whole bearing lifetime, but does not narrow as much as the IMS results at end of life. This larger spread in predictions demonstrates the volatility of the NN predictions on this dataset, as depending on the model, the prediction could be anywhere between 0 and 60 % at start of life and 50-100 % at end of life. The mean prediction has R^2 of 0.729 and RMSE of 0.15 which suggest the mean has a decent fit, however, it can be seen that the models tend to overestimate degradation early-mid life and underestimates mid-end life. If used to inform maintenance schedules, the start of life predictions could result in actions being taken too early where still usable components are prematurely replaced.

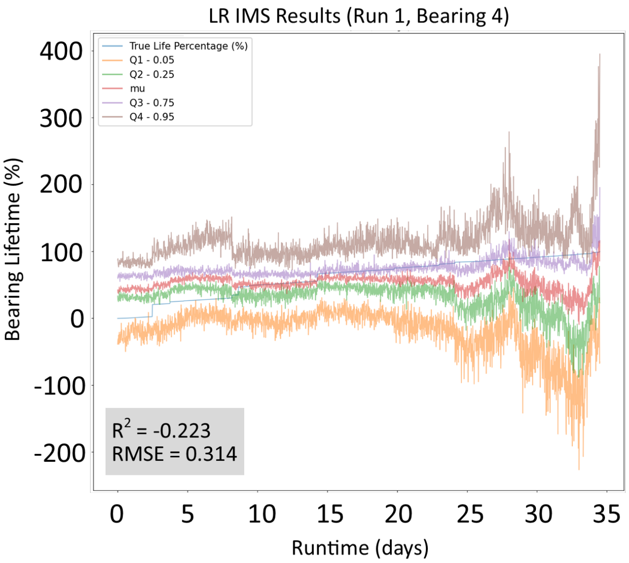


Figure 4. IMS Run 1, Bearing 4 Test Result Uncertainty (LR Model). R^2 = -0.223, RMSE = 0.314

Actions taken based upon the end of life predictions could be left too late, putting operators at risk of unplanned outages.

The fit of the LR models in figure 7 shows consistent estimations early-mid life, then a huge divergence of multiple lifetimes in positive and negative direction is observed in the final stages of the bearing life. This may be due to the rapid decay of the bearings, as the spectrograms show a rapid increase in vibration for some of the training data in the later stages of the experiment. As the end of life prediction is arguably the most crucial aspect of prognostics, these LR models could be considered a risk for any operator to employ in maintenance activities.

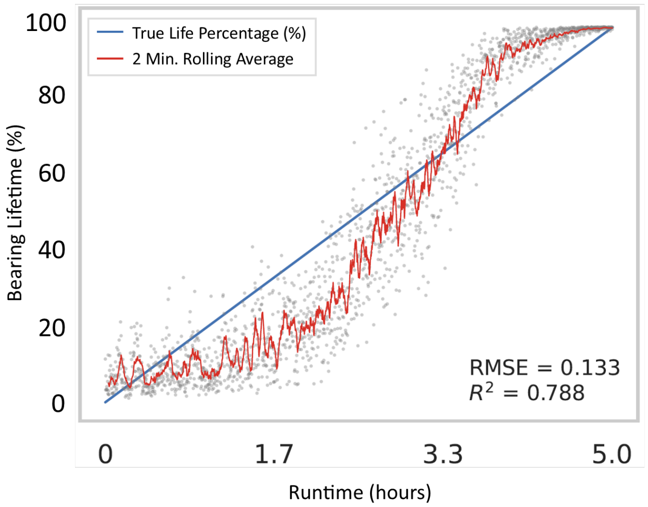


Figure 5. FEMTO Bearing 1_3 Test Results ((Hahn & Mechefske, 2022)). R^2 = 0.788, RMSE = 0.133

Pipeline Stage	Value	NN	LR
Max Models	NN - 460	198	-
	LR - 46	-	24
Sensor	Horizontal	197	12
	Vertical	1	12
Sampling	Normal	45	6
	- 1/8	43	6
	- 1/4	55	6
	- 1/2	55	6
Spec.Bins	10	69	8
	20	63	8
	40	66	8
HyperParam Search	10	77	-
	100	121	-

Table 5. Summary of FEMTO pipeline settings for LR and NN models (by successful model counts)

In the pipeline design summary in table 5, both the NN and LR models have relatively even contributions to the 198 successful NN models and 24 LR models from all settings for the sampling and spectrogram bin options, suggesting these do not have a great influence on the model performance. This is also true for the sensor alignment for the LR models, while for the NN models there is almost entirely self selected horizontal channel, as in (Hahn & Mechefske, 2022). This suggests that the horizontal sensor provides the most useful information for the NN model. Additionally, the NN has strong contribution from the larger hyperparameter search with a majority of models being chosen by the random search of 100. Finally the NN models have an acceptance rate of 43.0 % while the LR models have an acceptance rate of 52.2 %. Interestingly, while the mean NN performance produces better results for R^2 and RMSE, the LR models are more consistently performing above the set metric boundaries and being accepted into the testing stage. Despite their unsuitable design, the choice of metrics and bounds used to assess these models suggest they should be accepted, again suggesting that models require more diverse validation to determine their general suitability, or what situations they may be best suited for. This may also require an appreciation of the similarity of the training and testing data, as models that succeed at the training stage should be trusted to succeed in the testing or online monitoring stage.

This sensitivity analysis has demonstrated that the approach taken to data pipeline definition can have a significant impact on the accuracy of prognostic algorithms, with evidence for a specific bearing vibration case-study provided. This case-study suggests that when developing a data pipeline for this purpose valid models can be selected from a variety of plausible data pipeline configurations while resulting in a diverse range of learned RUL trajectories.

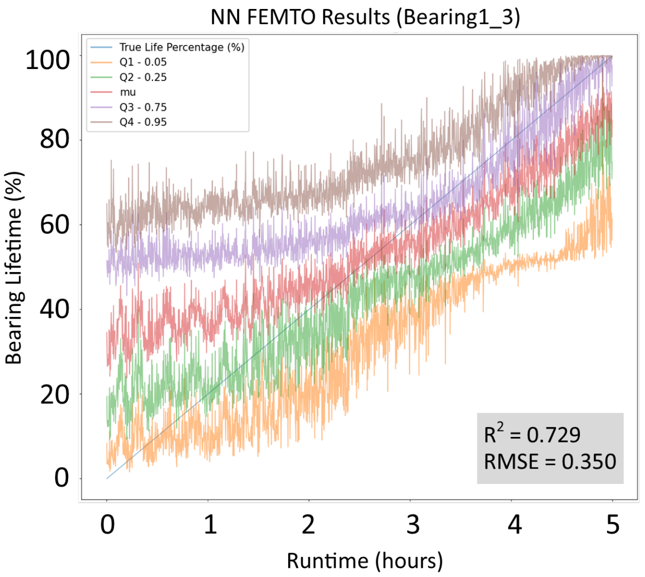


Figure 6. FEMTO Bearing 1_3 Test Result Uncertainty (NN Models). R^2 = 0.729, RMSE = 0.150

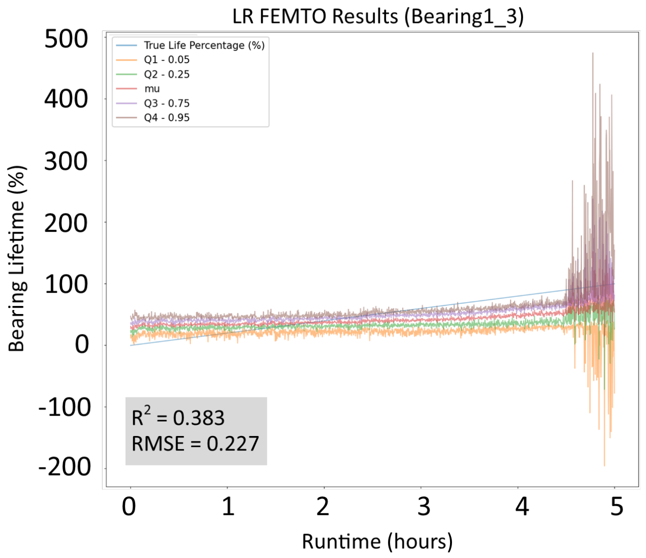


Figure 7. FEMTO Bearing 1_3 Test Result Uncertainty (LR Models). R^2 = 0.383, RMSE = 0.227

6. Conclusion and Future Work

Civil nuclear is a safety critical industry which cannot readily deploy data-driven analytics in decision-making processes without quantification of the uncertainties involved. Consequently, in this work an analysis of the impact of data acquisition pipeline design decisions on the performance of an existing hybrid RUL model for bearing prognostics was conducted. It was shown that the design decisions made at key stages of the data acquisition pipeline can create a large variance of potential RUL trajectories for both NN and LR models on both of the bearing run-to-failure datasets utilised in the study. The models were more sensitive to some design decisions than others, such as the available number of hyperparameter optimisation searches for the NN or the dimensionality of the input features for the LR model (on the IMS dataset). The presence of incompatible design decisions was not suggested by the results as many stages produced an equal number of successful models across the different design options. This suggests that valid models could be generated from completely different pipeline designs, which result in an entirely different learned RUL trajectory. Understanding how the data acquisition pipeline can impact on hybrid prognostic tools can allow nuclear plant operators to justify utilising resources towards reducing high uncertainty areas in the pipeline design to provide more confidence in applying these tools to support maintenance processes. This is of particular concern in the nuclear industry as ML algorithms applied to rotating plant deployed in nuclear engineering environments experience unique operating conditions, such as legacy data acquisition systems that have been upgraded over time without emphasis on the data that will be used for ML purposes.

The models were filtered by a requirement of $R^2 > 0.2$ and $RMSE < 0.35$ to remove unsuitable models before progressing to the testing stage, as in (Hahn & Mechefske, 2022). The results showed that the chosen metrics are not sufficient to definitively identify unsuitable models and are not descriptive enough to show the operator where model application should and should not be trusted. Additionally, the chosen training and testing data may not have been sufficiently comparable for LR type models, as shown by models that had been deemed acceptable in the training stage performing poorly on IMS testing data in figure 4.

To further develop this work, more analysis would be conducted on the impact of metric bias in the model selection process. Models were selected and ranked based on their R^2 and RMSE scores, but a different selection of shortlisted models may have been generated if different metrics had been used or prioritised. Additionally, if it was discovered that some models were more accurate for end of life predictions while other models are more suited for early-mid life, this may not be captured by summary statistics used to qualify the overall model usefulness. Additional methods to describe where the model is successful is needed to further justify the models use for specific prognostic stages, which could be aided by the application of explainability tools. Finally, for a more robust comparison, knowledge would be incorporated into different model types. This would provide more hybrid combinations to compare against, while investigating how model bias impacts the RUL prediction.

References

Abernethy, R. B. (2004). The new weibull handbook : reliability and statistical analysis for predicting life, safety, supportability, risk, cost and warranty claims..

Baur, M., Albertelli, P., & Monno, M. (2020, 03). A review of prognostics and health management of machine tools.. doi: 10.1007/s00170-020-05202-3

Canada Nuclear Safety Commission. (2012, November). *Maintenance programs for nuclear power plants.* Regulatory Document, RD/GD-210. (Online: nuclearsafety.gc.ca)

Coble, J., Ramuhalli, P., Bond, L., Hines, J., & Upadhyaya, B. (2015, 07). A review of prognostics and health management applications in nuclear power plants. *International Journal of Prognostics and Health Management*, *6*, 1-22. doi: 10.36001/ijphm.2015.v6i3.2271

Goebel, K., Eklund, N., & Bonanni, P. (2006, 01). Fusing competing prediction algorithms for prognostics. In (Vol. 2006, p. 10 pp.). doi: 10.1109/AERO.2006.1656116

Hahn, T. V., & Mechefske, C. K. (2022). *Knowledge informed machine learning using a weibull-based loss function.* Journal of Prognostics and Health Management. (Preprint available: https://doi.org/10.48550/arXiv.2201.01769, code available: https://github.com/tvhahn/weibull-knowledge-informed-ml)

Hart, E. (2018). *Wind turbine dynamics identification using gaussian process machine learning* (Unpublished doctoral dissertation). University of Strathclyde.

Jammu, N., & Kankar, P. (2011, 10). A review on prognosis of rolling element bearings. *International Journal of Engineering Science and Technology*, *3*.

JCGM Working Group 1. (2008, 09). *Evaluation of measurement data — Guide to the expression of uncertainty in measurement* (Tech. Rep.). JCGM.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B*, *63*(3), 425–464.

Kumar, S., Mukherjee, D., Guchhait, P., Banerjee, M. R., Srivastava, A. K., Vishwakarma, D., & Saket, R. (2019, 07). A comprehensive review of condition based prognostic maintenance (cbpm) for induction motor. *IEEE Access*, *7*, 90690-90704.

Lee, J. , Qiu, H. , Lin, J. and Rexnord Technical Services. (2007). IMS, University of Cincinnati. "Bearing Data Set", NASA Ames Prognostics Data Repository. (http://ti.arc.nasa.gov/project/prognostic-data-repository, NASA Ames Research Center, Moffett Field, CA)

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. doi: https://doi.org/10.1016/j.ymssp.2017.11.016

Liao, L., & Köttig, F. (2014). Review of hybrid prognostics approaches for remaining useful life prediction of engineered systems, and an application to battery life prediction. *IEEE Transactions on Reliability*, *63*(1), 191-207. doi: 10.1109/TR.2014.2299152

NASA Ames Prognostics Data Repository. (2012). *Femto bearing data set.* NASA Ames Research Center, Moffett Field, CA. (http://ti.arc.nasa.gov/project/prognostic-data-repository)

Nisbet, R., Elder, J., & Miner, G. (2009). Chapter 8 - advanced algorithms for data mining. In R. Nisbet, J. Elder, & G. Miner (Eds.), *Handbook of statistical analysis and data mining applications* (p. 151-172). Boston: Academic Press. doi: https://doi.org/10.1016/B978-0-12-374765-5.00008-5

Ramuhalli, P., Walker, C., Agarwal, V., & Lybeck, N. J. (2020). *Development of prognostic models using plant asset data* (Tech. Rep.). Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States); Idaho National . . . .

Rexnord Industries, LLC, Gear Group. (n.d.). Failure analysis gears-shafts-bearings-seals [Computer software manual].

Yung, C., & Bonnett, A. (2004). Repair or replace? *IEEE Industry Applications Magazine*, *10*(5), 48-58. doi: 10.1109/MIA.2004.1330770

Zhao, X., Kim, J., Warns, K., Wang, X., Ramuhalli, P., Cetiner, S., . . . Golay, M. (2021). Prognostics and health management in nuclear power plants: An updated method-centric review with special focus on data-driven methods. *Frontiers in Energy Research*, *9*. doi: 10.3389/fenrg.2021.696785