21st Australian International Aerospace Congress

ISBN number 978-1-925627-90-9

Please select category below:
Normal Paper ☐
Student Paper ☒
Young Engineer Paper ☐

Explainable Anomaly Detection with Neural Networks for Gear Condition Monitoring

N. Herwig[1], P. Borghesani[1], W. Smith[1], Z. Peng[1] and J. Antoni[1,2]

[1] *School of Mechanical & Manufacturing Engineering, UNSW Sydney, Australia*
[2] *Univ Lyon, INSA Lyon, LVA, UR677, 69621 Villeurbanne, France*

Abstract

Anomaly detection for gears can be used to overcome the scarcity of faulty condition data in many geared systems by using only healthy data to train a model. Utilising reasonable preprocessing in combination with the Deep Support Vector Data Description, the one class classification method offers high accuracy in anomaly detection. This is shown in this paper based on analysing a run-to-failure gear dataset. Additionally, this paper utilises Shapley Additive Explanations to highlight relevant input features used in anomaly detection to enhance the trust and explainability of the network. The entire framework offers a stable, accurate and interpretable approach to detecting anomalies/faults in gears without the need of historical fault data and expert intervention.

Keywords: Anomaly detection; Neural networks; Gear diagnostics; Fault detection; Signal processing

Introduction

Gears play a pivotal role in the reliable operation of rotational machines, and early fault detection is essential for avoiding costly failures and ensuring operational safety [1]. Traditional methods for fault detection include vibration signal analysis techniques, such as order tracking and synchronous averaging, which require expert knowledge for accurate fault identification [2]. Recent advances in machine learning (ML) offer automated solutions by leveraging large datasets, but most methods need labeled data for both healthy and faulty conditions, which are often unavailable [3]. This paper uses an anomaly detection (AD) approach with Deep Support Vector Data Description (Deep SVDD), allowing training on only healthy data, which overcomes the scarcity of fault data [4]. Additionally, Shapley Additive Explanations (SHAP) enhance the transparency and trust in the model's decision-making process.

This one-class classification was first introduced in [5], which serves as the fundament for the early AD networks. Over the years, AD methods have advanced from the less stable and inefficient approaches like One-Class SVM (OC-SVM) [6] to deep AD approaches [7,8]. In [9], the authors proposed a network which maps the processing data to a hypersphere. During training, the volume of this hypersphere gets minimised. Data which shows features distinct from the training data is mapped outside of the hypersphere and thus classified as an anomaly.

In condition monitoring (CM), many publications utilise AD methods to detect faults during operation. Regression long short-term memory (LSTM) architectures are often combined with a one-class support vector machine (SVM) to perform AD in CM [10]. In [11], the authors used the Deep SVDD [9] to detect anomalies in a helicopter vibration dataset. They also used Cyclic Spectral Correlation and Cyclic Spectral Coherence as inputs and a convolution neural network

in combination with dense layers to create a hypersphere of the healthy data. A similar approach is published in [12] for bearings. The authors also include a data enhancement method for improved accuracy. None of the publications offers a detailed explanation of their AD algorithm, raising concerns about the explainability and trustworthiness of the detected anomalies.

In this paper, the Deep SVDD [9] is utilised to map healthy preprocessed gear data. In the following, a run-to-failure test dataset is used to evaluate the performance of the technique. Finally, the network is analysed using the XAI method of SHapley Additive exPlanations (SHAP) [13] to ensure physically meaningful decision making and generalisability of the approach. The main contributions of the paper can be summarised as follows:

1. AD for gear vibration data using the Deep SVDD [9] one-class classification method and analytical preprocessing of the raw data.
2. Enhancing trust in the anomalies found using SHAP [13].

Methodology

The methodology used in this paper consists of three main steps: (1) signal preprocessing, (2) AD using Deep SVDD, and (3) SHAP-based explanation.

Pre-processing

Raw vibration signals from the gear are challenging to analyse due to their non-stationary nature. Preprocessing starts with order tracking, which resamples the signal based on the angular position of the reference shaft, followed by synchronous averaging to reduce noise.

In the resulting order-tracked signal, each sample corresponds to a specific angular position within a revolution of the shaft [14], and therefore the original f_s samples-per-second vibration signal $x(n/f_s)$ is transformed to its angular domain counterpart $x_{OT}\left(k/N_{spr}\right)$ with N_{spr} representing the number of samples per revolution chosen for the angular resampling, $k = 0, \dots, N_{spr}N_{rev} - 1$ and N_{rev} indicating the number of full revolutions of the shaft recorded in the signal.

To further enhance the data and reduce noise, synchronous averaging (SA) is used. The order tracked signal is segmented in single-revolution blocks (i.e., each block has a length N_{spr} samples), and then a synchronous average is computed across all samples with the same angular rotation, i.e.:

$$x_{SA}\left(\frac{k}{N_{spr}}\right) = \frac{1}{N_{rev}} \sum_{r=0}^{N_{rev}-1} x_{OT}\left(\frac{k}{N_{spr}} + r\right) \text{ with } k = 0, \dots, N_{spr} - 1 \qquad (1)$$

Since the duration of the SA signal is exactly one revolution of the shaft, its spectrum $X_{SA}(h)$ obtained via DFT contains only shaft harmonics, i.e., $|X_{SA}(h)|$ represents the amplitude of the h-th harmonic of the shaft frequency. The value $|X_{SA}(0)|$ is the mean of the acceleration signal, which is not used in the analysis of this paper.

The resulting amplitude spectrum $|X_{SA}(h)|$ with $h = 1, \dots, N_{spr}/2 + 1$ is used as an input for the neural network. For the sake of conciseness, in the following section the symbol $X_i(h)$ will be used for the h-th harmonic of the shaft frequency of the synchronous average $|X_{SA}(h)|$ of the i-th signal in a dataset, while $\mathbf{X}_i = \{X_i(1), \dots, X_i(N_{spr}/2 + 1)\}$ will represent the entire order spectrum for the same signal.

Deep SVDD

Healthy-only data is used to train a deep SVD for AD. The input to the network is the pre-processed data, i.e., the order-spectra obtained after order-tracking and synchronous averaging. Such healthy-only training dataset $\mathcal{X}_{train} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$ is composed of N training order-spectra, each with the same length $N_{spr}/2 + 1$.

The data is further processed by a neural network $\phi(\cdot;\mathcal{W})$ with three dense layers with trainable weights $\mathcal{W} = \{\mathrm{W}^{(1)}, \dots, \mathrm{W}^{(3)}\}$, so that, for each input $\mathbf{X}_i$ the network outputs:

$$\mathbf{Y}_i = \phi(\mathbf{X}_i;\mathcal{W}) \tag{2}$$

with $\mathbf{Y}_i = \{Y_i[1], Y_i[2]\}$, chosen to be bi-dimensional purely for presentation purposes.
The three subsequent dense layers have a decreasing number of neurons and specifically 50, 10 and 2. These numbers are manually set and may vary for different input sizes, as they cannot be precisely configured within a neural network architecture. The layers do not have a bias term and a ReLU function is chosen for the first two layers, since it avoids a non-zero saturation. These two conditions are necessary to avoid overfitting of the healthy-only dataset, as explained in [9]. The last layer has no activation function, so that it can map the input to the entire $\mathbb{R}^2$ space. Therefore, the concise expression of eq. (2) is more explicitly represented as:

$$\mathbf{Y}_i = \mathrm{W}^{(3)}\,\mathrm{ReLU}\left(\mathrm{W}^{(2)}\mathrm{ReLU}\left(\mathrm{W}^{(1)}\mathbf{X}_i\right)\right) \tag{3}$$

The objective function to be minimised by an ADAM optimiser is:

$$L = \min_{W}\left(\frac{1}{n}\sum_{i=1}^{n}\|\phi(x_i;\mathcal{W}) - \mathbf{c}\|^2 + \frac{\lambda}{2}\sum_{\ell=1}^{3}\left\|\mathrm{W}^{(\ell)}\right\|_F^2\right) \tag{4}$$

The first term is a Euclidean distance (squared) of the NN outputs $\mathbf{Y}_i$ from an arbitrary point $\mathbf{c} = \{c[1], c[2]\}$, which, according to [9] is chosen as the average output of the untrained network on the healthy dataset, i.e., the network with initialisation weights $\mathcal{W}_i$:

$$\mathbf{c} = \frac{1}{n}\sum_{i=1}^{N}\phi(\mathbf{X}_i, \mathcal{W}_{init}) \tag{5}$$

All other choices for $\boldsymbol{c}$ would obtain similar results, however, [9] suggests that this choice of $\boldsymbol{c}$ converges faster and is more robust. This term aims at encouraging the network during training to project all the healthy $\mathbf{X}_i$ into the same point $\mathbf{c}$ of the bidimensional output space, thus representing the core of the one-class adaptation of NNs.
The second term instead contains the Frobenius norm of the layer weights, i.e. if $w_{i,j}^{(\ell)}$ is the i-th row, j-th column term of the weight matrix $\mathrm{W}^{(\ell)}$

$$\left\|W^{\ell}\right\|_F^2 = \sum_i\sum_j\left|w_{i,j}^{(\ell)}\right|^2 \tag{6}$$

This term limits the growth (in absolute terms) of the dense layer weights, which, together with the conditions imposed on the layer structure and the non-zero value of c_i, further limits the overfitting of the healthy-only data. The hyperparameter $\lambda = 10^{-6}$ regulates the relative importance of the two components of the objective function. After training the deep one-class SVDD on healthy data, a radius R is defined based on the 99$^{\text{th}}$ percentile (p_{99}) of the squared Euclidean distance of the outputs $\mathbf{Y}_i$ from the centre $\mathbf{c}$. This choice is based on cross-validation of many different percentile settings and can be use case dependent. In the given case (run-to-failure test), false negatives (anomalies are present and not detected) are expected, as faults develop gradually with no clear failure onset. In other scenarios, where false negatives could have more severe consequences, the threshold would need to be adjusted accordingly to minimise the risk of missing critical anomalies.

$$R = \sqrt{\mathrm{p}_{99}\{\|\mathbf{Y}_i - \mathbf{c}\|^2, i = 1, \dots, N\}} \tag{7}$$

When testing a new signal with order-spectrum $\mathbf{X}_{test}$ for anomalies, the signal is classified as faulty ($C_{test} = 1$) if the distance of the network's output is larger than the radius R, and healthy ($C_{test} = 0$) otherwise:

$$C_{test} = \begin{cases} 1 & \text{if } \|\phi(\mathbf{X}_{test}; \mathcal{W}) - c\| > R \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

SHAP

In order to show which parts of the input are most relevant to the output computation of a NN, different feature importance techniques have been proposed in the literature. One of these approaches is SHAP [13]. When applied to a model such as a NN, this approach computes a SHAP value representing the relevance of each feature of the input to each feature of the output. The mathematics of SHAP are described in detail in [15].

Description of the data and its use in the AD method

The approach proposed in this paper is applied to the gear wear dataset described in [15]. The dataset was obtained on the Spur Gearbox teat-rig of the UNSW Tribology and Condition Monitoring group. A pair of healthy but non-hardened module 2 gears with 19 (input) and 52 (output) teeth were run for $3.25 \cdot 10^6$ cycles, with a vibration signal measured every 0.08 million cycles. The data was pre-processed using $N_{spr} = 100$ samples per revolution for the angular interpolation, ensuring that 2 harmonics of the gearmesh frequency are included in the resulting Nyquist range. Aliasing is prevented with an anti-aliasing filter. The resulting order spectra, without the zero-frequency term, have a length of 51 harmonics.

Tribological analysis of the surfaces conducted in parallel showed that tooth surface pitting started occurring after about $0.51 \cdot 10^6$ cycles (i.e. 15.7% of the total duration of the experiment). Considering that the surface mostly underwent initial smoothening through run-in during the first 15% of the test, the corresponding data (58 signals) were considered healthy and used to train the one-class network and initialise the value of **c** (eq. (5)**)**. The remaining 85% of the data (334 signals) is used for testing and SHAP-based interpretation. An overview of this entire process is visualised in Figure 1.

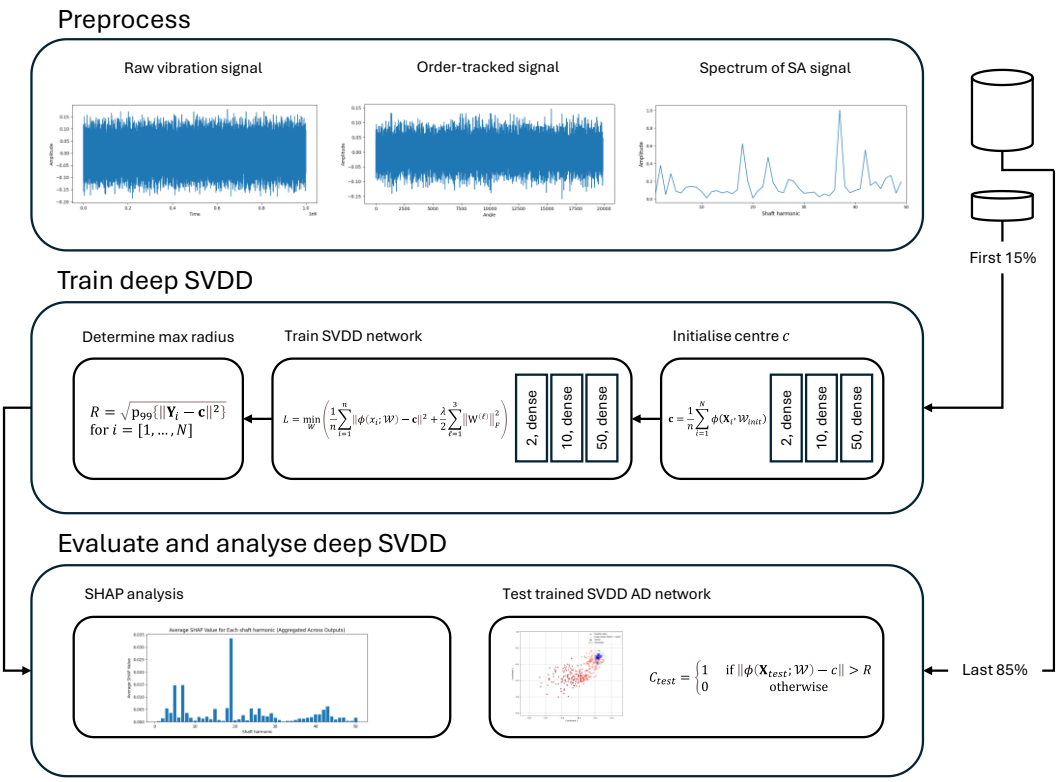

Figure 1: Overview of the proposed trustable AD gear signal method.

Results

In this section, the results of the AD and the SHAP analysis are presented. Since the arbitrary position of **c** is set based on the output of the untrained network on the healthy dataset, the random initialisation of the layer weights in the training process strongly affects it. Therefore, two examples of training runs are shown in Figure 2, resulting in two different positions of **c**.
The blue dots represent the output of the trained network for each of the training signals (healthy). The corresponding class boundary R (99^{th} percentile of the distance of the blue points from **c**) is shown as a green circle around **c**. Everything outside this circle is considered an anomaly, and only one of the training points is by definition classified as such.
The output of the network for each of the testing signals is represented by a red dot. The darkness of the dot indicates the actual progression of the test.
The results show how the method is effective in recognising a progressive drift of the data from the healthy condition, shown as a growing distance of the darker red points from the green circle. Only the very early faulty data is "misclassified" as non-anomalous, but this is not physically a problem in this situation, since there is no reason to distinguish between the last healthy signal and the first faulty one, given the gradual and relatively slow nature of the wear process.

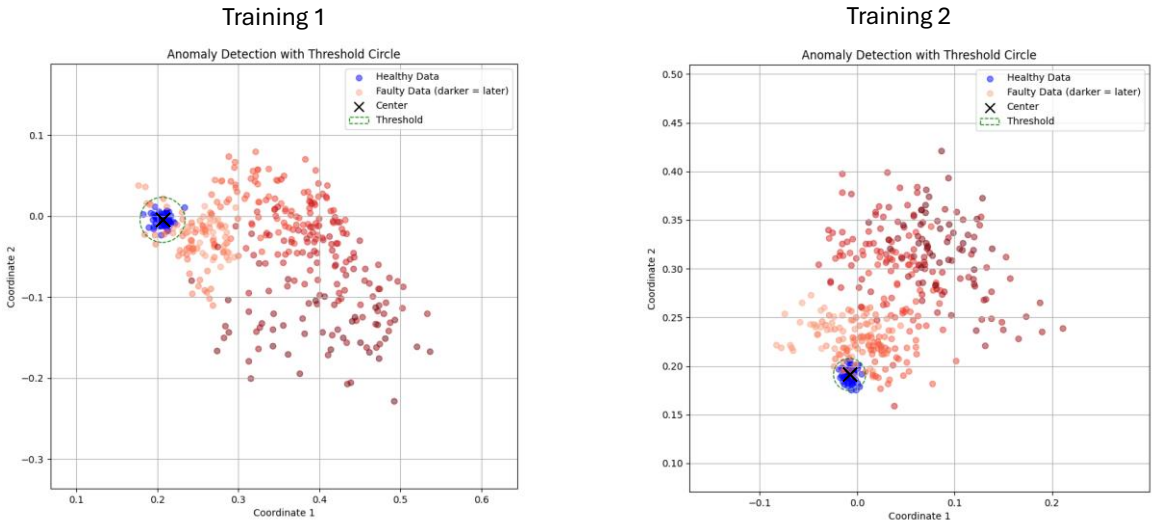

Figure 2: Deep SVDD test results for two different training runs with different random centre initialisation.

To justify the need for this relatively sophisticated approach, it is important to demonstrate that a very simple parameters such as RMS cannot be used with similar accuracy. An RMS study over the entire dataset shows that while the RMS slightly increases throughout the test, it remains in the range classified as healthy for about half of the faulty samples. The difference between maximum RMS in healthy and faulty cases is also not extreme (about 10%).
Figure 3 shows the SHAP values for each input feature (spectral harmonic) combined for the two feature outputs, i.e., the sum of the absolute SHAP values over the two output features. On the left these SHAP values are averaged over the training dataset, while on the right the average of the testing dataset is shown. Higher SHAP values for a specific shaft harmonic indicate a strong importance of that harmonic for the calculation of the two-dimensional output, and thus the classification of the signal.
Looking at Figure 3 (left), it can be seen that the relevance of the different harmonics for the training dataset is quite evenly spread, meaning that the algorithm is able to find a recurrent spectral pattern over the entire available harmonic range. The first and second gearmesh harmonics (19 and 38 shaft orders) are relatively prominent.
The results shown in Figure 3 (right) instead show a strong dominance of the first gearmesh harmonic (19^{th} shaft order), which seems to be the main contributor to the AD performed by the

algorithm. This is particularly reassuring, since it is known that evenly distributed wear such as that occurring in this experimental case is mainly affecting the first few gearmesh harmonics, thus providing physical justification to the outcomes of the algorithm.

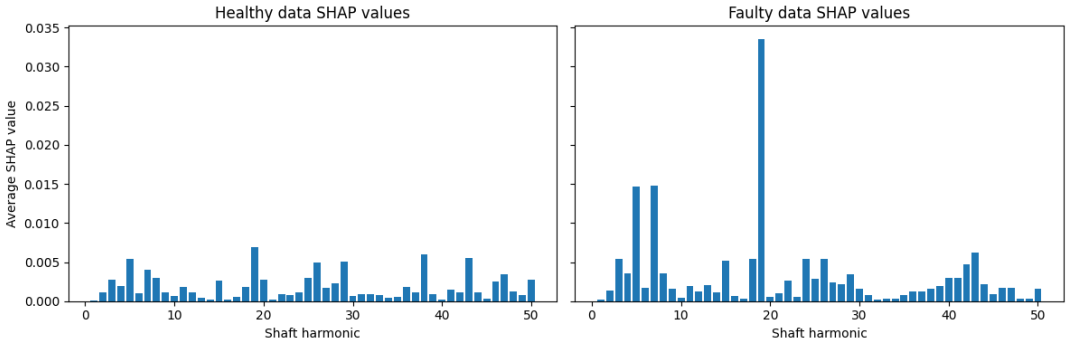

Figure 3: Average SHAP values combined for both outputs for the training data (left) and testing data (right)

To further confirm this finding, the average values of the inputs to the network, i.e., the average amplitudes of the shaft harmonics over the training and testing datasets are computed. It is clear that the 19th harmonic is the key differentiating fault symptom between the two classes.

Conclusion

In this paper a trustable gear AD framework is introduced. Starting with preprocessing the raw time domain vibration signals with order-tracking, synchronous averaging and Fourier transforming. This data is subsequently used to train the Deep-SVDD one class model with the first 15% of the run-to-failure data (considered 'healthy' data). The other 85% are used for testing the trained network (faulty data). The results show a high ability to show anomalies as the experiment progresses. Faults become more and more characteristic in the data, and this is picked up by the AD model well. Using SHAP it is shown that the model is focusing on physically meaningful signal components, such as the gearmesh harmonics, ensuring its effectiveness in detecting gear wear, offering significant potential for real-world applications in condition monitoring.

References

[1] W.J. Wang, P.D. McFadden, Early detection of gear failure by vibration analysis i. calculation of the time-frequency distribution, Mech. Syst. Signal Process. 7 (1993) 193–203. https://doi.org/10.1006/mssp.1993.1008.
[2] N.T. Du, N.P. Dien, Advanced Signal Decomposition Methods for Vibration Diagnosis of Rotating Machines: A Case Study at Variable Speed, in: N. Tien Khiem, T. Van Lien, N. Xuan Hung (Eds.), Mod. Mech. Appl., Springer Singapore, Singapore, 2022: pp. 393–400. https://doi.org/10.1007/978-981-16-3239-6_30.
[3] T.-D. Nguyen, H.-C. Nguyen, D.-H. Pham, P.-D. Nguyen, A distinguished deep learning method for gear fault classification using time–frequency representation, Discov. Appl. Sci. 6 (2024) 340. https://doi.org/10.1007/s42452-024-06033-7.
[4] J. Sharma, M.L. Mittal, G. Soni, Condition-based maintenance using machine learning and role of interpretability: a review, Int. J. Syst. Assur. Eng. Manag. 15 (2024) 1345–1360. https://doi.org/10.1007/s13198-022-01843-7.
[5] M.M. Moya, M.W. Koch, L.D. Hostetler, One-class classifier networks for target recognition applications, (1993). https://www.osti.gov/biblio/6755553.
[6] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the Support of a High-Dimensional Distribution, Neural Comput. 13 (2001) 1443–1471. https://doi.org/10.1162/089976601750264965.
[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. https://doi.org/10.1038/nature14539.
[8] S. Chaurasia, S. Goyal, M. Rajput, Outlier Detection Using Autoencoder Ensembles: A Robust Unsupervised Approach, in: 2020 Int. Conf. Contemp. Comput. Appl. IC3A, IEEE, Lucknow, India, 2020: pp. 76–80. https://doi.org/10.1109/IC3A48958.2020.233273.
[9] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, M. Kloft, Deep One-Class Classification, in: J. Dy, A. Krause (Eds.), Proc. 35th Int. Conf. Mach. Learn., PMLR, Proceedings of Machine Learning Research, 2018: pp. 4393--4402. https://proceedings.mlr.press/v80/ruff18a.html.
[10] K. Vos, Z. Peng, C. Jenkins, M.R. Shahriar, P. Borghesani, W. Wang, Vibration-based anomaly detection using LSTM/SVM approaches, Mech. Syst. Signal Process. 169 (2022) 108752. https://doi.org/10.1016/j.ymssp.2021.108752.
[11] C. Liu, K. Gryllias, A Deep Support Vector Data Description Method for Anomaly Detection in Helicopters, PHM Soc. Eur. Conf. 6 (2021) 9. https://doi.org/10.36001/phme.2021.v6i1.2957.
[12] L. Kou, J. Chen, Y. Qin, W. Mao, The Robust Multi-Scale Deep-SVDD Model for Anomaly Online Detection of Rolling Bearings, Sensors 22 (2022) 5681. https://doi.org/10.3390/s22155681.
[13] S. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, (2017). https://doi.org/10.48550/ARXIV.1705.07874.
[14] K.R. Fyfe, E.D.S. Munck, ANALYSIS OF COMPUTED ORDER TRACKING, Mech. Syst. Signal Process. 11 (1997) 187–205. https://doi.org/10.1006/mssp.1996.0056.
[15] H. Chang, P. Borghesani, W.A. Smith, Z. Peng, Application of surface replication combined with image analysis to investigate wear evolution on gear teeth – A case study, Wear 430–431 (2019) 355–368. https://doi.org/10.1016/j.wear.2019.05.024.