Filter-based feature selection for prognostics incorporating cross correlations and failure thresholds

Alexander Löwen[1], Peter Wissbrock[2], Amelie Bender[1], and Walter Sextro[1]

[1] *Paderborn University, Faculty of Mechanical Engineering, Chair of Dynamics and Mechatronics, Paderborn, 33098, Germany*
alexander.loewen@uni-paderborn.de
amelie.bender@uni-paderborn.de
walter.sextro@uni-paderborn.de

[2] *Lenze SE, Innovation Department, Aerzen, 31855, Germany*
peter.wissbrock@lenze.com

ABSTRACT

Historical condition monitoring data from technical systems can be utilized to develop data-driven models for predicting the remaining useful life (RUL) of similar systems, whereas the Health Index (HI) often is a crucial component. The development of robust and accurate models requires meaningful features that reflect the system's degradation process, enabling an accurate prediction of the system's HI. Traditionally, the identification of those is supported by one of various feature ranking methods. In literature, feature interdependencies and their transferability across various similar systems are not sufficiently considered in feature selection, exacerbating the challenge of HI prediction posed by the scarcity of data and system diversity in real-world applications. This work addresses this gaps by demonstrating how filter-based feature selection, incorporating failure thresholds and cross correlations, enhances feature selection leading to improved HI prediction. The proposed methodology is applied to a novel dataset* obtained from run-to-failure experiments on geared motors conducted as part of this study, which presents the aforementioned challenges. It is revealed that classical feature selection, consisting of feature ranking only, leaves potential untapped, which is utilized by the proposed selection methodology. It is shown that the proposed feature selection methodology leads to the best result with a RMSE of 0.14 in predicting the HI of a constructive different gearbox, while the features, determined by classical feature selection, lead to a RMSE of 0.19 at best.

* The dataset called Lenze-GD is accessible via: https://doi.org/10.5281/zenodo.11162448.

1. INTRODUCTION

Early fault detection of mechanical systems like gears and motors is an important topic for industrial production, helping companies to predict equipment failures, reduce downtime and to ensure the reliability and safety of industrial systems. The analysis of data from time series sensors like acoustic, vibration, position, or current is of great interest to monitor the health condition of machines and to predict failure in the mechanical systems life-cycle. The prognostics of the remaining useful lifetime (RUL) aims to predict operating time of a typical operational lifespan that a mechanical system has already passed and estimate the amount of the remaining useful life. In particular, vibration signals have been widely used for RUL-prognostics. However, the usage of signals acquired from inverts like the motor current reduces costs of installation and maintaining external sensors. Under the limitation of a drive system including an induction motor and an inverter with a sufficient data interface, the motor becomes the sensor.

A major challenge in developing accurate and robust RUL-prognostics is the limitation of data, especially in scenarios where abnormal observations are rare or difficult to obtain, referred to as data scarcity. In this study geared motors are focused, which are combinations of toothed-wheel-based gearboxes and of electric induction motors. To match the diverse requirements of customers, the geared motors can be configured and scaled individually. These customized geared motors can be used in a variety of different machine types, which also may be customized. In many real-world problems it is realistic that only a few or none run-to-failure data-collections are available and thus often only data from the healthy motor can be used for model training.

The work is structured as follows. Section 2 presents a comprehensive feature engineering methodology with focus on

feature selection to overcome data scarcity and address system differences. A multi-stage feature selection methodology is described followed by the machine learning (ML) models used and trained based on the selected features. ML algorithms employed are Gaussian Process Regression (GP), Linear Regression (LR), Multi-Layer-Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM). Next, a novel dataset from run-to-failure experiments on geared motors including gear-mesh and bearing failures is introduced in section 3. The experimental setup and the recorded data, which are obtained from a frequency inverter, are described. In section 4, the proposed feature engineering methodology is applied on the new data. In section 5 the advantages over the classical feature selection, consisting of feature ranking only, is shown resulting in the best root mean squared error (RMSE) of 0.14, in contrast to the classical selection's best RMSE of 0.19 in predicting the health index (HI).

2. METHODOLOGY

In this paper, a broadly used workflow for diagnostics and prognostics of technical systems is utilized, which comprises the elements data preprocessing, feature extraction and diagnostics or prognostics algorithm (Goyal, Mongia, & Sehgal, 2021; Ly, Tom, Byington, Patrick, & Vachtsevanos, 2009). Depending on the application, these elements are generally adapted and optimized to suit the circumstances of any given application. The methodology employed prioritizes a more generalized process. To address this limitation, feature engineering is focused wherein a wide range of features are computed, adapted and a multi-stage feature selection process is adopted to select subsequently the most relevant features. Data-driven algorithms are then trained with the selected features within a cross-validation process that includes hyperparameter optimization to predict the HI of the system. These steps are parameterized by means of the systems used for training and then applied to the system used for testing. The whole process is shown in Fig. 1 comprising feature extraction, feature processing, feature selection and model training including hyperparameter optimization, with particular focus on feature selection, whereas Fig. 2 shows the steps from feature processing to correlation analysis with more detail. The steps are described in the following.

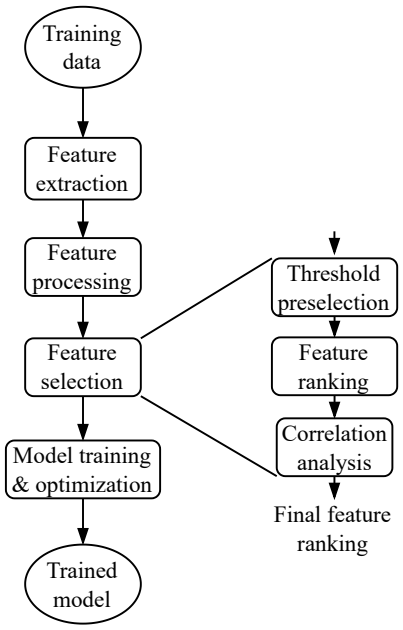

Figure 1. Overview of the applied training process.

2.1. Feature Extraction

Feature extraction is applied to each measurement and channel to extract information regarding system's degradation over time. To address a variety of a system's characteristics, a multitude of features are computed, aiming to encompass a wide range of potential applications where any given feature may capture the system's degradation process. To extract features from time series data, the publicly available Python package tsfresh is used (Christ, Braun, Neuffer, & Kempa-Liehr, 2018). tsfresh is utilized for an automatic extraction of time series features which comprises features from the time, frequency, and time-frequency domains.

In the review of (Goyal et al., 2021), several use cases regarding rotating mechanical systems are consolidated, highlighting the frequent utilization of the fast Fourier transform (FFT) for analysis. This underlines the capacity of FFT to extract information from data, particularly from systems with rotating components, to infer health-related insights. Therefore, additional frequency-dependent features are calculated by dividing the frequency spectrum into sections defined by a constant percentage bandwidth (CPB). Maximum and average FFT coefficients are extracted from the corresponding sections to capture amplitude changes in smaller frequency spectrums. A CPB analysis has been utilized, among other fields, in the field of acoustics (Gram-Hansen, 1991), providing the opportunity to efficiently consider the entire frequency spectrum within comprehensive feature extraction.

2.2. Feature Processing

Feature processing encompasses feature smoothing and feature scaling. Feature smoothing is utilized to reduce noise and variability from the feature data, making underlying patterns and trends more apparent. The moving average is often applied for this purpose. Feature scaling involves scaling the computed features based on the median value of their initial feature data points as shown in Eq. (1). Here, $f_{i,j}$ represents feature i of system j, $f_{i,j,init}$ contains the initial feature points, and $f^*_{i,j}$ denotes the scaled feature data. This process aims to eliminate unwanted influences and facilitate bet-

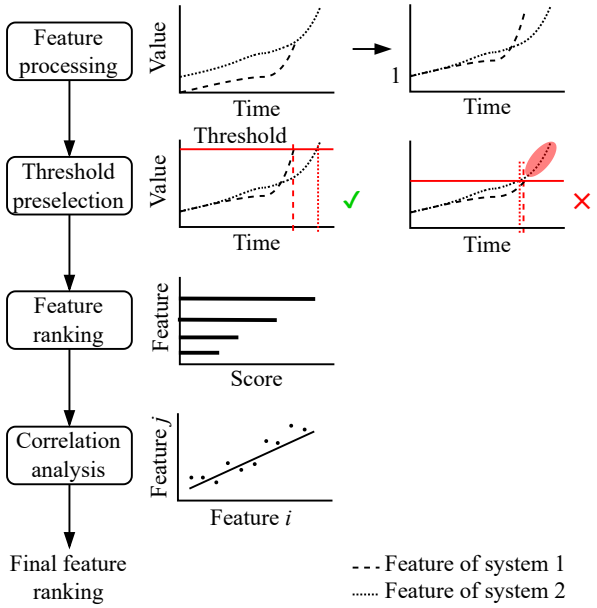

Figure 2. Overview and illustration of the feature processing and selection process.

ter comparability among different systems.

$$f^*_{i,j}(t) = \frac{f_{i,j}(t)}{\text{median}(f_{i,j,init})} \tag{1}$$

2.3. Feature Selection

Feature selection can be divided into filter, wrapper, embedded and hybrid methods (Hoque, Bhattacharyya, & Kalita, 2014) and is necessary for information concentration. This paper focuses on filter methods, as they are less computational intensive in general (Hoque et al., 2014). Feature selection utilized comprises the steps threshold preselection, feature ranking and correlation analysis. Threshold-based preselection retains features with similar failure thresholds and discards those without. It is followed by feature ranking and correlation analyses to remove highly correlated features. The steps are described in the following.

Threshold preselection: Subsequently on feature processing, a preliminary selection is conducted based on a common threshold value for each feature across the systems. Thresholds for system failure are often determined using a predefined HI, often as linear, e.g. in (Yang et al., 2016), or constructed based on selected features, e.g. in (Thoppil, Vasu, & Rao, 2021). In more rare cases, only one feature is directly used if it is sufficient to reflect the degradation process, provided that it can be used to define a system-wide failure threshold, e.g. in (Li, Huang, Gao, Zhao, & Li, 2023; Bender & Sextro, 2021). With limited data, it is difficult to evaluate

Table 1. Metrics considered to determine feature ranking.

Source	Mon.	Trend.	Rob.	Name
(Carino et al., 2015)	×			Spearman
(Nie et al., 2022)	×	×		Cori-Score
(Chen et al., 2019)	×	×	×	MTR_C
(Zhang et al., 2016)	×	×	×	MTR_Z

an individual feature for use as a reliable HI, as well as to construct a HI with respect to a failure threshold. Specific, multiple features that indicate a common threshold across the systems are often not explicitly sought out. To do this, a criterion based on the thresholds for each feature and system is introduced. Firstly, a threshold τ_i regarding Eq. (2) is calculated for each feature i with respect to the systems denoted by j. Here, $\widetilde{f}^*_{i,j,end}$ denotes the median value of the scaled feature points within the final portion, defined by α, of the RUL. The thresholds are reached at different points in the lifetime of each system. If the minimum reached lifetime, as determined by the specified threshold, is below β of the total lifetime of one of the systems, the feature is discarded. An example is given in Fig. 2, where a feature is marked with a cross, signifying that the systems 2 reaches the threshold prematurely, leading to the exclusion of this specific feature.

$$\tau_i = \begin{cases} \min\limits_j(\widetilde{f}^*_{i,j,\text{end}}) & \text{if } \widetilde{f}^*_{i,j,\text{end}} \geq 1 \\ \max\limits_j(\widetilde{f}^*_{i,j,\text{end}}) & \text{if } \widetilde{f}^*_{i,j,\text{end}} < 1 \end{cases} \tag{2}$$

Feature ranking: Feature ranking is crucial in predictive analysis as it allows to identify the most relevant and informative features. Evaluation metrics employed typically encompass assessment of monotonicity and trendability analysis (Carino, Zurita, Delgado, Ortega, & Romero-Troncoso, 2015; Nie, Zhang, Xu, Cai, & Yang, 2022). Moreover, these metrics can be combined with a metric to consider the robustness (Chen, Xu, Wang, & Li, 2019; Zhang, Zhang, & Xu, 2016). A short overview of considered metrics by source to perform feature ranking is given in Tab. 1 and described in the following.

In (Carino et al., 2015) the monotonicity is calculated using the Spearman correlation coefficient, while monotonicity in (Nie et al., 2022; Chen et al., 2019; Zhang et al., 2016) is assessed through the counts of positive and negative derivatives. Trendability is assessed usually through calculating the Pearson correlation coefficient (Chen et al., 2019; Nie et al., 2022; Zhang et al., 2016). Here, (Zhang et al., 2016) used smoothed feature values to encompass monotonicity and trendability, while all others evaluate the original feature data set. The robustness of a feature is assessed through comparison the raw feature values with their smoothed values (Chen et al., 2019; Zhang et al., 2016). The evaluation across multiple considered metrics is conducted using either the average score or the equally weighted sum. In this paper, all of the named fea-

Table 2. Fictional correlation matrix of the best 3 ranked features.

Feature	1	2
1	-	-
2	0.955	-
3	0.892	0.851

ture ranking methodologies are considered to get insight into the potential of the proposed feature selection methodology.

Correlation analysis: Correlation analyses, specifically the Pearson correlation, are often used, besides for feature selection, for similarity analyses (Guo, Li, Jia, Lei, & Lin, 2017; Nie et al., 2022). In this paper, the Pearson correlation is used to determine the similarity between features. Based on the similarity, highly similar features are classified as redundant and discarded, while the best-ranked features are retained. Tab. 2 provides a fictional example showing a correlation matrix for the ranked features 1, 2 and 3. Feature 2 correlates with a coefficient of 0.955 with feature 1. Feature 3 shows correlations of 0.892 and 0.851 with feature 1 and 2 respectively. A parameter can be used to specify which correlation is acceptable. Features that exceed this parameter across all systems are discarded, ensuring that only unique and informative features are retained. If the parameter in the example shown is set to 0.95, feature 2 is discarded, as it exceeds the parameter for feature 1.

2.4. Model training and test

Different ML algorithms are applied and optimized with regard to their hyperparameters applying a Bayesian optimizing algorithm provided by the scikit-optimize library (Head, Kumar, Nahrstaedt, Louppe, & Shcherbatyi, 2021). This technique is based on probabilistic modeling to explore the hyperparameter landscape and find the best parameter combinations (Garnett, 2023). The main goal is not to compare ML algorithms against each other. Instead, the focus lies on assessing the effectiveness and obtaining the best possible prediction result based on the introduced feature selection method. The ML algorithms employed include GP, LR, MLP with one hidden layer consisting of 100 neurons, RF and SVM from the sklearn library (Pedregosa et al., 2011). These algorithms are trained on processed and selected features and are optimized within a cross-validation process to predict a linear HI. The hyperparameter ranges used for optimization are given in the appendix, with standard values employed if a hyperparameter is unspecified.

The predictions are constrained between 0 and 1, where 0 denotes system failure. Evaluation of the models is based on the RMSE as calculated in Eq. (3). Here, $y_{true,i}$ denotes the true and $y_{predicted,i}$ the predicted HI for each observation i of n total observations.

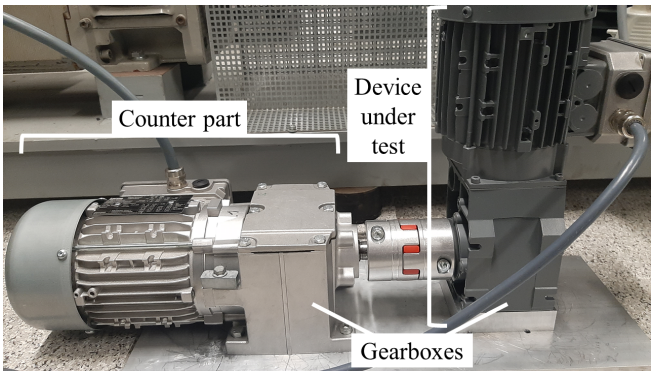

Figure 3. Experimental setup. On the left side, the counter part is shown, which is a helical geared motor. The geared motor on the right side is the device under test, which is a bevel gear.

Table 3. Overview of the gearboxes and their nominal values.

Name	**Type**	**Usage**	**Torque**	**Gear ratio**
H110	Helical	Counter part	110 Nm	28,738
B45	Bevel	Device under test	45 Nm	25,051
H45	Helical	Device under test	35 Nm	10,033

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{true,i} - (y_{predicted,i})^2} \qquad (3)$$

3. Case-Study

To facilitate the presented studies, a run-to-failure experiment for geared motors is introduced. A geared motor is installed in healthy condition and operated until it fails. Throughout the experiment, a data acquisition system is active to monitor the signals of all degradation states. In order to complete the experiment in limited time, the geared motors nominal torque is exceeded. The experiment is conducted three times in total and each with multiple operation states during measurement.

3.1. Experimental Setup

The mechanical part of the setup consist of a first geared motor, the device under test, and a second geared motor, the counter part, shown in Fig. 3. All gearboxes consist of two gear stages with in sum four toothed wheels. The function of the counter part is to create a load for the device under test. An overview of the nominal values of the gears is given in Tab. 3. Thereby the counter part has significant higher nominal torque, to make sure, that the device under test will cause failure, while the counter part stays in healthy condition. The actual torque is selected to lie in the mid of the finite life fatigue of the Woehler characteristic of the second and last gear stage of the device under test to accelerate degradation. During the experiment, the device under test runs with nominal speed.

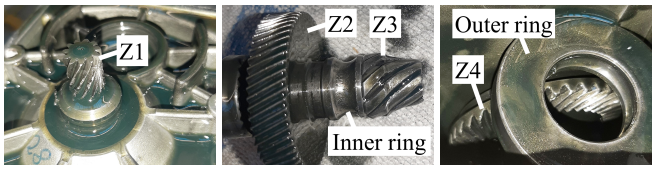

Figure 4. Failure of gearbox B45. Shown are the gears from left to right Z1 with moderate wear, Z2 with minor wear, Z3 with destructive wear and Z4 with moderate wear. As well as the destroyed bearings' inner ring next to Z3 and outer ring next to Z4.

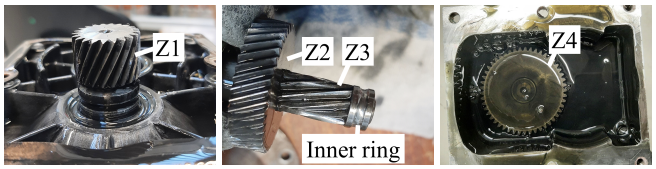

Figure 5. Failure of gearbox H45. Shown are the gears from left to right Z1 and Z2 with minor wear, Z3 with moderate wear and Z4 along with the gearbox full of the deteriorated oil. As well as the destroyed bearings inner ring next to Z3.

In sum, three run-to-failure experiments were conducted, one with B45 gearbox and two with H45 gearbox. In the following, one of the H45 gearboxes will be referred to as H45I and the other as H45II, if they are considered separately. A run-to-failure experiment ends when the gearbox failed, which means its transmission is interrupted. Here, gearbox failure occurred after around 200 hours (H45II) to 790 hours (B45). Subsequently, the gearboxes are opened to evaluate the failures. In the following, the gears are named beginning from the motor-shaft with Z1 transmitting over Z2 to the middle-shaft with Z3 transmitting to the output-shaft over Z4.

The B45 gearbox shows a destructive wear at the Z3, while Z1 and Z4 also show moderate wear, but they stay functional. Z2 only shows minor wear. All of which is shown in Fig. 4. This observation can be explained by the higher torque transmitted by Z3 and Z4 than the first gear stage with Z1 and Z2 and the higher rotation speed of Z1 and Z3 resulting in sum to the high wear of Z3. In addition, the bearing of the middle shaft most close to Z3 is destroyed.

The failure of the runs with H45 shows only minor wear at the gears, except Z3 which shows moderate wear, see Fig. 5. The failures are caused by destroyed bearings next to Z3. Overall, both gear wearing and a destroyed bearing in all cases is observed.

3.2. Data Acquisition

Once per hour, the steady operation of the experiment is interrupted to gather signal measurement of four operation states. These four states are aligned with the nominal values of the induction motor for star connection of the device under test. The states are the combinations of positive or negative nominal speed with nominal or idle torque. Each observation takes measurement with constant sampling rate of 8 kHz and for 2^{15} sampling points, which defines a time period of about 4 s. During this time period, 8 channels are stored in parallel, which are shown in Tab. 4.

Table 4. Overview of the channels acquired by the inverter.

Channel	Type
1	Direct current
2	Quadrature current
3	Effective current
4	Effective voltage
5	Quadrature voltage
6	Phase current U
7	Phase current V
8	Phase current W

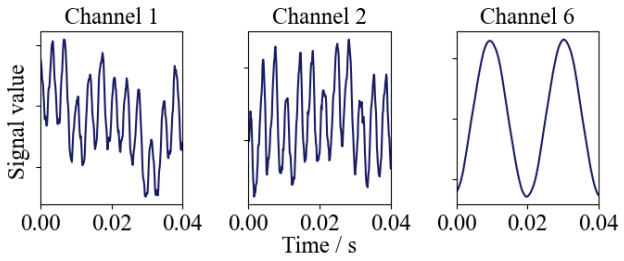

Figure 6. Overview of different derivatives of the current signal. Channel 1: direct current, Channel 2: quadrature current and Channel 6: phase current. All signals are in an internal normalization and thus unit free.

In contrast to vibration and acoustic signals, the original three phase currents are alternating, which may negatively influence some fault detection approaches. Further, the signals of at least two phases would be needed to cover all necessary information. To counter this, also the current in D-Q-coordinates as well as the effective current, calculated by the inverter, are stored, which is briefly shown in Fig. 6. Note that the direct current is related to the magnetic field, while the quadrature current is related to the motors torque. In addition, also the effective and quadrature target voltage are stored, however which are highly quantized and therefore may be of limited relevance.

4. Application

In this section, the presented methodology outlined in section 2 is applied on the gearbox data.

Firstly, only the data recorded at nominal speed in the loading direction with idle torque is considered. Additionally, the running-in process is discarded from the data. A running-in process is particularly well known for gears and causes volatile system behavior in the data shortly after machine commissioning. This can be caused by deforming or breakage of the highest asperities on the tooth surfaces (Feng et al.,

Table 5. Numbered overview of considered feature ranking methods.

No.	Name	Modified
1	Cori-Score	
2	Cori-Score	×
3	MTR_C	
4	MTR_C	×
5	MTR_Z	
6	Spearman	

Table 6. Minimum, maximum and mean value of the average RMSE for prediction on the training dataset within cross-validation across all feature selection variations.

Algorithm	Minimum	Maximum	Mean
GP	1.2e-9	9.8e-9	4.6e-9
LR	0.0956	0.2178	0.1458
MLP	0.0345	0.1220	0.0529
RF	0.0114	0.0314	0.0167
SVM	0.0571	0.0991	0.0662

2019). A running-in process is estimated at 50 hours. Therefore, the initial 50 feature data points, approximately two days of measurements, are rejected. Subsequently, 20,296 features are computed from the 8 channels of each gearbox data. The feature data set is then divided into a training and a test dataset, using the data from the H45 gearboxes for training and the B045 for testing. The target is to select meaningful features using the data from the H45 gearboxes to train ML algorithms, as discussed in section 2.4, and to apply them on the data from the B45 gearbox to finally predict its HI. In the following, the application of the proposed feature processing and selection methodology presented in sections 2.2 and 2.3 is described.

Within feature processing, the feature data undergoes smoothing, where a window size of 15 points seems appropriate. This window size is also used to determine the initial feature data points $f_{i,j,init}$ to scale the data. For threshold preselection, α is set to 1 % and β to 85 %. A small value for α can be selected as the running-in process has been removed. The value for β is chosen to consider the strong and varying increase of feature values towards the end of life. Threshold preselection leads to the exclusion of 19,572 features.

To rank the features, the feature ranking methods discussed in section 2.3 are employed. Due to the positive experience with the Spearman correlation specifically regarding capturing the HI of a system in (Aimiyekagbon, Bender, & Sextro, 2021), an additional version is employed, where the Spearman correlation is used for evaluating the monotonicity. An overview of the feature ranking methods is given by Tab. 5, where the additional versions are marked as modified. In the following, the numbers assigned in Tab. 5 are used as representatives for the mentioned ranking methods.

Lastly, for the correlation analysis to reject highly correlated features, the threshold value for the correlation coefficient is set to 0.98. A high value is chosen to remove strongly correlated features, thereby leave room for selection based on feature ranking. The selected threshold leads to the exclusion of 273 of 724 features. Subsequently, the top 5 ranked features are selected from the remaining 451 features, standardized and utilized for training and testing. A shuffle split with 5 splits is employed for cross-validation, as only two systems are given for training. To ensure the reproducibility of the results, the random seed is fixed. For optimization, 200 iterations are set.

For comparison purposes, additionally, the proposed feature selection methodology is replaced by feature ranking only. Feature selection consisting of feature ranking only represents the classical feature selection process, which is predominantly followed in the literature such as in (Carino et al., 2015; Nie et al., 2022). That means that out of the total of 20,296 features the top 5 ranked ones are used for training allowing a direct comparison with the proposed feature selection methodology.

5. RESULTS

The selected features, results and insights gained from further analysis are discussed in more detail in the following.

When inspecting the selected features, the channels 1, 2 and 3 show a significant higher relevance as they are selected 19, 24 and 10 times of 60 in sum respectively through feature selection. This observation leads to the conclusion that the current in D-Q-coordinates is particularly suitable for predicting the system's condition in contrast to the phase current. As assumed, the effective and quadrature voltage is of minor importance. Further, it can be observed that abrupt changes, reversed direction of feature progression and large differences in the endpoints between same features of the train and the test set cause confusion in prediction.

The minimal, maximal and mean RMSE of the predictions on the training data within cross validation is shown in Tab. 6 and Fig. 7 presents the prediction errors from predicting the HI of the gearbox B045. Primarily, all algorithms show low error values on the training dataset, which in combination with the results in Fig. 7 indicates, that some algorithms generalize better (RF) than other (MLP). MLP and SVM generate the highest RMSE, probably increased by the small amount of training data. SVM performs better evaluating the selected features from feature ranking only, although the predictions get particularly worse towards the end of life. The full potential of the MLP may not be exploited, as the iterations during its training and optimization are both limited to 200. In addition, the layer size and depth is not varied during optimization.

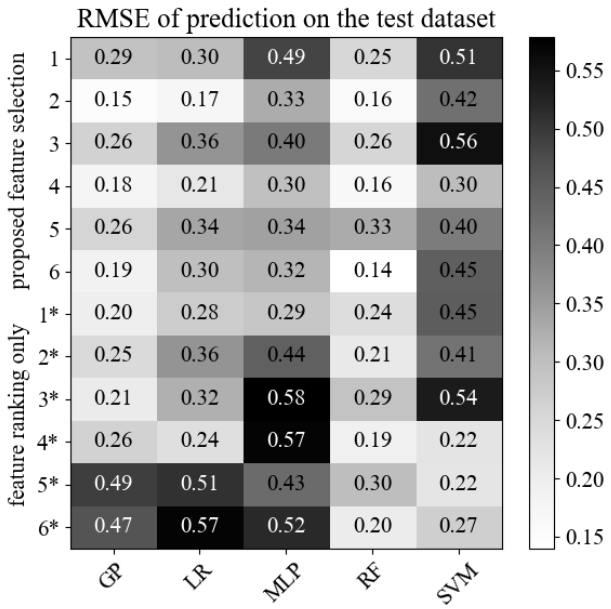

Figure 7. Results based on the estimation of the HI of gearbox B45, which provides the test data. The results are marked with an asterisk "*" when classical feature selection is applied and without when the proposed feature selection methodology is utilized.

The best performing algorithms are GP and above all RF. Especially noticeable is that the proposed selection method performs most effectively in combination with Spearman correlation for ranking the features as can be seen in rows 2, 4, and 6. To summarize, the best RMSE is reduced from 0.19 to 0.14 by around 26 % when the proposed feature selection method is utilized. The worst RMSE is reduced from 0.58 to 0.56 by around 3 %, whereby results with an RMSE of 0.5 and higher occur 6 versus 2 times and an RMSE of 0.4 and higher occurs 12 versus 7 times. Results with an RMSE of 0.2 and lower appear 2 versus 7 times. This indicates a higher robustness capabality using the proposed selection method.

The RF achieves the best result with an RMSE of 0.14 in row 6, where feature ranking within the proposed feature selection methodology is applied by assessing the Spearman correlation. The hyperparameters of the RF set by the hyperparameter optimization are given in Tab. 7 including a brief description. The selected features are shown in Fig. 8 where the feature values are plotted over the HI. The features are briefly described in Tab. 8. For detailed information on feature calculations, reference is made to the official documentation of tsfresh (Christ, Maximilian and Braun, Nils and Neuffer, Julius, 2016).

The prediction of the HI for the B45 gearbox data, generated by the RF trained on the presented features, is visualized in Fig. 9. The horizontal axis represents the actual HI values, while the vertical axis represents the predicted HI values. The diagonal line running from (1,1) to (0,0) represents the ideal prediction. The prediction of the test system shows a certain variance of the points, especially in the ranges 0.9 to 0.5 and 0.3 to 0.1 of the actual HI. The underestimated HI in the range 0.9 to 0.5 can be explained by the stronger gradient observed for the features 1 and 5. The overestimated HI in the range 0.3 to 0.1 is possibly caused by feature 2.

Table 7. Hyperparameter values and descriptions for the RF model set by the optimization algorithm.

Hyperparameter	Value	Description
n_estimators	149	Number of decision trees in the ensemble.
max_features	sqrt	Maximum number of features used to determine the best split. Here it is the square root of the number of features.
max_depth	27	Maximum depth of a single decision tree.
min_samples_split	1e-6	Minimum number of observations required to split a node in the decision trees. This number is defined by a fraction of the total number of observations.
min_samples_leaf	1e-6	Minimum number of observations required to form a leaf node. This number is defined by a fraction of the total number of observations.

Despite the observed variability, the prediction is deemed satisfactory considering the limited availability of training data and the structural differences between the systems for training and testing. The results presented underscore the ability of the proposed feature selection methodology in capturing the differences between the systems, especially in combination with the RF. Although certain challenges persist and continue to impact the overall results, the better results tend to align with the utilization of the proposed feature selection methodology, particularly shown in the upper half of the color map in Fig. 7.

6. Conclusion and Future Work

The effective use of available data is crucial, especially in scenarios characterized by data scarcity. The optimal use of available information is essential to improve the accuracy and reliability of prognostics and ensure efficient decision-making and resource allocation in the industry.

To tackle this challenge, comprehensive feature engineering with focus on feature selection is adopted, wherein the features are adapted to their initial values by scaling and feature selection is performed involving several successive steps. These steps encompass threshold-based preselection, feature ranking and cross-correlation analysis. Subsequently, training of ML-based models is conducted to predict the HI of the

Table 8. Description of the selected features obtained through the proposed feature selection methodology, wherein Spearman correlation was utilized for feature ranking.

Feature	Description
1	Value of the evaluated partial autocorrelation function at lag 6 of the quadrature current signal.
2	The highest order coefficient of a polynomial function the order 3 derived from the deterministic dynamics of the Langevin model, where 30 quantiles are used for averaging, based on the direct current data.
3	Complexity calculated by the Lempel-Ziv compression algorithm in the direct current data divided into 100 bins.
4	Custom feature explained in section 2.1, where direct current data is used. Bin 78 represents a frequency range from 26.12 to 26.86 Hz, where the FFT coefficients were aggregated by the mean.
5	The feature quantifies the maximum standard error of the linear trend over sections of length 5 in the direct current data.

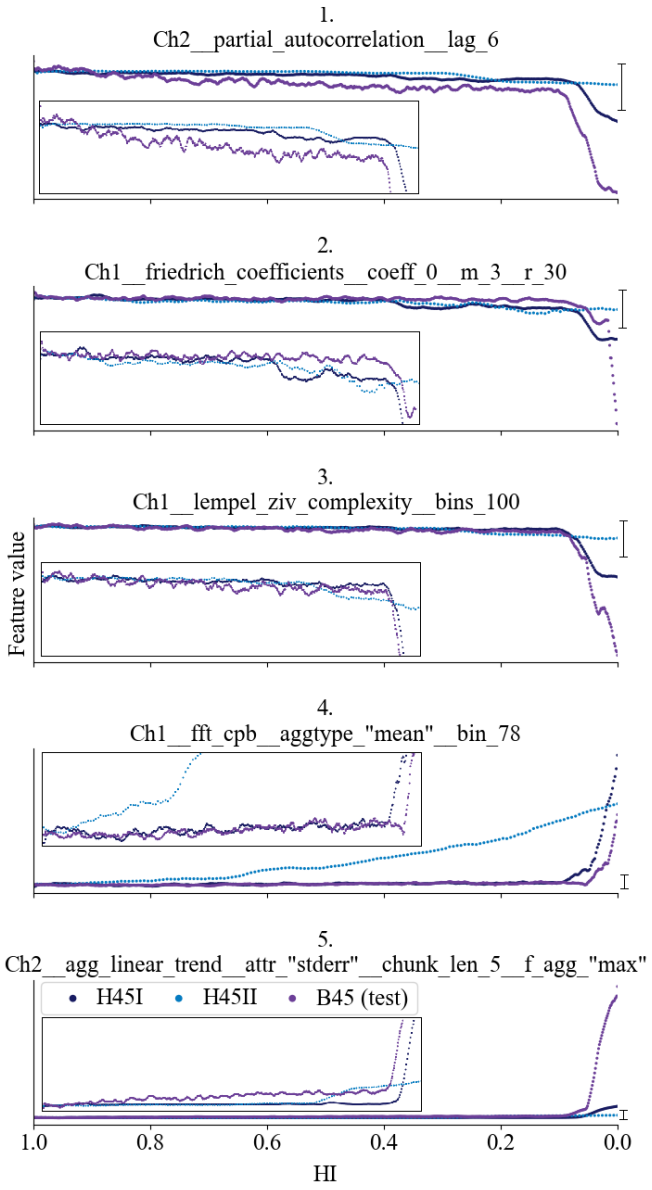

Figure 8. Best 5 features assessed by the proposed feature selection methodology based on the feature data from the H45 gearboxes, wherein the Spearman correlation was utilized for feature ranking. The boxes added indicate zoomed-in views of the features. The range is marked on the right edge. The feature values are unit free as they have been scaled.

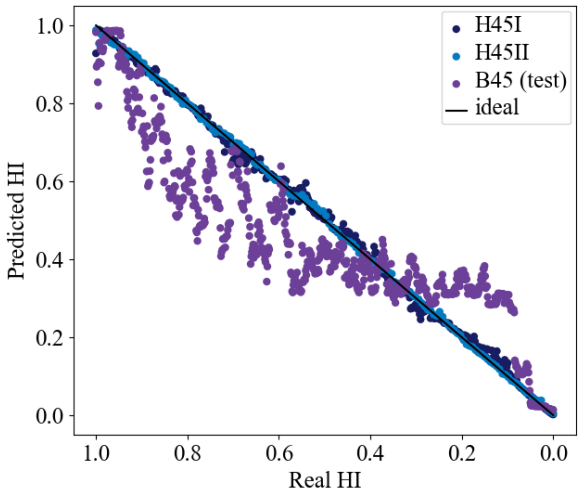

Figure 9. Test results generated by estimating the HI of the gearboxes H45I, H45II and B45, with the H45 gearboxes providing the training data and B45 gearbox the test data.

systems.

In order to evaluate the proposed methodology, a new dataset is introduced and utilized, which contains current, voltage and phase current data from run-to-failure experiments of gearboxes. The dataset is notable for considering two structurally different gearboxes and for addressing the challenge of data scarcity, as it is sourced from only three systems. The aim is to use the data from the two similar gearboxes to estimate and select features to infer the HI of the dissimilar gearbox over its entire operating time based on a ML algorithm. By publishing the novel dataset, other researchers are inspired to contribute to this specific problem setting.

It is observed that, the classical feature selection is able to select features capturing the degradation of the systems in some cases leading to an RMSE of 0.19 in the best case. However, the proposed feature selection methodology apparently supports overcoming system differences especially in combination with the RF by selecting appropriate features better leading to the best result overall with an RMSE of 0.14. Therefore, a great potential in applying the proposed methodology

to further problems in the field RUL-estimation is seen. It can enable more effective training based on ML training, as features are selected not only based on capturing the degradation of individual systems separately but also considering a common threshold for failure while avoiding redundancies.

The next research steps will include validation of the proposed methodology using further data or different test conditions to check the limitations, reliability and robustness of the results. The exploration of alternative methods, such as mutual information, should also be considered at the last step of the proposed feature selection process to replace the Pearson correlation analysis. These methods have the potential to enhance the methodology. Furthermore, the applicability of the proposed method to different types of gearboxes or even to other technical systems should be explored. This would contribute to demonstrating the scope and versatility of the proposed approach.

ACKNOWLEDGMENT

This research and development project is funded by the Ministry of Economy, industry, climate action and energy of the State of North Rhine-Westphalia (MWIKE) in the context of the Leading-Edge Cluster ‚Intelligent Technical Systems OstWestfalenLippe (it's OWL)' and supervised by Project Management Jülich (PtJ). The responsibility for the content of this publication lies with the author.

REFERENCES

Aimiyekagbon, O. K., Bender, A., & Sextro, W. (2021). On the applicability of time series features as health indicators for technical systems operating under varying conditions. *17. International Conference on Condition Monitoring and Asset Management (CM 2021)*.

Bender, A., & Sextro, W. (2021). Hybrid prediction method for remaining useful lifetime estimation considering uncertainties. *PHM Society European Conference*, *6*(1), 11. doi: 10.36001/phme.2021.v6i1.2843

Carino, J. A., Zurita, D., Delgado, M., Ortega, J. A., & Romero-Troncoso, R. J. (2015). Remaining useful life estimation of ball bearings by means of monotonic score calibration. In (pp. 1752–1758). doi: 10.1109/ICIT.2015.7125351

Chen, C., Xu, T., Wang, G., & Li, B. (2019). Railway turnout system rul prediction based on feature fusion and genetic programming. In (Vol. 151, p. 107162). doi: 10.1016/j.measurement.2019.107162

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

Christ, Maximilian and Braun, Nils and Neuffer, Julius. (2016). *tsfresh.* `https://tsfresh.readthedocs.io`. (Accessed: March 07, 2024)

Feng, P., Borghesani, P., Chang, H., Smith, W. A., Randall, R. B., & Peng, Z. (2019). Monitoring gear surface degradation using cyclostationarity of acoustic emission. *Mechanical Systems and Signal Processing*, *131*, 199–221. doi: 10.1016/j.ymssp.2019.05.055

Garnett, R. (2023). *Bayesian Optimization*. Cambridge University Press.

Goyal, D., Mongia, C., & Sehgal, S. (2021). Applications of digital signal processing in monitoring machining processes and rotary components: A review. In (Vol. 21, pp. 8780–8804). doi: 10.1109/JSEN.2021.3050718

Gram-Hansen, K. (1991). A bandwidth concept for cpb time-frequency analysis. In *[proceedings] icassp 91: 1991 international conference on acoustics, speech, and signal processing* (p. 2033-2036 vol.3). doi: 10.1109/ICASSP.1991.150803

Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017). A recurrent neural network based health indicator for remaining useful life prediction of bearings. *Neurocomputing*, *240*, 98–109. doi: 10.1016/j.neucom.2017.02.045

Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., & Shcherbatyi, I. (2021, October). *scikit-optimize: Sequential model-based optimization in python.* `https://zenodo.org/records/5565057`. Zenodo. (Last accessed: May 5, 2024)

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, *41*(14), 6371–6385. doi: 10.1016/j.eswa.2014.04.019

Li, Y., Huang, X., Gao, T., Zhao, C., & Li, S. (2023). A wiener-based remaining useful life prediction method with multiple degradation patterns. *Advanced Engineering Informatics*, *57*, 102066. doi: 10.1016/j.aei.2023.102066

Ly, C., Tom, K., Byington, C. S., Patrick, R., & Vachtsevanos, G. J. (2009). Fault diagnosis and failure prognosis for engineering systems: A global perspective. In (pp. 108–115). doi: 10.1109/COASE.2009.5234094

Nie, L., Zhang, L., Xu, S., Cai, W., & Yang, H. (2022). Remaining useful life prediction for rolling bearings based on similarity feature fusion and convolutional neural network. In (Vol. 44). doi: 10.1007/s40430-022-03638-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Thoppil, N. M., Vasu, V., & Rao, C. S. P. (2021). Health indicator construction and remaining useful life estimation for mechanical systems using vibration signal prognostics. *International Journal of System Assurance En-*

Table 9. Hyperparameter ranges for optimizing GP

Hyperparameter	Range
kernel	RBF, DotProduct, Matern, RationalQuadratic, WhiteKernel

Table 10. Hyperparameter ranges for optimizing MLP

Hyperparameter	Range	Distribution
activation	identity, logistic, tanh, relu	
solver	lbfgs, adam	
alpha	[1e-6, 1]	uniform
learning_rate	constant, adaptive, invscaling	
learning_rate_init	[1e-6, 1]	uniform

gineering and Management, *12*(5), 1001–1010. doi: 10.1007/s13198-021-01190-z

Yang, F., Habibullah, M. S., Zhang, T., Xu, Z., Lim, P., & Nadarajan, S. (2016). Health index-based prognostics for remaining useful life predictions in electrical machines. *IEEE Transactions on Industrial Electronics*, *63*(4), 2633–2644. doi: 10.1109/TIE.2016.2515054

Zhang, B., Zhang, L., & Xu, J. (2016). Degradation feature selection for remaining useful life prediction of rolling element bearings. In (Vol. 32, pp. 547–554). doi: 10.1002/qre.1771

BIOGRAPHIES

Alexander Loewen obtained his BSc and MSc in mechanical engineering at the Paderborn University, Germany. Since 2022, he is part of the Chair for Dynamics and Mechatronics at the Paderborn University. His research focuses on the automated training of ML based models for technical systems, particularly in anomaly detection, classification, and regression tasks under stationary operating conditions.

Peter Wissbrock is currently working as a researcher at the Innovation Department of Lenze SE. He obtained his BSc and MSc in electrical engineering at the OWL University of Applied Sciences and Arts, Germany. He is pursuing Ph.D. from Leibniz University Hannover, Germany in field of industrial analytics. He works in domain of drive train fault diagnosis, machinery anomaly detection and data acquisition infrastructure.

Amelie Bender studied mechanical engineering at RWTH Aachen University, Germany, and one semester abroad at the University of Newcastle, Australia. Since 2015 she is with the research group Dynamics and Mechatronics at Paderborn University, Germany. During her doctoral studies in mechanical engineering, her research focusses on condition monitoring of rubber-metal-bearings. She was awarded the academic degree Dr.-Ing. in 2021. As a team leader at the research group Dynamics and Mechatronics at Paderborn University, her research covers the topics condition monitoring, data analytics and reliability engineering.

Walter Sextro studied mechanical engineering at the Leibniz University of Hanover and at the Imperial College in London. Afterwards, he was development engineer at Baker Hughes Inteq in Celle, Germany and Houston, Texas. Back as research assistant at the University of Hanover he was awarded the academic degree Dr.-Ing. in 1997. Afterwards, he habilitated in the domain of mechanics under the topic Dynamical contact problems with friction: Models, Methods, Experiments and Applications. From 2004-2009 he was professor for mechanical engineering at the Technical University of Graz, Austria. Since March 2009 he is professor for mechanical engineering and head of the research group Dynamics and Mechatronics at the University of Paderborn.

APPENDIX

In Tabs. 9 to 12 the hyperparameter ranges are listed which where utilized for hyperparameter optimization of the regarding algorithm. For detailed information on hyperparameters, reference is made to the official documentation of scikit-learn (Pedregosa et al., 2011).

Table 11. Hyperparameter ranges for optimizing RF

Hyperparameter	Range	Distribution
n_estimators	[1, 200]	uniform
max_features	None, sqrt, log2	
max_depth	[1, 32]	uniform
min_samples_split	[1e-6, 1]	uniform
min_samples_leaf	[1e-6, 1]	uniform

Table 12. Hyperparameter ranges for optimizing SVM

Hyperparameter	Range	Distribution
C	[1e-2, 1e+3]	log-uniform
gamma	[1e-4, 1e+1]	log-uniform
kernel	linear, rbf	