Frequency-Domain Feature Analysis for Early Gear Damage Detection in Planetary Gearboxes

Lisa Binanzer[1], and Martin Dazer[2]

[1,2] *University of Stuttgart, Institute of Machine Components, Stuttgart, BW, 70569, Germany*

lisa.binanzer@ima.uni-stuttgart.de
martin.dazer@ima.uni-stuttgart.de

ABSTRACT

The earliest possible detection of pitting damage in gearboxes is a central objective of vibration-based condition monitoring. Machine learning enables the automated analysis of vibration signals, but reliable detection of very early pitting damage requires a detailed understanding of which frequency ranges and frequency resolutions contain damage-relevant information. This work applies machine learning as a data-driven analysis tool to systematically quantify the relevance of frequency-based vibration features for pitting damage detection and pitting size classification. The investigations are based on experiments with three identical single-stage planetary gearboxes and four defined pitting sizes ranging from 0.5 % to 4 %. The measured time signals are transformed into the frequency domain using the Fast Fourier Transform, and the amplitudes of individual frequency bins are used as features. Since the bin width depends on the selected segment length, the influence of frequency resolution on the identification of damage-relevant features is also analyzed. A tree-based gradient boosting algorithm is used for classification, and the importance of individual frequency features is quantified by permutation analysis. The evaluation follows a two-stage approach. First, healthy and damaged states are compared to identify generally relevant frequency features. Second, the healthy state is contrasted separately with each pitting size to determine when specific features become relevant and how their importance changes with increasing damage size. In addition, feature consistency across operating conditions, sensor positions, and the three identical gearboxes is examined. The results support the targeted selection of frequency-based features for subsequent machine-learning-based damage detection and damage size classification and provide guidance for sensor placement, frequency resolution, and measurement system design in application-oriented condition monitoring.

1. INTRODUCTION

Gearboxes are critical components in many industrial drive systems, including wind turbines, manufacturing equipment, and transportation systems. Gear damage can lead to high repair costs, unplanned downtime, and complex maintenance procedures. Due to their compact design and high power density, planetary gearboxes are widely used in industrial applications. To reduce operational risks, condition monitoring systems are increasingly applied to monitor gearbox condition and to support predictive maintenance, with vibration-based monitoring representing one of the most established approaches. By enabling fault detection during operation, such systems can support timely countermeasures, such as scheduled maintenance, load reduction, or controlled shutdowns, before damage progression leads to secondary damage or unplanned downtime.

For prognostics and health management (PHM), however, damage detection alone is not sufficient. Damage must be identified at a very early stage so that countermeasures can be initiated in time. This is particularly important because PHM aims not only to detect existing faults, but also to manage damage progression. Bertsche and Dazer (2022) describe health management as the control of damage according to the objective of the PHM solution. Early detection provides the time required to avoid harmful operating conditions, adapt system operation, and support maintenance planning. In this context, Gretzinger, Lucan, Stoll and Bertsche (2020) show that adaptive operating strategies can extend the lifetime of gear wheels without unnecessary performance loss. Very early damage detection can thus contribute to utilizing the remaining useful life more effectively while reducing the risk of unplanned downtime. Among the various gear failure modes, pitting is a typical fatigue-related tooth flank damage. According to the 2016 International Organization for Standardization [ISO] report, a relevant failure criterion for case-hardened gears is a pitting area of 4 % relative to the active tooth flank area of a single tooth. The present work therefore focuses deliberately on very small pitting damages,

specifically in the range from 0.5 % to 4 %, and thus on damages at and below this standardized failure criterion.

Against this background, a key question is which vibration-based features are actually informative for the early gear damage detection. The selection of suitable diagnostic features is itself described as a central challenge in gearbox fault diagnosis. Liu, Zhao, Zuo and Xu (2014) show that feature selection can reduce the feature space while improving fault-level diagnosis for planetary gearboxes. Zuber and Bajrić (2020) likewise identify the selection of informative features for fault detection and severity classification as a major challenge in data-driven gearbox diagnosis. Frequency-domain features are particularly suitable in this context, since they directly represent the distribution of discrete spectral components and thus provide a natural basis for data-driven diagnostic methods (Wang, Li, Xin and An, 2019). In addition, Wang, Huang and Wang (2021) show that multi-criteria feature selection combined with machine learning provides an effective basis for planetary gearbox fault diagnosis.

The objective of this work is not to develop the best possible detection or classification algorithm. Instead, machine learning is used deliberately as a data-driven analysis tool to quantify the importance of individual frequency-domain features in vibration spectra of planetary gearboxes. On this basis, frequency ranges and frequency resolutions that consistently contain damage-relevant information are identified.

For this purpose, a histogram-based gradient boosting (HGB) model is employed, since it provides a suitable framework for modeling structured feature spaces and for subsequent feature-importance analysis (Friedman, 2001). To quantify the contribution of individual frequency-domain features, permutation feature importance is used. This method evaluates the importance of a feature through the loss in predictive performance after targeted permutation and therefore allows a direct assessment of its contribution to classification performance (Breiman, 2001; Altmann, Toloşi, Sander and Lengauer, 2010). In this way, the machine-learning model is used here not primarily as an end in itself, but as an analysis instrument for identifying damage-relevant spectral components.

By systematically evaluating frequency-domain features across different operating conditions, sensor positions, and identical gearboxes, this work provides a basis for the targeted selection of vibration-based features for future machine-learning-based pitting detection and damage-size classification in planetary gearboxes. The results also provide guidance for the choice of suitable frequency resolutions, sensor positions, and measurement system configurations in application-oriented vibration-based condition monitoring.

2. EXPERIMENTAL DATASET

The frequency-domain feature analysis for early gear damage detection in planetary gearboxes is conducted using test data from (FVA Forschungvereinigung Antriebstechnik e.V., 2026).

Single-stage planetary gearboxes of type SP+ 100 MF from the Alpha Advanced Line by Wittenstein (alpha Advanced Line SP+, 2025) are used. These gearboxes originate from series production and exhibit low manufacturing tolerances, ensuring good comparability of the test results. They have a transmission ratio of 10, comprise three planet gears, and feature a fixed ring gear. The planet carrier side is the low-speed side, whereas the sun side is the high-speed side.

To compare different sensor positions, the test gearbox is equipped with three vibration sensors. The vibration signals are recorded at a sampling rate f_s of 250 kHz. The sensor at position “ring gear” is mounted with a magnetic base above the gear mesh. The sensor at position “bearing” is located above the sun-side bearing, and the sensor at position “adapter” is screwed to the adapter plate. Figure 1 shows the three sensor positions on the test gearbox.

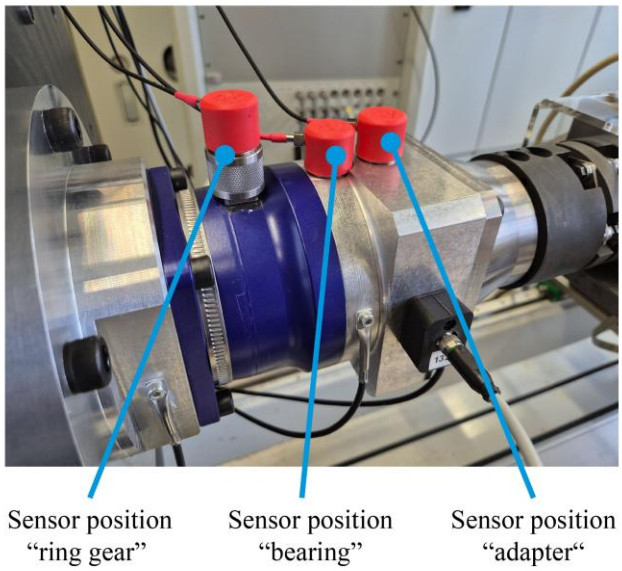

Figure 1: Sensor positions on the test gearbox.

In this study, four speed levels (610, 1140, 1670, 2200 rpm) and five torque levels (3, 6, 11, 16, 21 Nm) are considered. All combinations of these operating conditions are included in the frequency-domain feature analysis. All torque and speed values refer to the high-speed sun side.

At each operating point, a measurement of 100 s is recorded. During measurement, the gearbox temperature is maintained between 30 °C and 35 °C to keep the oil viscosity nearly constant across operating points and to avoid a significant influence on the measurement results.

Between test runs, the damage on the sun gear is progressively enlarged using a numerically controlled milling machine. Pitting sizes of 0.5 % (damage S), 1 % (damage M),

2 % (damage L), and 4 % (damage XL), each defined relative to the active tooth flank area of a single tooth, are produced.

Within the scope of the feature analysis, three gearboxes are considered. For each gearbox and each operating point, three independent datasets without damage are available. For gearbox 1, two independent datasets are available for each damage class (S, M, L, and XL), whereas for gearboxes 2 and 3, one independent dataset per damage class is available. An overview of the gearboxes, sensor positions, operating conditions, damage classes, and available datasets considered in this study is given in Table 1.

Table 1: Overview of the experimental datasets

Gearbox	1	2	3
Sensor positions	Adapter, bearing, ring gear		
Operating conditions	20 (4 speed levels, 5 torque levels)		
Independent healthy datasets per condition	3		
Damage classes	S, M, L, XL		
Independent damaged datasets per condition	2 per class	1 per class	1 per class

3. METHODS

This section presents the methodological framework used to investigate the relevance of frequency-domain vibration features for early pitting damage detection. The approach comprises data preprocessing, systematic feature screening, and the application of a histogram-based gradient boosting model. Subsequently, permutation feature importance is employed to quantify the relevance of individual frequency features. The overall workflow of the HGB-based frequency-domain feature analysis is summarized in Figure 2.

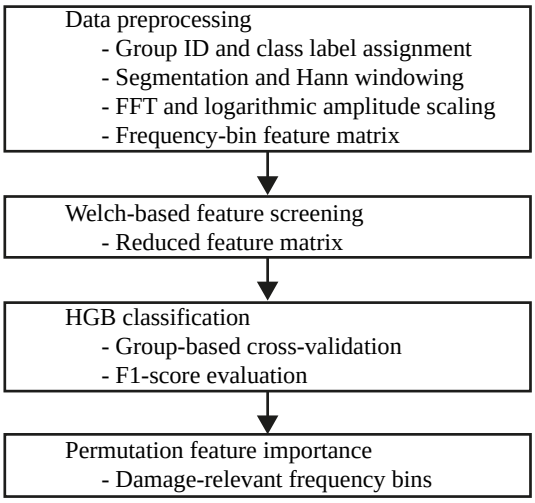

Figure 2: Workflow of the feature analysis.

3.1. Data preprocessing

The frequency-domain feature analysis is performed for each gearbox, each operating point, and each sensor. For a single analysis, independent datasets without damage and with damage are available. Each independent dataset has its own group ID. Additionally, each dataset is assigned the label “healthy (0)” or “damage (1)”.

The time-domain vibration signals of all datasets are partitioned into overlapping segments, with a 50 % overlap between adjacent segments, to generate multiple samples from each dataset while retaining local spectral information (Welch, 1967; Randall, 2021). This procedure increases the number of analyzable segments per dataset, improves statistical stability, and reduces the risk that relevant signal components are lost at the segment boundaries (Welch, 1967; Randall, 2021).

Since the subsequent Fast Fourier Transform is computed on finite signal segments, a Hann window is applied to each segment to smoothly taper the segment edges towards zero and thereby reduce spectral leakage caused by abrupt boundary discontinuities. The Hann window is selected because it provides a well-established compromise between leakage reduction and frequency resolution in spectral and vibration signal analysis (Harris, 1978; Randall, 2021).

Subsequently, each windowed segment is transformed into the frequency domain using a Fast Fourier Transform, resulting in a one-sided amplitude spectrum containing the positive frequency components from 0 Hz up to the Nyquist frequency of 125 kHz. The resulting spectrum consists of discrete frequency bins, where the width of a frequency bin is determined by the frequency resolution Δf. This is calculated from the segment length L and the sampling rate f_s:

$$\Delta f = \frac{f_s}{L} \tag{1}$$

In total, five segment lengths are investigated, resulting in five different frequency resolutions, as summarized in Table 2. The minimum segment length is selected such that, even at the lowest rotational speed, each segment still contains at least one damage overrolling event, while also corresponding to a power-of-two length for efficient FFT computation. Table 3 shows the corresponding numbers of damage overrolling events for the different rotational speeds and segment lengths. The larger segment lengths are defined as integer multiples of this minimum length in order to systematically analyze the influence of a higher number of damage overrolling events per segment and a finer frequency resolution, while still preserving computationally favorable transform sizes. This design reflects the fundamental trade-off of FFT-based spectral analysis, in which increasing the segment length improves frequency resolution but reduces temporal localization and alters the statistical properties of the segmented dataset (Welch, 1967). At the same time, the maximum segment length is limited such that a sufficient number

of segments remains available for the subsequent HGB-based evaluation.

Table 2: Frequency resolutions of the segment lengths

Segment length L	Frequency resolution Δf
$1 \cdot 2^{14} = 16384$	15.26 Hz
$2 \cdot 2^{14} = 32768$	7.63 Hz
$3 \cdot 2^{14} = 49152$	5.09 Hz
$4 \cdot 2^{14} = 65536$	3.81 Hz
$5 \cdot 2^{14} = 81920$	3.05 Hz

Table 3: Numbers of damage overrolling events per segment

Segment length L	610 rpm	1140 rpm	1670 rpm	2200 rpm
$1 \cdot 2^{14}$	1.80	3.36	4.93	6.49
$2 \cdot 2^{14}$	3.60	6.72	9.85	12.98
$3 \cdot 2^{14}$	5.40	10.09	14.78	19.46
$4 \cdot 2^{14}$	7.20	13.45	19.70	25.95
$5 \cdot 2^{14}$	9.00	16.81	24.63	32.44

To further process the spectral amplitudes, a logarithmic transformation of the form $\log(1 + A)$ is applied to the amplitudes A of frequency bins. This transformation compresses the dynamic range of the spectrum, reduces the dominance of large spectral peaks, and improves the visibility of weaker but potentially damage-relevant components, which is beneficial for vibration-based spectral analysis (Randall, 2021). The use of $\log(1 + A)$ additionally enables numerically stable handling of zero-valued amplitudes.

Each frequency bin now forms a feature, and the feature value is the log-transformed spectral amplitude of this frequency bin. By combining all segments, the feature matrix $X \in \mathbb{R}^{s \times k}$ is obtained, where s denotes the number of segments and k denotes the number of considered frequency bins. Each segment is additionally assigned its group ID and its label.

3.2. Feature Screening

After data preprocessing, a feature screening step is applied to reduce the very large number of frequency bins before the subsequent machine-learning analysis. Such screening is a well-established strategy for high-dimensional feature spaces, since it reduces dimensionality and improves computational efficiency before the subsequent learning algorithms are applied (Fan & Lv, 2008). In gearbox fault diagnosis, feature selection has been shown to reduce the feature space while improving fault-level classification performance (Liu, Zhao, Zuo and Xu, 2014). In the present study, this approach is also consistent with gradient-boosting-based workflows, where feature selection improves feature-relevance estimation while maintaining comparable predictive performance (Adler & Painsky, 2022).

Within the screening process, the group IDs are not considered. The screening is performed exclusively on the basis of the segment labels (healthy/ damage).

For each frequency feature j, the mean values of the healthy and damage segments, $\mu_h(j)$ and $\mu_d(j)$, as well as the corresponding variances, $\sigma_h^2(j)$ and $\sigma_d^2(j)$, are calculated. Based on these quantities, a Welch score $W(j)$ is determined to assess the discriminative power of each frequency feature. This score is derived from the core formulation of Welch's statistic, which is designed for comparing two groups without assuming equal variances (Welch, 1947). The score is defined as follows, where n_h and n_d denote the numbers of healthy and damage segments, respectively.

$$W(j) = \left| \frac{\mu_h(j) - \mu_d(j)}{\sqrt{\frac{\sigma_h^2(j)}{n_h} + \frac{\sigma_d^2(j)}{n_d}}} \right| \tag{2}$$

In the present work, however, no significance test is performed; instead, the Welch-based quantity is used solely as a ranking metric for univariate feature screening. Accordingly, high Welch scores indicate frequency features with large mean differences between healthy and damage segments relative to the within-class variance, whereas low scores indicate a strong overlap of the two classes or high within-class variability. After computing the Welch score for all frequency features, the 10 % of features with the highest scores are retained. These selected frequency features are then extracted from the original feature matrix X, resulting in the reduced feature matrix X_{scr}, which contains only the frequency bins with the highest Welch scores.

3.3. Histogram Gradient Boosting Model

The HGB model learns a mapping of the following form based on the feature matrix:

$$x_i \rightarrow P(y_i = 1 | x_i) = p_i \tag{3}$$

Here, x_i denotes the feature vector of the i-th segment and p_i represents the probability that damage is present in the considered segment i $(y_i = 1)$.

Gradient boosting is based on an ensemble of decision trees that is built iteratively (Friedman, 2001). This model class is well suited to the present frequency-domain feature space because tree-based boosting can capture nonlinear relationships and interactions between individual frequency bins without requiring an explicit transformation of the input features (Friedman, 2001). In contrast, linear models are limited in representing such interactions, while support vector machines and neural networks generally provide a less direct relation between individual input features and the model decision. Tree-based models have also been shown to remain highly competitive for typical tabular data compared with

deep learning approaches (Grinsztajn, Oyallon and Varoquaux, 2022; Borisov, Leemann, Seßler, Haug, Pawelczyk and Kasneci, 2024). In each boosting iteration, a new tree is fitted to reduce the errors of the current ensemble, while previously trained trees remain unchanged (Friedman, 2001).

In HGB, the continuous feature values are discretized into histogram bins before the split search, so that only a limited number of candidate split thresholds must be evaluated for each feature (Ke, Meng, Finley, Wang, Chen, Ma, Ye and Liu, 2017). This reduces the computational cost of tree construction and makes the method efficient for high-dimensional feature spaces (Ke et al., 2017). During tree construction, all available features are considered as split candidates at each node, and the split with the largest reduction in the current loss function is selected. Tree growth stops once the maximum depth is reached or no further split yields a relevant reduction in the prediction error.

The log-loss function L_i is used to quantify the deviation between the predicted damage probability p_i and the true class label y_i. For binary classification, log loss corresponds to the negative Bernoulli log-likelihood and is therefore a suitable loss function for probabilistic classification in gradient boosting (Friedman, Hastie and Tibshirani, 2000; Friedman, 2001). For a segment i, the loss is defined as follows:

$$\begin{aligned} &L_i \\ &= -(y_i \log(p_i) \\ &+ (1 - y_i) \log(1 - p_i)) \end{aligned} \tag{4}$$

Log loss penalizes confident but incorrect predictions particularly strongly. For a validation set with N segments, the mean validation loss L_{val} is calculated as follows:

$$L_{val} = \frac{1}{N} \sum_{i=1}^{N} L_i \tag{5}$$

To control model complexity and optimize model performance, empirical hyperparameter tuning is performed for the HGB model. The final hyperparameter configuration used in this study is summarized in Table 4.

Table 4: Hyperparameter HGB Model

Parameter	Value
Learning rate	0.1
Maximum iterations	100
Maximum tree depth	8
Histogram bins	255
Validation fraction	0.1
Early stopping patience	10
Early stopping tolerance	10^{-7}

Training and performance evaluation are carried out using group-based k-fold cross-validation with $k = 3$. Cross-validation is a well-established strategy for model assessment, as it enables performance estimation across multiple train-test partitions instead of relying on a single data split (Stone, 1974; Arlot &Celisse, 2010). In the present study, the group IDs of the measurement datasets are distributed across three folds, and in each fold the training set is formed from the group IDs not contained in the corresponding test set. The split is chosen such that both the training set and the test set contain at least one healthy dataset and one damage dataset. This procedure prevents segments from the same measurement dataset from appearing simultaneously in training and test data and thereby reduces the risk of data leakage. At the same time, the use of three different train-test partitions makes the analysis more robust and less dependent on a single data split.

Model evaluation is performed fold-wise by applying each trained model to the corresponding held-out test set. For each test segment, a label prediction of the form Healthy or Damage is generated and compared with the true label. With Damage defined as the positive label, the entries of the confusion matrix, i.e., True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), are determined. On this basis, precision and recall are first calculated:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Subsequently, the F1-score is determined as the harmonic mean of both quantities:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

The F1-score is suitable for the present binary classification task, as it considers both the precision of damage detection and the completeness of detection. For each of the three folds, an individual F1-score is obtained, and the final model performance is reported as the arithmetic mean of these three fold-wise F1-scores.

3.4. Permutation Feature Importance

Based on the previously determined model performance, the importance of the individual frequency features is assessed using permutation feature importance. Since each feature corresponds to one frequency bin, this method quantifies how strongly a specific frequency bin contributes to damage classification. In contrast to the F1-score, which evaluates only the overall predictive performance of the model, permutation feature importance is used to determine the contribution of individual frequency bins to this performance (Breimann, 2001; Altmann, Toloşi, Sander and Lengauer, 2010). This is also consistent with recent work in the gradient-boosting

context, where explicit feature-importance analysis is used to improve the interpretation of tree-based boosting models (Adler & Painsky, 2022).

The permutation is performed on the test dataset of the corresponding fold. To limit the computational effort, the number of evaluated test segments is restricted to 50 % of the fold-specific test data, with a minimum of 500 and a maximum of 1500 segments. For a given feature j, the amplitude values of the corresponding frequency bin are randomly permuted across the selected segments, while all remaining features are kept unchanged. As a result, the value distribution of this frequency bin is preserved, whereas its original assignment to the individual segments and their labels is removed. For each fold m, the original F1-score on the unchanged test dataset is first determined and used as the reference value $F1_m^{orig}$. Subsequently, a permuted version of the test dataset is generated for feature j, and the F1-score is recalculated, yielding $F1_{m,j,r}^{perm}$ for repetition r. To reduce random fluctuations and obtain a more robust estimate of feature importance, a total of $r = 10$ independent permutations is performed. The feature-specific importance for fold m, feature j, and repetition r is then defined by the corresponding decrease in the F1-score and is denoted as $I_{m,j,r}$:

$$I_{m,j,r} = F1_m^{orig} - F1_{m,j,r}^{perm} \tag{9}$$

A large decrease in the F1-score indicates that the corresponding frequency bin contributes substantially to the model decision, whereas little or no change indicates a low contribution. For each fold m and each feature j, the arithmetic mean of the ten repetition-specific importance values $I_{m,j,r}$ is then calculated, resulting in one fold-specific importance value. The final importance of frequency bin j, denoted as I_j^{final}, is subsequently calculated as the arithmetic mean of these three fold-specific importance values. This final quantity is used to evaluate the importance of individual frequency bins for damage classification.

3.5. Healthy vs. specific damage class

In addition to the global Healthy-vs.-Damage analysis, a damage-size-specific evaluation is performed to analyze for which damage size specific frequency features become relevant and how their importance evolves with increasing damage size. For this purpose, the healthy condition is considered separately in comparison with each individual damage class, i.e., S, M, L, and XL. For each of these binary classification tasks, the dataset is restricted to healthy segments and to segments of the respective damage class, while all other damage classes are excluded. In contrast to the global evaluation, the Welch-based screening is repeated separately for each Healthy-vs.-specific-damage-class analysis on the corresponding reduced dataset. As a result, the reduced feature matrix is determined individually for each damage class and is therefore not necessarily identical to that obtained for the global Healthy-vs.-Damage analysis. The model thus no longer learns the separation between healthy and the entirety of all damage conditions, but specifically the distinction between healthy and the respective damage class.

Accordingly, the interpretation of the evaluation metrics also changes. The F1-score now describes the detection performance for the respective damage class relative to the healthy condition. Likewise, permutation feature importance quantifies the importance of the individual frequency features specifically for this binary classification task, i.e., for distinguishing between healthy and the respective damage class.

This damage-size-specific evaluation is carried out exclusively for gearbox 1, since only for this gearbox multiple independent measurement datasets are available for each damage size. Owing to the reduced number of available group IDs in each Healthy-vs.-specific-damage analysis, only group-based k-fold cross-validation with $k = 2$ can be performed. All remaining processing steps remain unchanged compared to the global analysis.

4. RESULTS

This chapter first presents the results for the distinction between the healthy condition and all damage conditions combined in order to identify generally damage-relevant frequency features. It then considers the healthy condition in comparison with the individual damage classes separately to analyze when specific frequency features become relevant and how their relevance evolves with increasing damage size.

4.1. Healthy vs. all damages

For visualization, the permutation importances are represented as an importance density per hertz for each segment length. This quantity is obtained by dividing each permutation importance by the corresponding frequency resolution. The curves represent the frequency-dependent distribution of positive permutation importances for the different rotational speeds. To obtain condensed importance distributions, the importance density per hertz is first aggregated into 5 kHz intervals and subsequently summed over all operating conditions with identical rotational speed. This representation highlights frequency regions in which the model consistently identifies diagnostically useful information.

Figure 3 shows the feature importance density for Gearbox 1 at the ring gear sensor position with a segment length of 81,920. The highest importance values are clearly concentrated below approximately 25 kHz for all rotational speeds. Beyond this range, the importance values decrease markedly. However, no substantial differences in the overall course of the feature importance density can be observed between the individual rotational speeds across the entire frequency range.

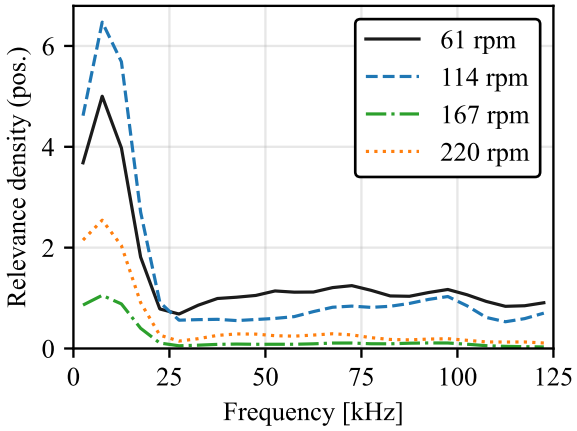

Figure 3: Gearbox 1, ring gear, L=81,920.

In contrast, the distribution at the adapter sensor position appears less uniform, as shown in Figure 4. In particular, at the lowest rotational speed, no clear trend can be identified across the frequency range. For the other rotational speeds, however, the curves likewise flatten toward higher frequencies. This decrease occurs at different frequency ranges, roughly between 20 and 40 kHz, depending on the rotational speed. A comparable behavior is observed for the bearing sensor position. At the lowest rotational speed, the importance values are distributed over almost the entire frequency spectrum. At the higher rotational speeds, the importance values decrease more clearly with increasing frequency, although the onset of this decline varies and is less consistent than at the ring gear sensor position.

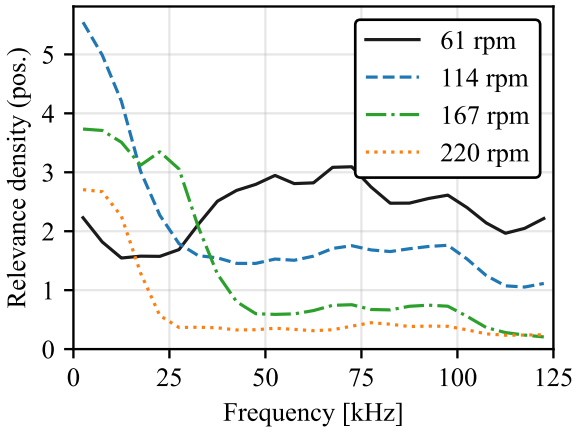

Figure 4: Gearbox 1, adapter, L=81,920.

Figure 5 shows the results for the bearing sensor position. Here, a comparatively uniform pattern is again visible, with the highest importance values occurring below approximately 25 kHz. An exception is the rotational speed of 610 rpm, for which no clear trend can be identified over the frequency range.

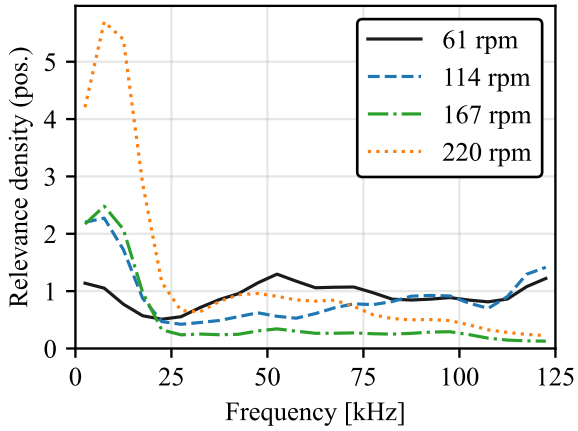

Figure 5: Gearbox 1, bearing, L=81,920.

A similar trend is observed for Gearboxes 2 and 3. At the adapter sensor position, the curves are again more diffuse and less consistent across the frequency range. In contrast, the ring gear and bearing sensor positions show a clearer and more consistent pattern, indicating that the most important frequency components are predominantly located below 25 kHz.

To provide an overall comparison across the investigated gearboxes, Figure 6 shows the positive relevance density aggregated over all sensor positions, segment lengths, rotational speeds, and torque levels. Prior to aggregation, the permutation importances are divided by the corresponding frequency resolution and grouped into 5 kHz frequency bands to enable a consistent comparison across segment lengths. The three curves represent the individual gearboxes and show that the majority of diagnostically relevant information is concentrated below approximately 25 kHz.

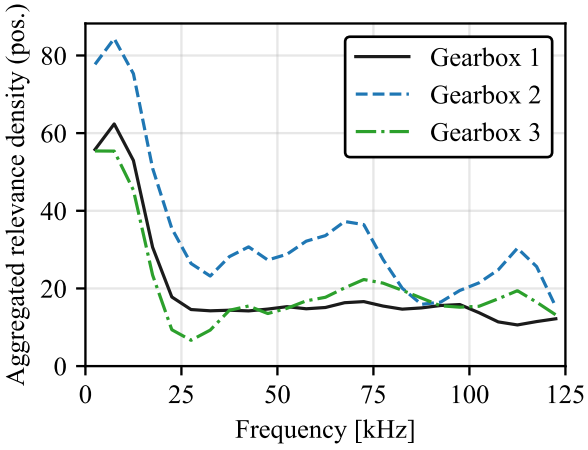

Figure 6: Aggregated relevance density across all sensor positions, segment lengths and operating conditions.

To account not only for the frequency-dependent importance distributions but also for overall model performance, the F1

scores are additionally considered. Across all gearboxes, the ring gear sensor position consistently provides very good classification performance for the discrimination between healthy and damaged conditions. Moreover, the F1 scores allow a direct comparison of the investigated segment lengths. Figure 7 presents the F1 scores for Gearbox 1 at the ring gear sensor position for all investigated segment lengths. Figure 8 and Figure 9 show the corresponding results for Gearboxes 2 and 3.

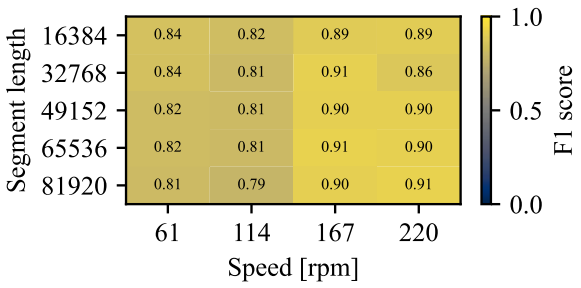

Figure 7: F1 scores, Gearbox 1, ring gear.

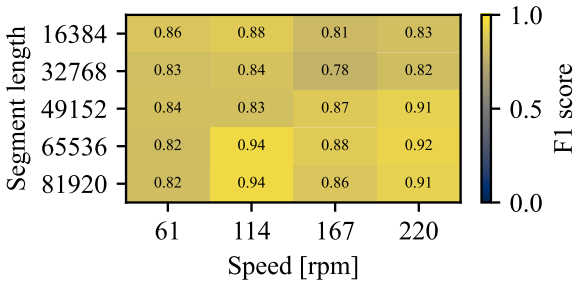

Figure 8: F1 scores, Gearbox 2, ring gear.

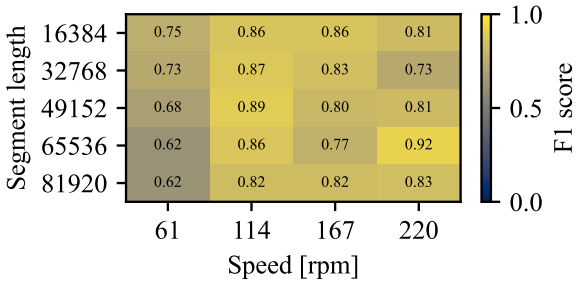

Figure 9: F1 scores, Gearbox 3, ring gear.

Overall, good classification performance is observed for all gearboxes and all segment lengths. The influence of segment length on the F1 score is smallest for Gearbox 1. Here, very good classification between healthy and damaged conditions is achieved consistently across all rotational speeds and segment lengths. Gearbox 2 also yields very good F1 scores. At the lowest rotational speed, shorter segment lengths result in higher F1 scores, whereas the tendency is reversed at the higher rotational speeds. For Gearbox 3, the lowest F1 scores are obtained at the lowest rotational speed of 610 rpm. However, the same tendency is observed at this operating point, with shorter segment lengths leading to better F1 scores. At the higher rotational speeds, the F1 scores are also good, but no clear dependence on segment length can be identified.

To additionally assess the robustness of the classification results, the standard deviation of the F1 score across the cross-validation folds is considered for Gearbox 1. This gearbox is particularly suitable for such an analysis because two independent datasets are available for each damage class, enabling a more reliable assessment of the robustness and reproducibility of the classification results. Figure 10 shows the corresponding F1 score standard deviations for all investigated segment lengths and rotational speeds. Overall, the standard deviations range between 0.04 and 0.11, indicating generally stable classification performance across the investigated operating conditions.

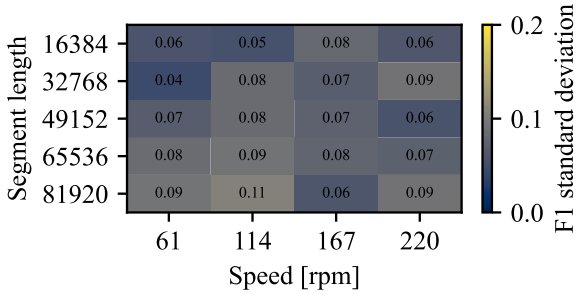

Figure 10: F1 standard deviation, Gearbox 1, ring gear.

4.2. Healthy vs. specific damage class

The damage-class-specific plots analyze the spectral signatures associated with the different damage classes. They are based on the permutation importance of the individual frequency features, which is converted into an importance density per hertz. For this purpose, each permutation importance is divided by the corresponding frequency resolution. The resulting importance densities are aggregated into frequency bins with a width of 1 kHz and summed over all operating points with identical rotational speed. For each rotational speed, the aggregated importance distributions of the different damage classes are displayed together. The plots therefore allow a qualitative analysis of how damage-related frequency regions differ between the individual damage classes. In particular, they make it possible to identify frequency regions in which the importance increases systematically with increasing damage size or in which spectral signatures shift within the frequency range. The plots thus provide indications of damage-related changes in the vibration signatures for the different damage classes.

Across rotational speeds, segment lengths, and damage classes, differences in the magnitude of the importance density can be observed. However, the overall shape of the curves remains similar. In general, the highest importance densities are located below 15 kHz, followed by a flattening toward higher frequencies. This behavior becomes more pronounced with increasing rotational speed, increasing segment length, and increasing damage size. Figure 11 illustrates this behavior exemplarily for Gearbox 1 at the ring gear sensor position, at 1140 rpm and a segment length of 81,920. When only the frequency range up to 15 kHz is considered, clear differences between the damage classes become visible. In particular, pitting size S exhibits a narrower-band importance distribution than the other damage sizes. In contrast, damage class XL shows the broadest distribution, with elevated importance values over a frequency range of approximately 3 to 15 kHz.

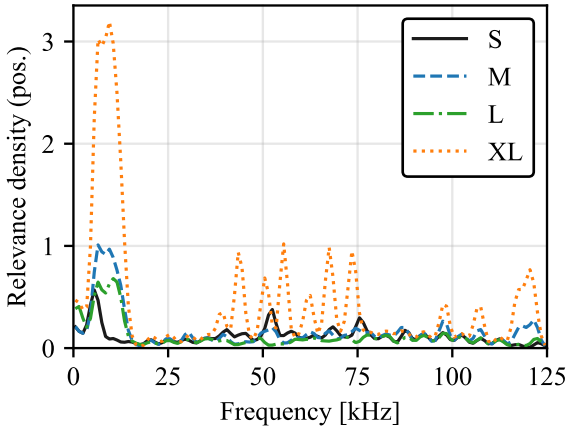

Figure 11: Gearbox 1, 1140 rpm, ring gear, L=81,920.

Figure 12 shows the corresponding results for 2200 rpm. At this higher rotational speed, the differences are again most pronounced for pitting size S. In contrast to the behavior at 1140 rpm, damage class S now also exhibits a broader-band importance distribution between approximately 3 and 15 kHz.

A more differentiated picture emerges when considering the F1 scores. Here, clear differences with respect to segment length and damage size become apparent. Figure 13 shows the F1 scores for Gearbox 1 at 1140 rpm and the ring gear sensor position. For pitting sizes S and M, the highest F1 scores are obtained at the shortest segment length. For damage class L, the best classification performance is achieved at the longest segment length. The results for damage class XL are less consistent across the investigated segment lengths. Overall, however, damage classes L and XL can be classified more reliably than S and M.

Figure 14 presents the results for 2200 rpm. At this higher rotational speed, damage class S can be detected more reliably. For damage class S, the F1 scores tend to be higher for longer segment lengths. The results for damage class M are mixed, whereas damage class L again tends to benefit from longer segment lengths. For damage class XL, in contrast, lower segment lengths tend to yield better results.

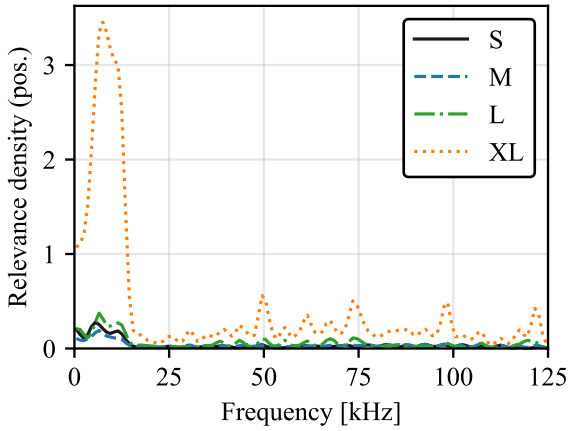

Figure 12: Gearbox 1, 2200 rpm, ring gear, L=81,920.

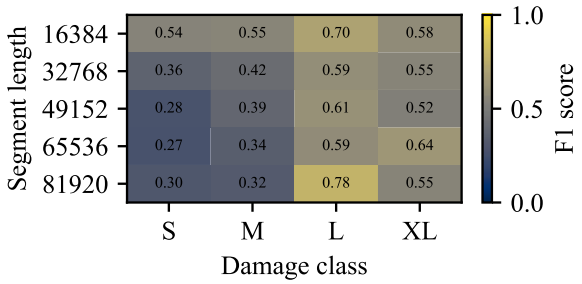

Figure 13: F1 scores, Gearbox 1, 1140 rpm, ring gear.

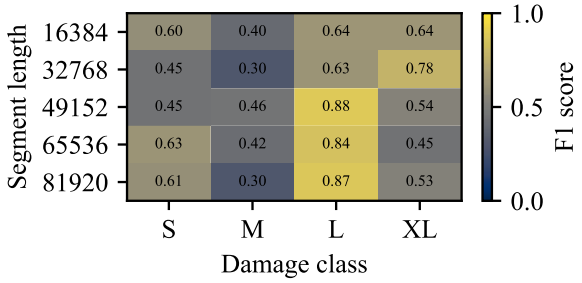

Figure 14: F1 scores, Gearbox 1, 2200 rpm, ring gear.

5. DISCUSSION

The present results show that, for the global classification task healthy vs. damage, in which all damage sizes are combined into a single damage class, the ring gear sensor position provides the clearest and most consistent frequency-dependent importance distributions. Across all investigated

rotational speeds, high importance values are concentrated below 25 kHz, followed by a pronounced decrease around 25 kHz and substantially lower values in the remaining frequency range. A similar behavior is observed for the bearing sensor position, with the exception of the lowest rotational speed of 610 rpm, at which no clear trend can be identified and the importances are distributed over a broader part of the spectrum. At the adapter sensor position, the results are less distinct overall. In particular, no clear tendency is visible at the lowest rotational speed, while at higher rotational speeds the importances decrease with increasing frequency, although not at the same characteristic threshold as observed for the other sensor positions. This general pattern is consistent across all three investigated gearboxes.

Despite these local deviations in the individual sensor- and speed-specific results, the aggregated overview (see Figure 6) reveals a much clearer overall trend for all three investigated gearboxes. While the adapter sensor position and the lowest rotational speed partly lead to more diffuse relevance distributions, their influence does not dominate the aggregated representation. Instead, the aggregation over all sensor positions, segment lengths, rotational speeds, and torque levels consistently shows for all three gearboxes that the frequency components most relevant for separating healthy and damaged states are predominantly concentrated below 25 kHz.

When the importance distributions are considered together with the classification performance, the ring gear sensor position emerges as the most suitable sensor location for the discrimination between healthy and damaged states in all investigated gearboxes. This sensor position consistently yields the highest F1 scores and thus the most reliable classification results. Overall, the F1 scores are very good, with the main exception occurring for Gearbox 3 at the lowest rotational speed. The influence of segment length on the classification performance remains limited. For Gearbox 1, no systematic influence of segment length can be identified and no specific segment length leads to consistently better classification results. For Gearboxes 2 and 3, however, shorter segment lengths lead to better performance at the lowest rotational speed. This indicates that, under these operating conditions, a less finely resolved frequency spectrum is more suitable for the classification between healthy and damaged conditions. Apart from this effect at low rotational speed, the influence of segment length is generally small and the overall classification performance remains high.

A more differentiated picture emerges for the classification task healthy vs. specific damage class. Here, the results indicate that the most important frequency components for the classification are located primarily between 3 and 15 kHz. At lower rotational speeds, the relevant frequencies for the smallest damage size S are concentrated in an even narrower range of approximately 3 to 6 kHz. The F1 scores further show that the smaller damage sizes S and M tend to be more difficult to classify at lower rotational speeds. With increasing rotational speed, the classification performance improves, particularly for damage size S. In addition, the larger damage classes L and XL can be classified more clearly than the smaller damage classes. In contrast to the global healthy-vs.-damage task, the influence of segment length is less straightforward in this case and initially appears diffuse.

To investigate this aspect further, the frequency range is restricted to 3 to 15 kHz and the F1 scores are recalculated using this reduced spectrum. At 1140 rpm, shown in Figure 15, this restriction leads to a clear increase in F1 scores across all damage sizes. Under this condition, a more distinct tendency with respect to segment length becomes visible: higher F1 scores are generally obtained for shorter segment lengths. This indicates that, at lower rotational speeds, a frequency spectrum with lower resolution is more suitable for the classification task than a very finely resolved spectrum. At 2200 rpm, shown in Figure 16, the restricted frequency range again leads to increasing F1 scores for all damage sizes. For the smallest damage size S, the highest F1 scores are obtained at the shortest segment length. For the larger damage sizes, however, the tendency changes, and the F1 scores increase with increasing segment length. This indicates that, at higher rotational speeds and with increasing damage size, more finely resolved frequency spectra are advantageous for classification.

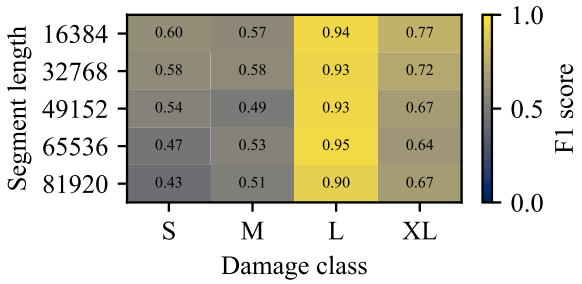

Figure 15: F1 scores, Gearbox 1, 1140 rpm, ring gear, 3-15 kHz.

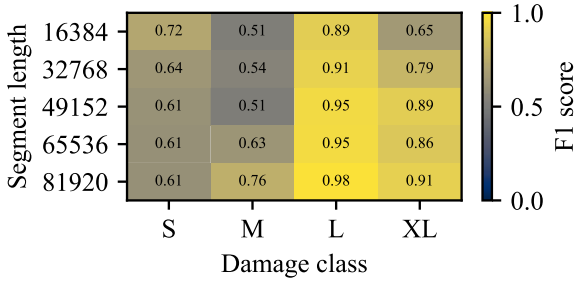

Figure 16: F1 scores, Gearbox 1, 2200 rpm, ring gear, 3-15 kHz.

Overall, the results show that the restriction of the frequency range is essential for the classification between healthy condition and specific damage classes. In addition, the choice of segment length depends on the rotational speed and therefore also on the spectral resolution required under the respective operating condition. This dependency is particularly relevant for the more demanding classification task healthy vs. specific damage class, whereas it is much less critical for the more general task healthy vs. damage.

6. CONCLUSION

In conclusion, this work shows that HGB in combination with permutation importance provides an effective data-driven framework for identifying damage-relevant spectral components for early pitting detection in planetary gearboxes. Rather than focusing on the development of an optimized classification model, the study systematically reveals which frequency ranges, sensor positions, and frequency resolutions of the recorded vibration data contain diagnostically useful information for two classification tasks: the global distinction between healthy and any damage, and the more detailed distinction between healthy and a specific damage size.

Across all investigated gearboxes, the ring gear sensor position proves to be the most suitable measurement location for the global classification task healthy vs. any damage, yielding consistently high F1 scores and showing that the most informative spectral components are predominantly located below 25 kHz. This finding is further supported by the aggregated relevance-density analysis, which reveals a consistent concentration of diagnostically relevant spectral information below 25 kHz across all three investigated gearboxes. For the more demanding classification task healthy vs. a specific damage size, the relevant information is concentrated more narrowly between 3 and 15 kHz. In this case, classification performance depends more strongly on operating condition and frequency resolution, indicating that the choice of spectral resolution should be adapted to rotational speed and diagnostic objective. In particular, coarser frequency resolution tends to be more advantageous at lower rotational speeds, whereas finer frequency resolution becomes more beneficial at higher rotational speeds and for larger damage sizes.

Overall, the results provide practical guidance for vibration-based condition monitoring of planetary gearboxes, especially for the detection and classification of very small pitting damages in the range of 0.5 % to 4 %. They show that a targeted restriction of the frequency range, an appropriate selection of sensor position, and an application-specific choice of frequency resolution can improve diagnostic performance and support the design of more effective monitoring systems for early-stage gear damage.

REFERENCES

Adler, A. I., & Painsky, A. (2022). Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy, 24(5), 687.* doi:10.3390/e24050687

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26(10),* 1340–1347. doi:10.1093/bioinformatics/btq134

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79. doi:10.1214/09-SS054

Bertsche, B., & Dazer, M. (2022). Zuverlässigkeit im Fahrzeug- und Maschinenbau: Ermittlung von Bauteil- und System-Zuverlässigkeiten. Springer. doi:10.1007/978-3-662-65024-0

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 35(6),* 7499–7519. doi:10.1109/TNNLS.2022.3229161

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. doi:10.1023/A:1010933404324

Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 70(5),* 849–911. doi:10.1111/j.1467-9868.2008.00674.x

FVA Forschungvereinigung Antriebstechnik e.V., & Binanzer, L. (2026). *FVA Heft 1698, FVA1036I, Sensorkonzept und Datenverarbeitung mittels künstlicher Intelligenz zur Frühsterkennung von Schäden und deren Lokalisation in Getriebeanwendungen* (Forschungvereinigung Antriebstechnik e.V., Ed.). Frankfurt.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29(5),* 1189–1232. doi:10.1214/aos/1013203451

Friedman, J. H., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *Annals of Statistics, 28*, 337–407. doi:10.1214/aos/1016218223

Gretzinger, Y., Lucan, K., Stoll, C., & Bertsche, B. (2020). Lifetime extension of gear wheels using an adaptive operating strategy. *In Proceedings IRF2020: 7th International Conference Integrity-Reliability-Failure* (pp. 703–710).

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *In Advances in Neural Information Processing Systems 35* (pp. 507–520). doi:10.52202/068431-0037

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE, 66(1),* 51–83. doi:10.1109/PROC.1978.10837

ISO International Organization for Standardization (Ed.). (2016). *ISO 6336-5, Calculation of load capacity of spur*

and helical gears – Part 5: Strength and quality of materials (ISO 6336-5:2016(E)).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 3149–3157), December 4–9, 2017, Long Beach, CA, United States.

Liu, Z., Zhao, X., Zuo, M. J., & Xu, H. (2014). Feature selection for fault level diagnosis of planetary gearboxes. Advances in Data Analysis and Classification, 8, 377–401. doi:10.1007/s11634-014-0168-4

Randall, R. B. (2021). Basic signal processing techniques. In R. B. Randall, Vibration-based condition monitoring (pp. 53–96). Hoboken, NJ, United States: Wiley. doi:10.1002/9781119477631.ch3

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological), 36(2)*, 111–133. doi:10.1111/j.2517-6161.1974.tb00994.x

Wang, J., Li, S., Xin, Y., & An, Z. (2019). Gear fault intelligent diagnosis based on frequency-domain feature extraction. *Journal of Vibration Engineering & Technologies, 7*, 159–166. doi:10.1007/s42417-019-00089-1

Wang, Z., Huang, H., & Wang, Y. (2021). Fault diagnosis of planetary gearbox using multi-criteria feature selection and heterogeneous ensemble learning classification. *Measurement, 173, 108654*. doi:10.1016/j.measurement.2020.108654

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34(1–2)*, 28–35. doi:10.1093/biomet/34.1-2.28

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. doi:10.1109/TAU.1967.1161901

WITTENSTEIN alpha GmbH. (2025). alpha Advanced Line SP+: Technical data sheets. Igersheim, Germany: WITTENSTEIN alpha GmbH. Retrieved from https://www.wittenstein.de/download/alpha-advanced-line-sp-de.pdf

Zuber, N., & Bajrić, R. (2020). Gearbox faults feature selection and severity classification using machine learning. *Eksploatacja i Niezawodność – Maintenance and Reliability, 22(4)*, 748–756. doi:10.17531/ein.2020.4.19