

Condition Indicator-Based Gaussian Process Regressor With Automatically Labelled Images

Rik Vaerenberg^{1,2}, Rui Zhu^{1,2}, Henrique Duarte Vieira de Sousa^{1,2}, Toby Verwimp^{1,2}, Wenzhi Liao³, Hao Wen^{1,2},
Evenela Maria Dima^{1,2}, Konstantinos Gryllias^{1,2}

¹ *Division LMSD, Department of Mechanical Engineering, KU Leuven, Leuven, Belgium*

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

³ *Flanders Make vzw, Leuven, Belgium*

ABSTRACT

The IJPHMNA 2026 data challenge aimed to predict pitting progression in gears, where the two main encountered difficulties were the lack of ground truth, with only images of the tooth flanks provided, requiring further processing for a label to be obtained, and the limited amount of given image acquisitions during training which required a model that could be trained on a low amount of labeled data.

In order to get a ground truth label, a two-stage deep learning pipeline for gear pitting detection and characterization was used (Liao & De Geest, 2024), trained using in-house image acquisitions, and fine-tuned through manually labeled images of Experiment A. In the pipeline's first stage, a YOLOv11-based segmentation model is employed to accurately segment the tooth-flank region from the input gear images, thereby removing irrelevant background regions and providing a focused region of interest for subsequent defect analysis. The tooth-flank segmentation model achieves an accuracy of more than 99%, demonstrating its effectiveness in reliably localizing the gear surface under inspection.

In the second stage, the extracted tooth-flank regions are processed using a YOLOv11 segmentation model to detect and delineate individual pitting defects at the pixel level. Compared with tooth-flank segmentation, pitting segmentation is more challenging because of the small size, irregular morphology, and relatively low contrast of pitting defects, resulting in a segmentation accuracy of approximately 70%. The generated pitting masks are subsequently fed into the characterization module, where quantitative and morphological features, including the number, area, size, location, and shape of pits, can be extracted for further assessment of gear sur-

face damage. This two-stage strategy separates surface localization from fine-grained defect segmentation, enabling more targeted and quantitative analysis of gear pitting.

The original dataset contains only sparse visual inspections, meaning that ground-truth (GT) pitting-area measurements are available for a limited number of signal samples, at the end of individual runs. To increase the number of labeled samples available for subsequent analysis and model development, these visual observations were retained as GT anchors, while signal samples between two consecutive inspections were assigned reference labels through linear interpolation with respect to cumulative elapsed time. This procedure provides a continuous estimate of pitting-area progression between observed GT points and addresses the lack of GT labels for most measurements.

Advanced cyclostationarity-based indicators have been used as the inputs for the following machine learning model. Gear-meshing signals are generally characterized as first-order cyclostationary (CS1) because of their deterministic periodicity. However, second-order cyclostationary (CS2) components are also present in gear-mesh-induced vibrations (Feng, Smith, Randall, Wu, & Peng, 2022). These components are associated with stochastic excitations arising from frictional interactions and asperity contacts during sliding between the meshing tooth surfaces. Therefore, CS2 components can provide additional information on changes in tooth-surface condition and can be exploited to construct advanced monitoring indicators. Building on this, the raw gearbox vibration signal is first decomposed. The dominant CS1 components are estimated using time synchronous averaging (TSA) and subsequently subtracted from the raw signal. The resulting residual signal contains the CS2 components associated with tooth-surface interactions. Demodulation spectra, including the squared envelope spectrum (SES) and enhanced envelope spectrum (EES), are then computed from the decomposed

Rik Vaerenberg et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

signal components (Zhu, Van Maele, Poletto, De Baets, & Gryllias, 2024). Finally, indicators are extracted by tracking the amplitudes at characteristic frequencies, such as the shaft rotational frequency, gear mesh frequency, and their harmonics. Furthermore, the indicator introduced in (Walluk et al., 2025), ALR, was also used as an input, due to its good performance in the original work.

Based on these signal processing indicators and the interpolated GT values, a Gaussian process regression (GPR) model was trained using a compound kernel created by multiplying the radial basis function and constant kernels and then adding a white kernel, giving a scaled squared-exponential function plus a noise term. GPR was chosen as it features a relatively quick training time, allowing for extensive hyperparameter optimization. In addition, a neural network was considered, but it showed a tendency to overfit. In order to select the hyperparameters of the model, two training-validation splits were made, one where experiment B was taken as validation, and one where experiment C was taken as validation. The average loss between these training and validation splits was used to optimize all the hyperparameters of the model. To keep the scales of the different indicators similar, a z-score normalization was performed. As the total training dataset was too large to fit in memory for the GPR, the training data was split into non-overlapping sets of 673 measurements, optimized using the validation dataset, which gave an ensemble of models, where the final prediction was the average of this ensemble.

Initial tests showed a significant domain shift in the indicators between the runs, which is hypothesized to be caused by the opening of the cover for acquiring the images. Following the ideas of (Walluk et al., 2025), the indicator values were subtracted by their first value in every run. Next, it was also observed that the indicators were temperature-dependent, so the initial value of every run was chosen as the first point when the temperature crossed over 95°C. Next to the indicators, the pitting area labels were then also transitioned to a relative value by subtracting the pitting area of the previous run.

This then finally resulted in an ensemble of GPR models, which took the indicator value, relative to the start of the run as input, and predicted the pitting degradation. The final degradation curve was then obtained by taking the average predicted pitting degradation of the final 10 acquisitions of a run and taking the cumulative sum over the different runs.

As it was clear from the start that the organizers' internal validation metric had different scales compared to ours, a simple scaling scheme was added. An all-zeros prediction was submitted which gave a MSE with respect to only predicting zeros. Then the final degradation curve was scaled so to have the same MSE with the zeros-baseline.

Results obtained on the training dataset showed good behavior at the start and the middle of the run, but had difficulties adjusting to the steep curve presented near the end of some runs. Other challenges include the difference in labeling scheme done between the authors and the dataset provider, leading to the scale mismatch between the output and the scoring metric. In addition the variability of the tooth flank positioning with respect to the camera made the tooth flank identification more challenging.

ACKNOWLEDGMENT

This work was supported by Flanders Make, the strategic research center for the manufacturing industry, in the context of the QED project. Additionally, the authors would like to gratefully acknowledge the support of the China Scholarship Council, the support of the Fonds Wetenschappelijk Onderzoek Vlaanderen (FWO) under the research grants no. 1SE9325N & 1SE0125N, the support of the European Commission for the Marie Skłodowska Curie program through the Horizon Europe DN PATRON project under grant agreement no. GA 101120172, and the support of the Clean Energy Transition Partnership, funded from the European Union's Horizon Europe research and innovation program under grant agreement no. GA 101069750, and VLAIO, via the Hybrid Wind project.

REFERENCES

- Feng, K., Smith, W. A., Randall, R. B., Wu, H., & Peng, Z. (2022, February). Vibration-based monitoring and prediction of surface profile change and pitting density in a spur gear wear process. *Mechanical Systems and Signal Processing*, 165, 108319. doi: 10.1016/j.ymssp.2021.108319
- Liao, W., & De Geest, R. (2024, June). Active learning for gear defect detection in gearboxes. *PHM Society European Conference*, 8(1), 10. doi: 10.36001/phme.2024.v8i1.4050
- Walluk, M., Tucker, J., Hood, A., Horney, P., Jones, A., Matthews, W., & Nenadic, N. (2025, October). Characterizing surface-damage progression of spur gears with vibration and oil data. *Annual Conference of the PHM Society*, 17(1). doi: 10.36001/phm-conf.2025.v17i1.4363
- Zhu, R., Van Maele, D., Poletto, J. C., De Baets, P., & Gryllias, K. (2024). Development and evaluation of vibration indicators for gear pitting detection. In *Proceedings of the international conference on noise and vibration engineering (isma 2024) and international conference on uncertainty in structural dynamics (usd 2024)* (pp. 508–521). Leuven, Belgium: KU Leuven, Department of Mechanical Engineering.