A Preprocessing and Modeling Approach for Gearbox Pitting Severity Prediction under Unseen Operating Conditions and Fault Severities.

Rik Vaerenberg[1,2], Douw Marx[1,2], Seyed Ali Hosseinli[1,2], Fabrizio De Fabritiis[1,2], Hao Wen[1,2], Rui Zhu[1,2], and Konstantinos Gryllias[1,2,3]

[1] *Department of Mechanical Engineering, Division LMSD, KU Leuven,*
Celestijnenlaan 300, Box 2420, 3001 Leuven, Belgium
rik.vaerenberg@kuleuven.be
konstantinos.gryllias@kuleuven.be

[2] *Flanders Make@KU Leuven, Belgium*

[3] *Leuven.AI - KU Leuven Institute for AI, B-3000 Leuven, Belgium*

ABSTRACT

Gear pitting is a common gearbox failure mode that can lead to unplanned machine downtime, inefficient power transmission and a higher risk of sudden catastrophic failure. Consequently, there is strong incentive to create machine learning models that are capable of detecting and quantifying the severity of gearbox pitting faults. The performance of machine learning models is however highly dependent on the availability of training data and since training data for a wide variety of different operating conditions and fault severities is rarely available in practice, machine learning models must be designed to be robust to unseen operating conditions and fault severities. Furthermore, models should be capable of identifying data outside of the training data distribution and adjusting the confidence in a prediction accordingly. This work presents a strategy for pitting severity estimation in gearboxes under unseen operating conditions and fault severities in response to the PHM North America 2023 Conference Data Challenge. The strategy includes the design of dedicated validation sets for quantifying model performance on unseen data, an investigation into the most appropriate preprocessing methods, and a specialized convolutional neural network with an integrated out-of-distribution detection model for identifying samples from foreign operating conditions and fault severities. The results show that the best models are capable of some generalization to unseen operating conditions, but the generalization to unseen pitting severities is more challenging.

1. INTRODUCTION

This section presents a brief overview of prior work in pitting severity estimation in gearboxes, describes the pitting severity dataset of PHM2023 data challenge and provides a short overview of the contents of the rest of the article.

1.1. Pitting Severity Estimation in Gearboxes

Gear transmission systems play an important role in rotating machinery, which are used in various industrial applications, ranging from high-power wind turbines (Salameh, Cauet, Etien, Sakout, & Rambault, 2018) to aircrafts (Cartocci, Napolitano, Costante, Valigi, & Fravolini, 2022).

To ensure the continuous and reliable operation of these machines, condition based maintenance practices have been developed to reduce unnecessary maintenance (Lei et al., 2018), and reduce machine downtime (Lee, Wu, Zhao, Ghaffari, & Liao, 2014). To this end, various methods for detecting and quantifying the severity of gear pitting deterioration have been studied in recent years.

Traditional methods for vibration-based condition-based maintenance that rely on signal processing have proven to be very

https://doi.org/10.36001/IJPHM.2024.v15i1.3808

successful in this task (Elasha et al., 2014; Teng, Wang, Zhang, Liu, & Ding, 2014) (Öztürk, Sabuncu, & Yesilyurt, 2008). However, the outputs of these signal processing techniques often require interpretation by experienced operators.

As a competing alternative to these traditional methods based on signal processing, machine learning (ML) techniques have shown promising results for pitting fault detection and severity quantification in recent years (Chen, Li, & Sanchez, 2015; Kumar, Parey, & Kankar, 2023) (Qu, He, Deutsch, & He, 2017; Medina et al., 2019). By processing large amounts of data collected from machines in healthy and faulty conditions, ML algorithms can identify hidden fault patterns in data that does not require manual feature engineering as required in signal processing methods (Hendriks, Dumond, & Knox, 2022). In particular, Convolutional Neural Networks (CNN) (LeCun et al., 1989), have demonstrated remarkable versatility across a broad spectrum of applications of condition monitoring, including the detection and diagnosis of gear pitting faults. (Li, Li, Zhao, Qu, & He, 2020) proposed 1D separable convolution with a residual connection network to diagnose gear pitting, achieving high classification accuracy on different operating speeds. In another work, a model that combines a CNN with gated recurrent units effectively identified gear defects using acoustic emission and lateral vibration data (Li, Li, Qu, & He, 2019). Additionally, CNN's frequently serve as foundational elements for more sophisticated network designs, enabling the creation of intricate models for diverse challenges. For example, (Qin, Wang, & Xi, 2022) utilizes a CNN as the underlying structure of CycleGAN to identify gear pitting through image data.

Despite the success of ML approaches, the effectiveness of supervised machine learning models as applied in condition monitoring is highly dependent on the availability of data at different operating conditions and fault severities, since these data driven methods assume that labeled training samples are available (Schmidt & Heyns, 2019). Specifically, data-driven supervised machine learning models have trouble generalizing to data distributions that are different from the data they were trained on (Shimodaira, 2000) (Lakshminarayanan, Pritzel, & Blundell, 2017) (Louizos & Welling, 2017). This distribution shift is present when the model encounters data from unseen operating conditions and fault severities. Therefore, it has been proposed to use Out-of-Distribution (OOD) detectors when deploying machine learning models to flag samples out of the training distribution (Bishop, 1994).

In this work, the strengths of traditional signal processing methodologies are combined with CNN's for the task of predicting pitting severity from vibration data. The presented approach aims to evaluate model performance and the generalization capabilities of the machine learning models to unseen operating speeds and fault severities. Ultimately, this rigorous evaluation enables the selection of an optimal combination of signal processing and machine learning methodologies.

1.2. Description of PHM2023 Data Challenge Dataset and Data Exploration

This investigation into gearbox pitting severity estimation is conducted as part of the PHM North America 2023 Data challenge (Prognostics and Health Management Society, 2023). Participating teams are required to diagnose the pitting fault severity of a gearbox from a three-axis vibration measurement.

1.2.1. Dataset Description and Visualisation

The dataset comprises of vibration signals from gear pitting experiments of increasing severity conducted on a one-stage gearbox test rig. The gearbox has spur gears with a speed reduction ratio of 1.8:1, containing a driving gear with 40 teeth and a driven gear with 72 teeth. The pitting severity of the gears was artificially increased by manual drilling operations on the gear tooth faces without any disassembly of the gearbox between consecutive tests (Prognostics and Health Management Society, 2023). Measurements were collected under various operating conditions and pitting severities including the gearbox in healthy condition. The operating conditions include speeds from 100 to 2000 rpm and torque levels from 200 to 900 Nm. During measurement, longer time series signals were collected for lower rotational speed conditions to ensure sufficient data points per shaft rotation. Therefore, the signal duration varies, with approximately 12 seconds signals for speeds of 100-200 rpm, 6 seconds signals for speeds of 300-1000 rpm, and 3 seconds for speeds exceeding 1200 rpm. The vibration signals were sampled at a rate of 20480 Hz, and included a total of 3 measured channels for the horizontal, axial, and vertical accelerations respectively.

The data challenge imposed a dataset split to evaluate participants, with the training set comprising only 78 operating conditions across 7 pitting severity levels even though the test set contained data from 81 operating conditions and 11 pitting severity levels. This means that some pitting severity levels (5, 7, 9, 10) and operating conditions (1500RPM, 1800RPM, 2400RPM) are excluded from the training set, thereby requiring models to generalize to unseen operational conditions and fault levels. The data available in the training set is shown in Table 1.

As a further introduction to the dataset, the time series, the Power Spectral Density (PSD) (Welch, 1967) and the first principal components of the PSD are visualized for different fault severities and operating conditions. From the time series plots of the training data set in Figure 1, it is clear that pitting severity cannot easily be diagnosed from the magnitude of the raw time series data alone. For instance, increasing pitting severity does not necessarily imply larger peak-to-peak values as shown for the 100rpm-300N and 600rpm-50N data

Table 1. Splitting of training data for different operating conditions. Operating conditions present in the training data are indicated with ✓. Operating conditions only present in the test set are indicated with ✗. No data is available for blank cells.

Speed (RPM)	**Torque (Nm)**					
	50	100	200	300	400	500
100	✓	✓	✓	✓	✓	✓
200	✓	✓	✓	✓	✓	✓
300	✓	✓	✓	✓	✓	✓
400	✓	✓	✓	✓	✓	✓
500	✓	✓	✓	✓	✓	✓
600	✓	✓	✓	✓	✓	✓
700	✓	✓	✓	✓	✓	✓
800	✓	✓	✓	✓	✓	✓
900	✓	✓	✓	✓	✓	✓
1000	✓	✓	✓	✓	✓	✓
1200	✓	✓	✓	✓	✓	✓
1500	✗	✗	✗	✗	✗	✗
1800	✗	✗	✗	✗	✗	
2100	✓	✓	✓	✓		
2400	✗	✗	✗	✗		
2700	✓	✓	✓			
3000	✓	✓	✓			
3600	✓	✓				

in Figure 1.

The PSD of the vibration signal at different loads and operating conditions are shown in Figure 2. The subplots show a comparison of different operating torques for a healthy condition, different operating torques for a faulty condition, and different fault conditions for a fixed operating condition, respectively. While a relation between the peak value in the PSD and load at a fixed speed can be observed for healthy samples (Figure 2a), the same pattern is not identifiable for faulty samples (Figure 2b). Finally, in the PSD of the vibration signal for different pitting severities under a fixed operating condition (Figure 2c), there is also no clear relation visible between the peak value and the pitting severity.

The challenge of separating the different pitting severity states is further made apparent in Figure 3 showing the first two principal components of training data PSDs after a Box-Cox transform (Box & Cox, 1964). There tends to be an overlap between fault classes, especially at high speeds. Furthermore, similar pitting severities do not necessarily lie in the same vicinity in the first two principal components.

1.2.2. Competition Scoring Metrics

The 2023 PHM North America Data challenge competition used a dedicated scoring system that rewarded correct predictions and severely penalized predictions that are far from the true pitting severity. Hereby the competition participants are encouraged to create models that do not make incorrect predictions confidently. The score per observation is defined as:

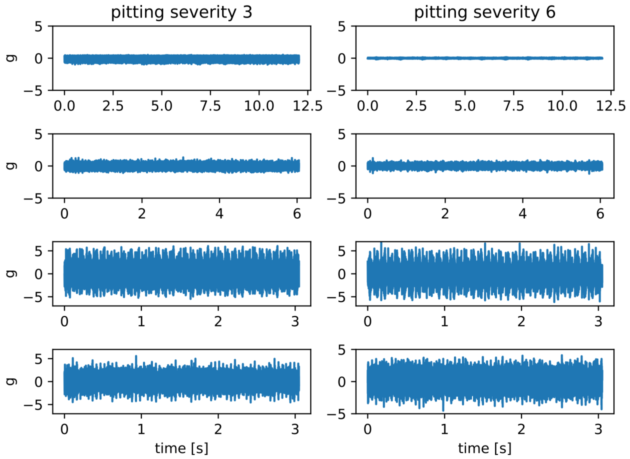


Figure 1. Vibration measurements at different speeds and loads for pitting severity 3 and 6 in the training set. From top to bottom: 100rpm-300N, 600rpm-50N, 1200rpm-500N, 2100rpm-100N.

$$\text{Score}_{\text{ observation}} \coloneqq \text{c} \sum_{i=0}^{10} \text{p}_i \, \text{s}_i \tag{1}$$

where c $\in \{0.2, 1\}$ is a confidence factor of the prediction and $\text{p}_i \in [0, 1]$ is the probability of the sample belonging to pitting severity i as predicted by the model. The pitting severity score s_i is the reward or penalization based on the distance between pitting severity i and the true state of the observation according to Table 2.

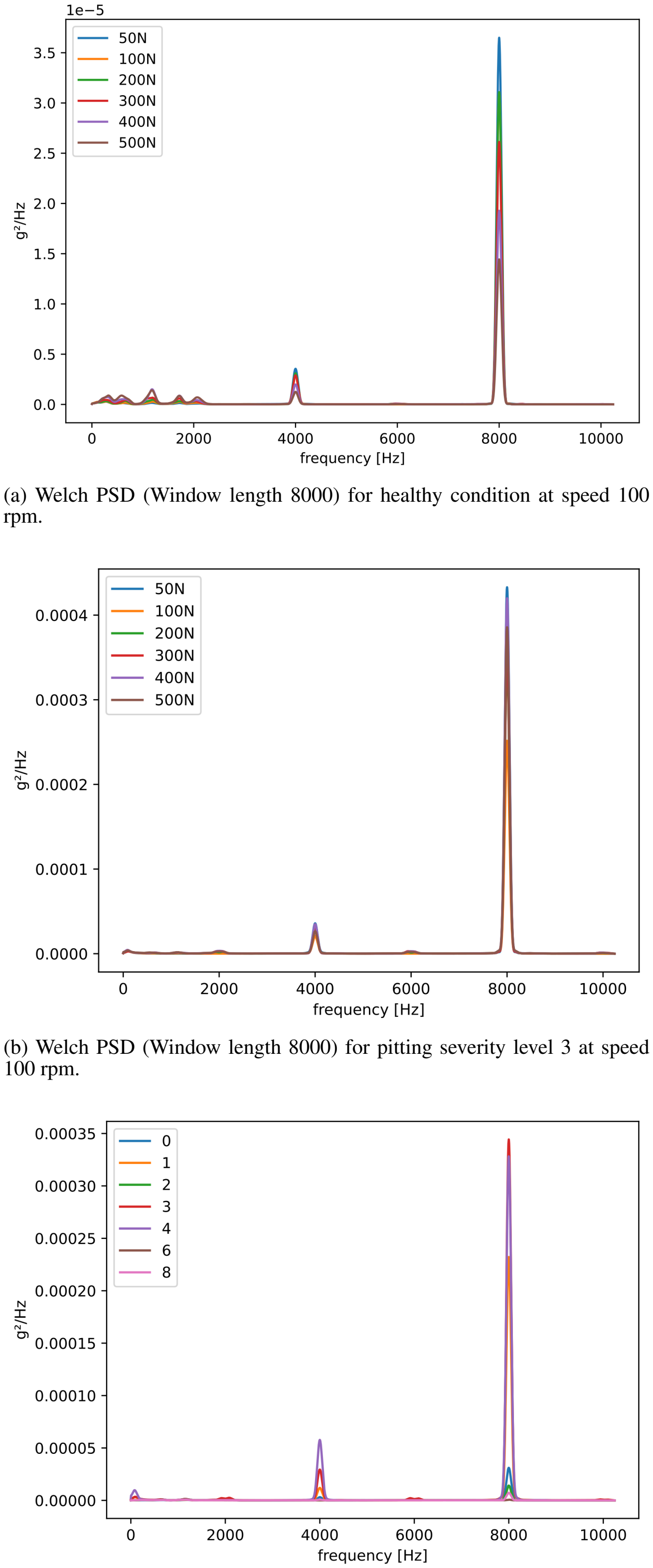
This results in a scoring scheme where random prediction of the observation pitting severity, on average, results in a negative score, irrespective of the true pitting severity of the observation. In fact, for a balanced testing data set the expected value of the score is -0.81. This scoring system encourages participants to not make incorrect predictions confidently. Specifically, the confidence factor c can allow participants to distinguish between observations for which the model prediction is considered reliable (confident prediction) or unreliable (unconfident prediction).

Finally, for each observation, a constraint on p_i is applied such that the sum of the predicted probabilities for a signal should be equal to or less than one:

$$\sum_{i=0}^{10} \text{p}_{\,i} \leq 1 \tag{2}$$

2. MODEL DESIGN APPROACH

This work presents a strategy for effective machine learning model design for predicting pitting severities with the added challenge of unseen pitting severities and unseen operating



(a) Welch PSD (Window length 8000) for healthy condition at speed 100 rpm.

(b) Welch PSD (Window length 8000) for pitting severity level 3 at speed 100 rpm.

(c) Welch PSD (Window length 8000) for severities in the training set at speed 100 rpm and load 200N.

Figure 2. Welch PSD (Window length 8000) for the measured vibration signals at different operating conditions and fault severities.

Table 2. Pitting severity score table: point system based on proximity to the true answer. Predictions far from the true pitting severity are heavily penalized.

Distance from true state	**pitting severity score s**
0 (correct prediction)	1.0
1	0.5
2	0
3	-0.5
4	-1.0
5	-1.5
6	-2.0
7	-2.5
8	-3.0
9	-3.5
10	-4.0

conditions in the test set.

As shown schematically in Figure 4, this approach follows an experimental design approach based on carefully selected validation sets. Each of the components are discussed in greater detail though the remainder of the paper. Firstly, Section 3 describes the procedure for creating dedicated validation sets for evaluating candidate machine learning models on data from unseen operating conditions and fault severities. Thereafter, Section 4 introduces several preprocessing approaches that are applied during model development. Following this, the machine learning classification model with an integrated out-of-distribution detector is considered in Section 5. The wide range of design choices outlined in Figure 4 are then evaluated on the previously designed validation sets. These results are presented in Section 6 together with the performance of the final selected model on the test dataset.

3. DESIGN OF VALIDATION SETS TO MIMIC TEST SETS WITH UNSEEN DATA

The modeling strategies proposed in this work are compared based on model performance on carefully designed validation sets which are created in an effort to emulate the test data that contains signals from unseen operating conditions and fault severities. A comprehensive set of preprocessing techniques, normalization schemes and machine learning models are evaluated and the optimal strategy is then selected based on its performance on each of these validation sets. This section describes how the validation sets are designed such that they mimic the true test conditions.

As mentioned in Section 1.2, the model will be evaluated on a test set containing data from the same operating conditions and pitting severities as training, with different operating conditions and with different fault severities. To evaluate the generalization of a model to these unseen conditions, can-

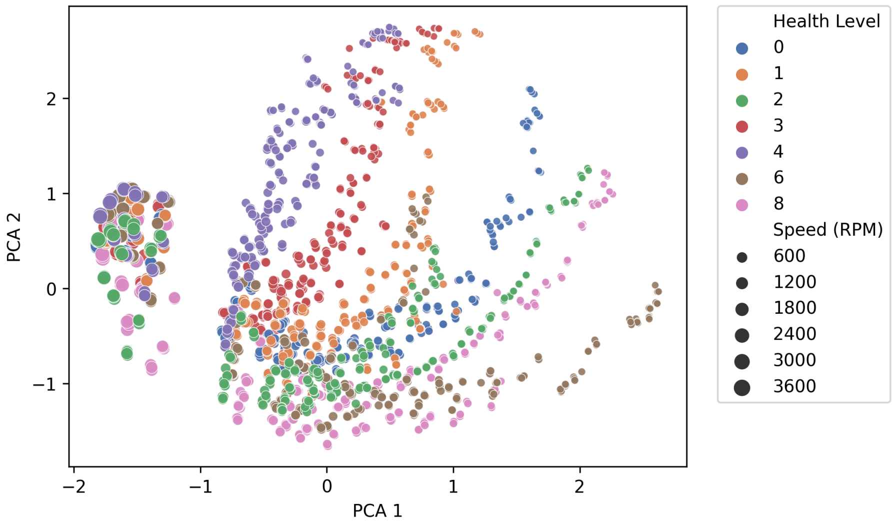


Figure 3. The first two principal components of the training data PSD's after Box Cox transform. PSD window length: 8000.

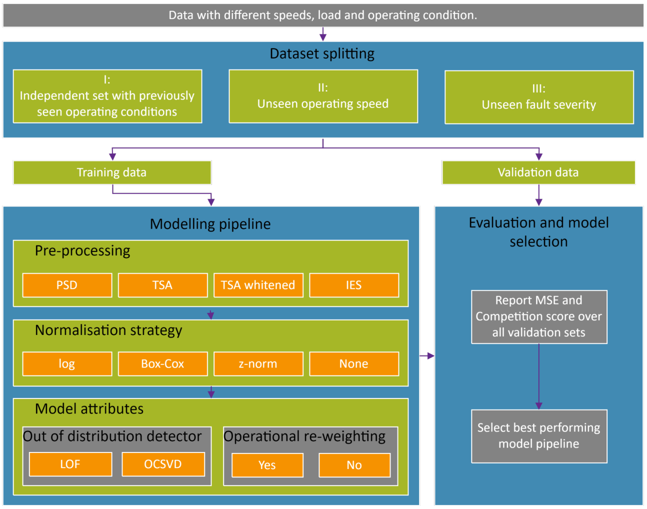

Data with different speeds, load and operating condition.
Dataset splitting
I:
Independent set with previously seen operating conditions
II:
Unseen operating speed
III:
Unseen fault severity
Training data
Validation data
Modelling pipeline
Pre-processing
PSD
TSA
TSA whitened
IES
Normalisation strategy
log
Box-Cox
z-norm
None
Model attributes
Out of distribution detector
LOF
OCSVD
Operational re-weighting
Yes
No
Evaluation and model selection
Report MSE and Competition score over all validation sets
Select best performing model pipeline

Figure 4. The proposed approach for model design.

didate models and preprocessing strategies are validated on a selected subset of the training set. Therefore, three scenarios are considered: seen operating conditions and pitting severities (normal supervised learning), unseen operating conditions, and unseen pitting severities during training. Out of these scenarios, three validation cases are created out of the training data: validation set I with 20% of the training signal randomly separated off, validation set II with the speeds of 900rpm and 1200rpm separated off and validation set III with fault severity two and eight separated off.

In validation set II the decision was made to use the speeds of 900rpm and 1200rpm as validation data based on the fact that the unseen speed-load conditions in the test set are not at low speed. In fact, the lowest unseen speed in the test data is 1500 rpm. Thus, validation set II focuses on high speed conditions which still contain torque levels up to 500Nm. Validation set III is created to highlight the challenge of unseen pitting conditions in the test set. As in the test set the highest pitting severity is not included in training, highest pitting severity from training (eight) is used for validation in validation set III. As not every unseen pitting condition in the test set are high severities, pitting condition two is included in validation set III as well.

The models are first evaluated on their performance on validation set I. The purpose of validation set I is to check for overfitting of the model on the training data and to evaluate the expected performance on the in-distribution test data. Secondly, the models are evaluated on the unseen speed conditions by the use of validation set II to check the generalization capabilities to unseen operating conditions. Finally, validation set III is used to determine which model could generalize best to unseen pitting severities.

4. DATA PREPROCESSING

In an attempt to combine the strengths of signal processing and machine learning methods, several signal processing and normalization techniques were evaluated to determine which diagnostic features are most suitable as input for the machine learning models to be evaluated.

4.1. Signal Processing Methods as Data Preprocessing

A total of four established signal processing approaches were evaluated to obtain informative features as input for the machine learning models. The first preprocessing technique under consideration is the Power Spectral Density (PSD), estimated by the Welch method (Welch, 1967). This method

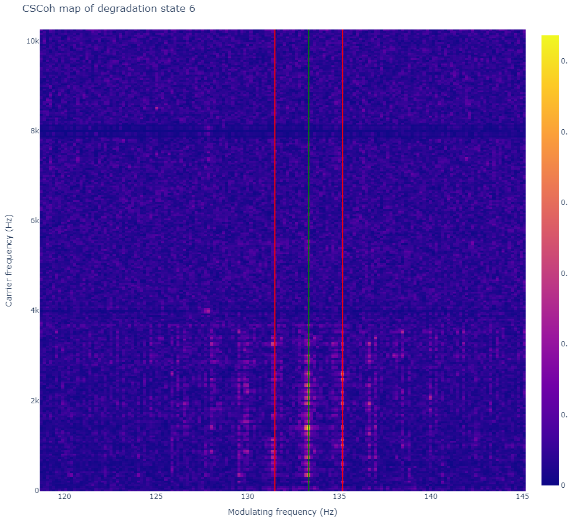


Figure 5. The CSCoh map of degradation state 6 with the gear mesh frequency indicated in green and the shaft frequency sideband in red.

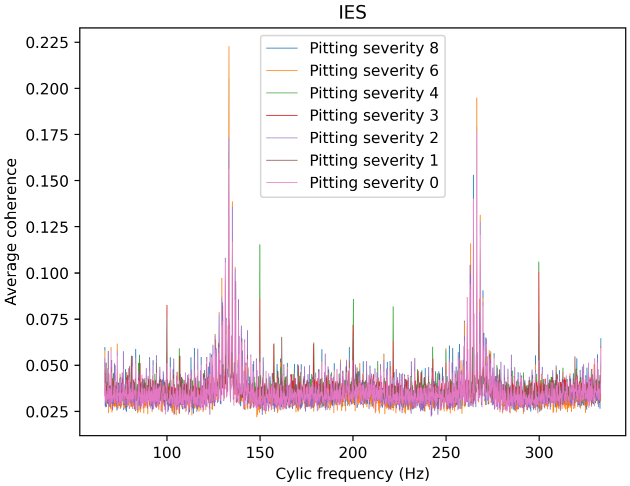


Figure 6. The IES with a spectral frequency range from 300Hz to 3700Hz for the 200rpm and 100Nm operating conditions from half the gearmesh frequency to two and a half times the gear mesh frequency with eight pitting severity levels.

gives an estimate of how the vibration signal energy is distributed in the frequency domain and has been effectively used as a preprocessing technique in machine learning gearbox fault recognition (Su, Tomovic, & Zhu, 2019). The PSD is calculated with a window length of 8000 points over a signal segment of 40960 points. Due to the fact that the signal segment of 40960 points is shorter than a full measurement in the training set, a data augmentation scheme can be implemented where the start point of the signal segment used to compute the PSD moves with a stride of 5120 samples over the full measured signal. This results in multiple different PSD representations per training signal and thus in an increase of training samples.

The second preprocessing technique applied in this work is the Cyclic Spectral Coherence (CSCoh), (Antoni, 2007). The CSCoh reduces the vibration signals to a bi-variable map which gives an indication of the signal energy associated with a given modulation between cyclic frequency (fault characteristic frequency) and spectral frequency (fault carrier frequency). One simple and effective procedure to analyze the bi-variable map is to integrate along spectral frequency to obtain a spectrum showing the energy associated with a given cyclic frequency over a range of spectral frequencies. For instance, the Improved Envelope Spectrum (IES) can be generated by integrating the magnitude of the CSCoh within a specific narrow spectral frequency band (Mauricio, Smith, Randall, Antoni, & Gryllias, 2020).

Figure 5 shows a section of the CSCoh map where the gear mesh frequency is indicated in green and the shaft frequency sidebands indicated in red. To monitor gear pitting under various speed and load conditions, the most salient fault information is expected at the gear mesh frequency and its sidebands. It is clear from Figure 5 that the modulation of the gear mesh frequency and its sidebands is strongest in the frequency band of 300Hz to 3700Hz and therefore this spectral frequency band is selected to generate the IES. In this way, the cyclostationarity-based feature derived from the bi-variate CSCoh map is used as the input to the pitting severity prediction model.

To maximize the fault information of the processed signal, only a range of frequencies of the IES from half of the gearmesh frequency to two and a half times the gear mesh frequency is used. The two largest peaks in this feature vector (See Figure 6) are related to the gearmesh frequency, with the expectation that the amplitude and number of sidebands spaced at regular multiples of the shaft frequency should be indicative of a gear pitting fault. This range of the IES is used as input for the ML for all the operational speeds, meaning that the gearmesh peaks and sidebands tend to show up at the same position in the feature vector provided that the amount of frequency bins in this range remain the same. This proposed feature representation is chosen in an effort to help the machine learning network to generalize well to unseen operational conditions.

The CSCoh indicator, and consequently the IES, has a fixed frequency resolution which is the sampling frequency divided by the signal length (Antoni, Xin, & Hamzaoui, 2017). Since measured signals in the dataset originate from different operating speeds, the frequency range of interest in the IES differ from each other, which brings inconsistency of the input dimensions for the network. To address this issue, the amount of frequency bins in a frequency range expected to contain the

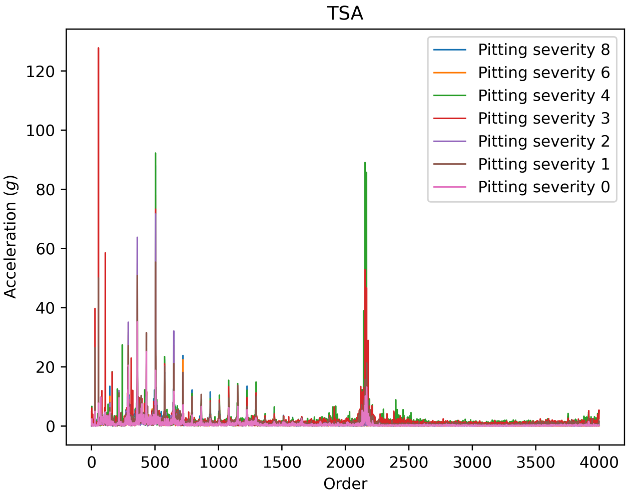


Figure 7. The frequency representation of the time synchronous averaged signal for the 200rpm and 100Nm operating conditions for eight degradation states

most salient fault information is kept constant. This is done by selecting only a part of the full signal length in computing the IES, thereby controlling the frequency resolution. In future work other approaches could be proposed in an effort to efficiently reduce the dimension of the IES representation in the frequency domain.

The final preprocessing technique considered in this work is the time synchronous average (TSA), which is used to extract signal components that are periodic with the observed gear such as the pitting faults. This operation is performed by averaging a series of angle-synchronized signal segments, each corresponding to one period of a synchronous signal. Specifically, the raw signal is firstly segmented based on the available tacho signal (Mcfadden & Toozhy, 2000), and is then resampled with a fixed dimension. Finally, each of the resampled segments are averaged. As an additional preprocessing candidate, we further evaluate a TSA preprocessing scheme that includes cepstrum prewhitening before the resampling and averaging (Borghesani, Pennacchi, Randall, Sawalhi, & Ricci, 2013). The resampling step results in a representation in the angular domain with a fixed size, which allows for element-wise averaging and also allows for a fixed input size for downstream machine learning models. The fixed signal length in this work is chosen to be 8000 points. Finally, since all other methodologies represent the signal in the frequency domain the TSA signal is transformed into the frequency domain using the fast Fourier transform. Figure 7 shows the frequency representation of the time synchronous averaged signal for one of the operating conditions.

4.2. Normalization and Scaling Candidates

Different data normalization schemes can greatly influence the input data distribution on which neural networks are trained, which can have a significant influence on network training and performance. Therefore, different normalization and scaling schemes are evaluated in this work.

The first normalization scheme under investigation is z-normalization. In this approach, data is centred by subtracting its mean and scaled with respect to the standard deviation as calculated over the full training dataset. This results in the transformation shown in Eq. (3), where μ is the mean and σ the standard deviation which is calculated over the full training dataset per frequency bin.

$$x = \frac{x - \mu}{\sigma} \tag{3}$$

Furthermore, the Box-Cox normalization (Box & Cox, 1964) approach is further considered as a normalization candidate. The Box-Cox normalization scheme is a method that transforms positive non-normally distributed data to be more Gaussian-like. This property makes it well suited for normalizing positive amplitude spectra as used in this investigation. The Box-Cox approach utilizes a power transformation to stabilize data variances and make the data distribution more symmetric. The transformation is defined as:

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(x), & \text{if } \lambda = 0 \end{cases}$$

Here, x is the input data, and λ is the transformation parameter. The optimal λ value is determined through maximum likelihood estimation.

As a final preprocessing candidate, the frequency representations are also re-scaled using a simple log transformation. Here, the motivation is similar as for the Box-Cox transform, ensuring that the strictly positive spectral data has more uniformly distributed variance.

5. CLASSIFICATION AND OUT-OF DISTRIBUTION MODEL DESIGN

In this section, the machine learning models that analyze the preprocessed data discussed in the previous section are introduced. This includes a convolutional neural network for fault severity classification, as well as an out-of-distribution detection model to identify samples for which the model should not make confident predictions. All the machine learning hyperparameters are set based on experimentation on the previously defined validation sets.

5.1. Machine Learning Models

The machine learning models considered in this work are CNN's with 3 layers of one dimensional convolutional layers, followed by a flatten layer, one fully connected layer, and an output layer with a single output which was trained in an ordinal regression fashion (Rosenthal & Ratna, 2022) (LeCun et al., 1989) (McCullagh, 1980). By using the ordinal regression instead of a softmax layer with cross entropy loss the relative ordering of the different classes is preserved.

The first convolutional layer has 5 filters with a kernel size of 100 and a stride of 10. The second has 10 filters with a kernel size of 50 and a stride of 10. The final convolutional layer has 20 filters with a kernel size of 20 and a stride of 2. The fully connected layer has 20 output nodes.

Initial trails on the validation tests showed that the model has a tendency to over-fit on the training set and thus several measures are taken to prevent overfitting. First of all dropout was used with a factor of 0.1 on the first convolutional layer and 0.03 on the second convolutional layer (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012). Then a weight decay of 10^{-6} is added (Loshchilov & Hutter, 2017). Finally, to reduce the number of weights, the same weights were used for the convolutional layers over the data of the three different axes of the accelerometer.

The CNN's were trained using an Adam optimizer (Kingma & Ba, 2014) with a learning rate of 10^{-4} for 60 epochs when the PSD with data augmentation is used and 1020 epochs when one of the preprocessing schemes without data augmentation is used. In this way, the same number of training steps are used between different preprocessing techniques.

Further, to address potential problems arising from the unequal distribution of the operating speeds and of the operating loads a loss reweighting scheme is investigated (Cui, Jia, Lin, Song, & Belongie, 2019). The weight assigned to an operating speed is equal to the average amount of samples per operating speed divided by the amount of samples for the operating speed in question. In the same way, a weighting scheme is implemented for the operational loads. The final loss weight used for a given speed and load combination is then the weight from the operating speed multiplied by the weight from the operating load.

The model used in the final submission is based on an ensemble of the same model with five different random initializations. Using a homogeneous ensemble of models by combining the model results tend to improve model generalizability (Ganaie, Hu, Malik, Tanveer, & Suganthan, 2022).

During the validation phase, it was observed that the final outputs of the deep ordinal model are miscalibrated, which means that the predicted probability distribution did not match well with the expected empirical distribution of the labels. The model was very unsure about its prediction and thus always outputted a very high entropy distribution, where the predicted probabilities for all classes were similar. However, in most of the cases on Validation Set I the true label was predicted to have the highest probability and thus the model should have been more certain.

To solve this miscalibration a post-processing step is introduced which transforms the final output into a discrete distribution where one class has $p_i = 1.0$ and all other classes have $p_i = 0$. During the experimental phase, some experiments were carried out using temperature scaling (Guo, Pleiss, Sun, & Weinberger, 2017). However, it was observed that extremely low temperature showed the best results and thus the temperature scaling was simplified to a one-hot distribution. During validation, it was observed that this post-processing step increased the average score. Additionally, to handle uncertainty in the output of the model an out-of-distribution (OOD) detector is considered as will be discussed in the next section.

5.2. Out-of-Distribution Detection Methodologies

As will be shown in the next section, the models are able to perform well on in-distribution data (Validation test I) and unseen speeds (Validation test II) but not on unseen health conditions (Validation test III). Therefore, an out-of-distribution (OOD) detector is implemented to detect unseen pitting severities and to indicate that these predictions are uncertain. The OOD detector is trained to consider data from the training set as normal and other unfamiliar data as anomalous. Specifically, the latent features from the final layer of the machine learning model is used as input to the anomaly detection model.

Two different anomaly detection models are considered which are the One-Class SVM (Alam, Sonbhadra, Agarwal, & Nagabhushan, 2020) and the Local Outlier Factor (LOF) (Breunig, Kriegel, Ng, & Sander, 2000).

The One-Class SVM is a variant of the popular Support Vector Machine (SVM) classification algorithm. It is trained on normal instances to create a boundary that encompasses the majority of the normal data points. To do this, the algorithm identifies a hyperplane that separates the normal data from the outliers, with the margin between the healthy and faulty samples defined by a parameter ν that governs the proportion of training data considered as outliers, thereby allowing for a flexible adjustment of the outlier detection sensitivity. The formulation of the One-Class SVM involves solving a convex optimization problem, which results in finding a hyperplane that maximizes the margin while minimizing the classification errors for the normal instances. Points lying on or outside the decision boundary are considered outliers. The One-Class SVM is implemented with a radial basis function kernel with a gamma value of 0.9 and a ν value of 0.01.

Table 3. Average score and MSE of the machine learning model with different preprocessing techniques and different normalization schemes on different validation sets. Best performing scores are reported in **bold**. Numbers in brackets indicate the standard deviation over 5 independent trials. Set I: In-distribution data, Set II: Unseen speed, Set II: Unseen fault severity.

Preprocessing	Normalization	Set I		Set II		Set III	
		Score $\uparrow$	MSE $\downarrow$	Score $\uparrow$	MSE $\downarrow$	Score $\uparrow$	MSE $\downarrow$
PSD	Log	**0.969 (0.009)**	**0.062 (0.018)**	0.278 (0.042)	1.444 (0.085)	**0.002 (0.020)**	**1.996 (0.040)**
	Box-Cox	0.956 (0.006)	0.089 (0.012)	0.493 (0.117)	1.015 (0.223)	-0.138 (0.023)	2.276 (0.046)
	z-norm	0.949 (0.008)	0.101 (0.016)	-0.312 (0.153)	2.624 (0.306)	-0.055 (0.070)	2.109 (0.139)
	None	0.875 (0.043)	0.251 (0.087)	-0.364 (0.123)	2.729 (0.246)	-0.304 (0.047)	2.609 (0.093)
TSA	log	0.835 (0.015)	0.330 (0.031)	0.151 (0.087)	1.699 (0.175)	-0.376 (0.053)	2.752 (0.106)
	boxcox	0.836 (0.013)	0.328 (0.026)	-0.010 (0.058)	2.019 (0.115)	-0.456 (0.099)	2.912 (0.197)
	z-norm	0.841 (0.015)	0.317 (0.030)	0.043 (0.117)	1.914 (0.234)	-0.348 (0.049)	2.695 (0.099)
	None	0.850 (0.026)	0.301 (0.051)	0.031 (0.058)	1.938 (0.117)	-0.363 (0.064)	2.726 (0.127)
TSA whitened	log	0.755 (0.015)	0.489 (0.030)	0.508 (0.091)	0.984 (0.182)	-0.418 (0.043)	2.836 (0.086)
	boxcox	0.750 (0.054)	0.500 (0.108)	0.398 (0.066)	1.203 (0.133)	-0.414 (0.049)	2.828 (0.097)
	z-norm	0.844 (0.027)	0.311 (0.054)	**0.653 (0.070)**	**0.694 (0.139)**	-0.142 (0.046)	2.285 (0.092)
	None	0.861 (0.006)	0.279 (0.013)	0.624 (0.092)	0.752 (0.185)	-0.153 (0.036)	2.306 (0.072)
IES	Log	0.509 (0.015)	0.982 (0.029)	-0.002 (0.049)	2.004 (0.099)	-0.518 (0.049)	3.037 (0.098)
	Box-Cox	0.684 (0.016)	0.633 (0.032)	0.302 (0.047)	1.396 (0.095)	-0.395 (0.057)	2.790 (0.114)
	z-norm	0.724 (0.022)	0.552 (0.043)	0.304 (0.056)	1.393 (0.113)	-0.287 (0.067)	2.573 (0.134)
	None	0.672 (0.019)	0.656 (0.037)	0.153 (0.035)	1.694 (0.069)	-0.468 (0.077)	2.936 (0.155)

The second outlier model used is the Local Outlier Factor (LOF) model. Unlike One-Class SVM, LOF does not require a training phase and evaluates the local density of instances to identify outliers. The LOF algorithm assesses the density of a data point by comparing its local density to the densities of its neighbors. Data points with significantly lower local densities compared to their neighbors are identified as outliers. The local density is determined based on the distances between data points and their k-nearest neighbors where for this work 100 neighbours are considered. In the training data 1% of the data is considered anomalous which is used to set the threshold on the local density. This threshold is then used to determine if a sample is out-of-distribution or not.

When a sample is flagged as OOD the model will set the value of c as used in Eq. (1) to 0.2 with the intention of minimizing the penalty from the misclassified OOD samples.

6. RESULTS

This section presents the results of different models for the validation sets described in Section 3 and the performance of the selected model on the hold out test set.

6.1. Results on Specially Designed Validation Sets

The average score and the mean square error (MSE) between the predicted pitting severity and the actual severity on the designed validation sets are shown in Table 3 with the standard deviation in brackets. The results show that PSD-based approaches tend to have the highest score on the in-distribution validation set (Validation set I). One possible explanation for this could be due to the data augmentation scheme used as part of the PSD based approach. This data augmentation scheme results in more training samples which can reduce overfitting and thus increase generalizability on in-distribution data.

However, for the unseen operating conditions (Validation set II) some of the TSA methods are outperforming PSD-based methods.

One of the possible causes could be that not every speed and not every load is present in an equal amount in the training dataset of the PSD methodology due to the data augmentation. This is due to the fact that, for the high operating speeds the signal length is reduced (only 3s captured instead of 12s) and thus when using the proposed data augmentation approach for PSD data there is a disproportionate amount of training samples for the low operating speeds. Therefore, the reweighting scheme proposed in Section 5 is used to remedy this. The results for models that do and do not use the re-weighting scheme are shown in Table 4. Making use of the reweighting scheme results in an increase on the average score of 0.015 on set II. It can be noted that despite this performance increase for the PSD based model, the best TSA based model still achieves a higher score.

It is clear that all the models perform best on the in-distribution data (Validation test I) which was expected based on the previous literature. Most models also perform better on the unseen speeds (Validation set II) than the unseen states (Validation set III).

To account for the reduced performance in validation set III two different OOD detection models are considered in an effort to decrease penalizations of wrong predictions as described in Section 5.2. When an anomaly detection model flags a sample as anomalous the prediction is considered an uncertain prediction, which scales the score with a factor of

Table 4. Performance of PSD with log scaling model with and without operational weighting in training. Set I: In-distribution data, Set II: Unseen speed, Set III: Unseen fault severity.

Weighting	Set I		Set II		Set III	
	Score	MSE	Score	MSE	Score	MSE
No	0.969 (0.009)	0.062 (0.018)	0.278 (0.042)	1.444 (0.085)	0.002 (0.020)	1.996 (0.040)
Yes	0.968 (0.005)	0.064 (0.009)	0.293 (0.081)	1.415(0.163)	0.030 (0.049)	1.939 (0.099)

Table 5. Performance of PSD with log scaling model with and without several out-of-distribution detectors. Set I: In-distribution data, Set II: Unseen speed, Set III: Unseen fault severity.

OOD detector	Set I	Set II	Set III
	Score	Score	Score
No OOD detection	0.968 (0.005)	0.293 (0.081)	0.030 (0.049)
One-Class SVM	0.952 (0.015)	0.099 (0.034)	0.050 (0.017)
LOF	0.945 (0.012)	0.164 (0.095)	0.007 (0.054)

0.2 as discussed in Section 1.2.2.

The performance on the three validation sets with the OOD detection model on a model with PSD based preprocessing and log scaling is reported in Table 5. The table shows that using OOD model results in a drop in performance on Validation set I. However, using an OCSVM OOD detector a smaller drop in score can be observed in comparison to the LOF model. The One-Class SVM model also allows for an increase in the performance on validation set III and thus correctly flags samples with negative scores as OOD.

Ultimately, a model with PSD-based preprocessing with logarithmic scaling was selected, as this approach leads to the highest performance on validation sets I and III. The model further included a loss reweighting scheme during training with respect to the operating conditions as described in Section 5.

Additionally, it could be observed that the model had difficulties with generalizing to unseen pitting severities and thus an out-of-distribution detector was implemented. Table 5 shows that this did not impact the results on the in-distribution data (validation set I) significantly. However, the out-of-distribution detector did decrease performance on the unseen operational conditions and thus it was not used when a test sample came from an unseen speed-load combination.

6.2. Results on Data Competition Unseen Test Sets

The final model submitted to the competition was evaluated on two unseen test sets of which the labels were not publicly available but were scored by the organizers. Ultimately, an ensemble of 5 randomly initialized models trained on PSD data with a log re-scaling was submitted. The model achieved a final total score of 282.2 on the test set used as a criterion by the organizers and 213.3 on test set on which participants could evaluate their models beforehand. The maximum score achievable on this dataset is 800 (800 signals with each a maximum score of one). Despite the heavy penalization of incorrect predictions as explained in Section 1.2.2, the model achieves a respectable positive score that is consistent with the results on the internal test sets listed in Table 3. Considering that the test set contained 800 samples translating to an average score of 0.35, which is in between the results on the different validation sets. Thus, the performance of the validation sets seems in line with the performance on the test sets. To conduct a quantified analysis on the representativeness of the validation sets, the relative ratio of unseen pitting severities would be needed which is currently unknown.

7. CONCLUSION

In this work, a strategy is presented to design a model for predicting gearbox pitting severity under unseen operating conditions and fault severities.

As part of this strategy a convolutional neural network with an ordinal loss criterion, trained on the power spectral density data is proposed as a possible solution. This model was selected after rigorous model evaluation using three validation sets that are carefully designed to evaluate how well the model generalizes to unseen operation conditions and fault severities. The results show that the proposed model is successful at generalizing to unseen operating conditions whilst generalizing to unseen fault severities remains a challenge.

The model design approach exploits the use of validation sets that provide insights for making modeling choices and tend to be indicative of the final score on the test set. Thus, this work illustrates the need for careful experimental design in preprocessing and machine learning methods for condition monitoring.

In future work, additional signal processing methods could be evaluated as preprocessing options since the result showed notable differences in model performance depending on the signal processing method used for preprocessing. Additionally, data augmentation techniques could be considered to further enhance model generalization. Finally, ensembles based on different preprocessing techniques could be considered to capture the benefits of different signal processing techniques on different cases as the TSA methodology did outperform the PSD methodology on unseen speed-load conditions. The idea of using an ensemble could also be used to augment the OOD detector, where a different OOD detector with different parameters could be designed to be used on unseen speeds.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support of the European Commission under the Marie Sklodowska-Curie program through the ETN MOIRA project (GA 955681). This work was supported by Flanders Make, the strategic research center for the manufacturing industry, in the context of DGTwin Prediction and QED projects and the Flemish government under the "AI Flanders" programme. The authors also gratefully acknowledge the financial support provided by the FWO - Fonds Wetenschappelijk Onderzoek – Vlaanderen under the project G0A3123N.

REFERENCES

Alam, S., Sonbhadra, S. K., Agarwal, S., & Nagabhushan, P. (2020). One-class support vector classifiers: A survey. *Knowledge-Based Systems*, *196*, 105754. doi: 10.1016/j.knosys.2020.105754

Antoni, J. (2007). Cyclic spectral analysis of rolling-element bearing signals: Facts and fictions. *Journal of Sound and vibration*, *304*(3-5), 497–529.

Antoni, J., Xin, G., & Hamzaoui, N. (2017). Fast computation of the spectral correlation. *Mechanical Systems and Signal Processing*, *92*, 248–277.

Bishop, C. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image and Signal Processing*, *141*, 217-222(5).

Borghesani, P., Pennacchi, P., Randall, R., Sawalhi, N., & Ricci, R. (2013). Application of cepstrum prewhitening for the diagnosis of bearing faults under variable speed conditions. *Mechanical Systems and Signal Processing*, *36*(2), 370–384. doi: 10.1016/j.ymssp.2012.11.001

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. doi: 10.1111/j.2517-6161.1964.tb00553.x

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93–104). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/342009.335388

Cartocci, N., Napolitano, M. R., Costante, G., Valigi, P., & Fravolini, M. L. (2022). Aircraft robust data-driven multiple sensor fault diagnosis based on optimality criteria. *Mechanical Systems and Signal Processing*, *170*, 108668. doi: 10.1016/j.ymssp.2021.108668

Chen, Z., Li, C., & Sanchez, R.-V. (2015). Gearbox Fault Identification and Classification with Convolutional Neural Networks. *Shock and Vibration*, *2015*, 1–10. doi: 10.1155/2015/390134

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9268–9277).

Elasha, F., Ruiz-Cárcel, C., Mba, D., Kiat, G., Nze, I., & Yebra, G. (2014). Pitting detection in worm gearboxes with vibration analysis. *Engineering Failure Analysis*, *42*, 366–376. doi: 10.1016/j.engfailanal.2014.04.028

Ganaie, M., Hu, M., Malik, A., Tanveer, M., & Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151. doi: 10.1016/j.engappai.2022.105151

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

Hendriks, J., Dumond, P., & Knox, D. (2022). Towards better benchmarking using the CWRU bearing fault dataset. *Mechanical Systems and Signal Processing*, *169*, 108732. doi: 10.1016/j.ymssp.2021.108732

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, A., Parey, A., & Kankar, P. K. (2023). Supervised Machine Learning Based Approach for Early Fault Detection in Polymer Gears Using Vibration Signals. *MAPAN*, *38*(2), 383–394. doi: 10.1007/s12647-022-00608-8

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Lee, J., Wu, F., Zhao, W., Ghaffari, M., & Liao, L. (2014). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. *Mechanical Systems and Signal Processing*, 21. doi: http://dx.doi.org/10.1016/j.ymssp.2013.06.004

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: 10.1016/j.ymssp.2017.11.016

Li, X., Li, J., Qu, Y., & He, D. (2019). Gear pitting fault diagnosis using integrated cnn and gru network with both vibration and acoustic emission signals. *Applied Sciences*, *9*(4), 768.

Li, X., Li, J., Zhao, C., Qu, Y., & He, D. (2020). Gear pitting fault diagnosis with mixed operating conditions based on adaptive 1D separable convolution with residual connection. *Mechanical Systems and Signal Processing*, *142*, 106740. doi: 10.1016/j.ymssp.2020.106740

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Louizos, C., & Welling, M. (2017). Multiplicative normalizing flows for variational Bayesian neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 2218–2227). PMLR.

Mauricio, A., Smith, W. A., Randall, R. B., Antoni, J., & Gryllias, K. (2020). Improved envelope spectrum via feature optimisation-gram (iesfogram): A novel tool for rolling element bearing diagnostics under non-stationary operating conditions. *Mechanical Systems and Signal Processing*, *144*, 106891. doi: https://doi.org/10.1016/j.ymssp.2020.106891

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127.

Mcfadden, P., & Toozhy, M. (2000). Application of synchronous averaging to vibration monitoring of rolling element bearings. *Mechanical Systems and Signal Processing*, *14*(6), 891–906. doi: 10.1006/mssp.2000.1290

Medina, R., Cerrada, M., Cabrera, D., Sanchez, R.-V., Li, C., & Oliveira, J. V. D. (2019). Deep Learning-Based Gear Pitting Severity Assessment Using Acoustic Emission, Vibration and Currents Signals. In *2019 Prognostics and System Health Management Conference (PHM-Paris)* (pp. 210–216). Paris, France: IEEE. doi: 10.1109/PHM-Paris.2019.00042

Öztürk, H., Sabuncu, M., & Yesilyurt, I. (2008). Early Detection of Pitting Damage in Gears using Mean Frequency of Scalogram. *Journal of Vibration and Control*, *14*(4), 469–484. doi: 10.1177/1077546307080026

Prognostics and Health Management Society. (2023). *PHM North America 2023 Conference Data Challenge.* Online data repository. (Accessed on 2023-10-17. https://data.phmsociety.org/phm2023-conference-data-challenge)

Qin, Y., Wang, Z., & Xi, D. (2022). Tree CycleGAN with maximum diversity loss for image augmentation and its application into gear pitting detection. *Applied Soft Computing*, *114*, 108130. doi: 10.1016/j.asoc.2021.108130

Qu, Y., He, M., Deutsch, J., & He, D. (2017). Detection of Pitting in Gears Using a Deep Sparse Autoencoder. *Applied Sciences*, *7*(5), 515. doi: 10.3390/app7050515

Rosenthal, E., & Ratna, S. (2022). *Spacecutter.* https://github.com/EthanRosenthal/spacecutter, https://www.ethanrosenthal.com/2018/12/06/spacecutter-ordinal regression/.

Salameh, J. P., Cauet, S., Etien, E., Sakout, A., & Rambault, L. (2018). Gearbox condition monitoring in wind turbines: A review. *Mechanical Systems and Signal Processing*, *111*, 251–264. doi: 10.1016/j.ymssp.2018.03.052

Schmidt, S., & Heyns, P. S. (2019). An open set recognition methodology utilising discrepancy analysis for gear diagnostics under varying operating conditions. *Mechanical Systems and Signal Processing*, *119*, 1–22. doi: 10.1016/j.ymssp.2018.09.016

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90*(2), 227-244. doi: https://doi.org/10.1016/S0378-3758(00)00115-4

Su, X., Tomovic, M. M., & Zhu, D. (2019). Diagnosis of gradual faults in high-speed gear pairs using machine learning. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *41*, 1–11.

Teng, W., Wang, F., Zhang, K., Liu, Y., & Ding, X. (2014). Pitting Fault Detection of a Wind Turbine Gearbox Using Empirical Mode Decomposition. *Strojniški vestnik – Journal of Mechanical Engineering*, *60*(1), 12–20. doi: 10.5545/sv-jme.2013.1295

Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, *15*(2), 70–73.