An Introduction to 2023 PHM Data Challenge: The Elephant in the Room and an Analysis of Competition Results

Yongzhi Qu[1], Jesse William[2], Abhinav Saxena[3], Neil Eklund[4], and Scott Clements[5]

[1]*University of Utah, Salt Lake City, UT, 84112 USA*
yongzhi.qu@utah.edu
[2]*Global Technology Connection, Inc., Atlanta, GA, 30339 USA*
jwilliams@globaltechinc.com
[3]*GE Aerospace Research, Niskayuna, NY, 12309 USA*
asaxena@ge.com
[4] *Oak Grove Analytics, LLC, Philadelphia, PA 19151 USA*
neil.eklund@gmail.com
[5]*Lockheed Martin, Fort Worth, TX, 76108 USA*
n.s.clements@ieee.org

ABSTRACT

The trend in diagnostics and prognostics for PHM is shifting toward explainable data-driven models. However, complex engineered systems are typically challenging to develop entirely explainable models for, whether they are grounded in physics or data-driven techniques. Consequently, the development of machine learning models, including hybrid variants capable of both interpolation and extrapolation, holds significant promise for enhancing the practicality of system simulation, analysis, modeling, and control in industry. The primary objective of this data challenge is to encourage contributions that expand the scope of model generalization beyond the training domain. The second aim of this data challenge is to quantify model uncertainty and methods to incorporate it into predictions. For most PHM tasks, clear guidance of the required action is ideal. To issue a definitive guidance to end users, it is useful to quantify uncertainty for the whole model. This data challenge addresses both estimation and uncertainty.

1. OBJECTIVE

This year's data challenge focuses on estimating gearbox degradation levels in a gearbox operated under a variety of conditions. Participants are scored based on both accuracy and confidence derived from estimated uncertainties in estimation.

2. DATA CHALLENGE TASK

Although a challenging task for many data-driven approaches, to be practical for real-world applications, a model should generalize to previously unseen operational conditions and fault levels. Participants are also required to express measures of confidence in model predictions. Such confidence measures might be used to determine whether these predictions can be trusted or not before taking any downstream actions.

The overall data challenge task is to develop a fault severity estimate using the data provided. The training dataset includes measurements under varied operating conditions from a healthy state as well as six known fault levels. The testing and validation datasets contain data from eleven health levels, which include a healthy state and 10 degradation/fault levels. Data from some fault levels and operating conditions are excluded from the training datasets to mirror real-world conditions where data may only be available from a subset of the operating envelope. The training data are collected from a range of different operating conditions under 15 different rotational speeds and six different torque levels, while the test and validation operating conditions span 18 different rotational speeds and six different torque levels.

The data challenge requires fault level estimation for three regimes of the operational envelope:

1. Samples from conditions seen in the training dataset.

2. Samples from conditions not seen in the training dataset, but within the range of operational conditions and fault levels seen in the training set, i.e., interpolation.

3. Fault level estimation from conditions not seen in the training dataset and outside the training range for fault levels, i.e., extrapolation.

Both a fault level estimate and a corresponding confidence level are required from the model. Such confidence may be used in deciding whether a prediction should lead to an action reconfiguration, (inspection, repair, etc.) or no action if the confidence was below pre-decided acceptable threshold. In real settings such thresholds would be determined based on operational risks and business models, however, this challenge requires participants to focus on developing methods to assess confidence in their models and implicitly learn thresholds such that overall accuracy can be maximized. Accuracy calculation with incorporating confidence is explained in Section 4.

3. PROBLEM AND DATA DESCRIPTION

3.1. Experimental setup

A brief overview of the data collection process is provided here. Full details are provided in the papers referenced (Li, Qu, and Nichifor, *et al.*, 2018-2022).

The gear pitting experiments were performed on a one-stage gearbox installed in an electronically closed transmission test rig. The gearbox test rig includes two 45 kW Siemens servo motors. One of the motors can act as the driving motor while the other can be configured as the load motor. Motor 1 is the driving motor in this experiment. The overall gearbox test rig, excluding the control system, is shown in Fig. 1. The testing gearbox is a one stage gearbox with spur gears. The gearbox has a speed reduction rate of 1.8:1. The input driving gear has 40 teeth, and the driven gear has 72 teeth. Detailed gear parameters are provided in Table 1.

A tri-axial accelerometer was attached on the gearbox case close to the bearing house on the output end as shown in Figure 3. X, Y, Z are horizontal, axial and vertical, separately.

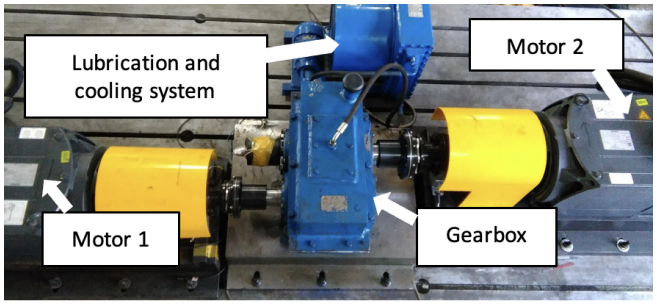

Figure 1. Experiment test rig for gearbox dynamic meshing stiffness analysis

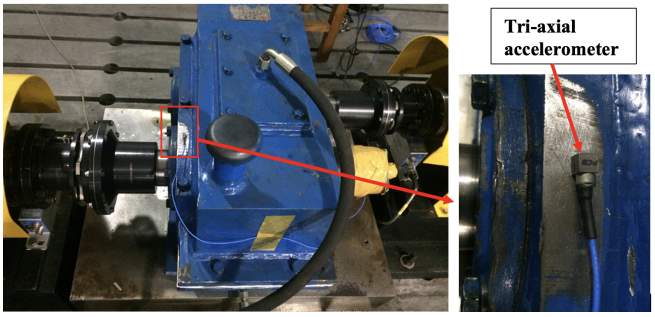

Figure 2. Vibration measurement with Tri-axial accelerometer

Table 1. List of gear parameters for the tested gearbox

Gear parameter	Driving gear	Driven gear
Tooth number	40	72
Module	3 mm	3 mm
Base circle diameter	112.763 mm	202.974 mm
Pitch diameter	120 mm	216 mm
Pressure angle	20°	20°
Addendum coefficient	1	1
Coefficient of top clearance	0.25	0.25
Diametral pitch	8.4667	8.4667
Engaged angle	19.7828°	19.7828°
Circular pitch	9.42478 mm	9.42478 mm
Addendum	4.5 mm	3.588 mm
Dedendum	2.25 mm	3.162 mm
Addendum modification coefficient	0.5	0.196
Addendum modification	1.5 mm	0.588 mm
Fillet radius	0.9 mm	0.9 mm
Tooth thickness	5.8043 mm	5.1404 mm
Tooth width	85 mm	85 mm
Theoretical center distance	168 mm	168 mm
Actual center distance	170.002 mm	170.002 mm

Both healthy and gradually pitted gear under various operating conditions were tested and the vibration signals collected. Five sets of data were collected. Symbol '●' indicates that the data samples for this setting are provided for training while '∘' indicates the data are hidden from training but will appear in testing and validation.

One or more gear teeth are manually degraded using a drill bit through the lube oil cover without any disassembly and assembly of the gearbox or test rig. Degradation severity increases in levels from 0 to 10.

To sample enough data points in terms of revolutions, longer time series data are collected for lower rotational speed conditions. For 100-200 rpm, the sampling time is about 12s; for 300-1000 rpm, the sampling time is about 6s; for 1200 rpm and above, the sampling time is about 3s.

All data are sampled with a sampling rate of 20,480 Hz. horizontal, axial, and vertical accelerometers were sampled separately, along with a tachometer signal. All signals were time synchronized.

The tachometer (laser reflective tachometer) outputs one pulse per revolution. This is measured on the output shaft, which is 5/9 of the input shaft speed. For example, for input shaft speed at 35Hz and output shaft speed will be 19.44 Hz.

3.2. **Data Description**

Out of all data marked as black (78 operating conditions throughout 7 health levels), on average, 3.69 repetitions for each operating condition of each fault level are included in the training data set. A total of 2016 data files were included in the training. Pitting degradation levels 5, 7, 9, 10 are omitted from the training data set.

Table 2. Operation conditions of the experiments (low speed)

Speed & Torque	100	200	300	400	500	600	700	800	900	1000
50	●	●	●	●	●	●	●	●	●	●
100	●	●	●	●	●	●	●	●	●	●
200	●	●	●	●	●	●	●	●	●	●
300	●	●	●	●	●	●	●	●	●	●
400	●	●	●	●	●	●	●	●	●	●
500	●	●	●	●	●	●	●	●	●	●

Table 3. Operation conditions of the experiments (median to high speed)

Speed & Torque	1200	1500	1800	2100	2400	2700	3000	3600
50	●	○	○	●	○	●	●	●
100	●	○	○	●	○	●	●	●
200	●	○	○	●	○	●	●	
300	●	○	○	●	○			
400	●	○	○					
500	●	○						

4. EVALUATION METRICS

For the submission of data challenge results, a probability based prediction is required for each predicted label. The probability can be distributed across multiple labels, with a sum of probability equal to or less than one. A binary confidence level is required to be included for the label classification & prediction for each sample, with 0 and 1 mean low and high confidence, respectively. The exact rewards or penalty also depends on the how far the predicted label is from the underlying true label as specified in the following:

$$SS_{total} = \sum_{i=1}^{n} w_i S_i$$

where, SS_{total} is the total score, n is total number of testing/validation samples, $w_i = 1$ for confidence level of 1, $w_i = 0.2$ for confidence level of 0, $S_i = \sum_{k=0}^{10} Q_{i,k} P_{i,k}$, with $Q_{i,k}$ equals to the score for prediction of sample i at distance k, while $P_{i,k}$, is the reported probability the label at the distance k as shown in Table 4. For example, the ideal prediction S_i will have a $P_{i,k}$ =100% at distance 0, which means $P_{i,0} = 1$ for $k = 0$, $P_{i,k} = 0$ for $k \neq 0$. In this case, we have $S_i = 1$, otherwise, $S_i < 1$. Accordingly, the highest total score will be n depending on the testing and validation data size. In this data challenge, the highest possible testing score is 800, and highest validation score is 812.

Table 4. Prediction score for each sample based on the distance from the true labels.

Distance from true label (k)	Points (Q)	Reported prediction Probability (P)
0 (correct prediction)	1.0	P_0
1	0.5	P_1
2	0	P_2
3	-0.5	P_3
4	-1.0	P_4
5	-1.5	P_5
6	-2.0	P_6
7	-2.5	P_7
8	-3.0	P_8
9	-3.5	P_9
10	-4.0	P_{10}

A high confidence level will be scored with a higher weight for the final sum of score, while a low confidence level will be scored with a lower weight. Similarly, a wrong prediction with a high confidence level will also be graded with a higher penalty.

5. SUMMARY

A total of 52 teams registered and 20 teams completed the data challenge. The final scores of all twenty are summarized in Table 5. Top ten teams were invited to submit a brief description of the technical approach taken. A panel of experts evaluated the summaries independently on criteria of data-preprocessing steps, algorithmic novelty, treatment of uncertainty, and creativity. A final score incorporating test set performance and method scores was used to identify top five

finalists. Winners will be chosen from finalists based on conference presentations on their detailed approach and discussions.

For readers' reference, the following list indicates different approaches adopted by five teams out of the top ten finalists. They are the ones who registered for the conference for presentation and submitted their summary to the conference proceedings. Further details of the methodologies can be found in data challenge summary papers included in the PHM 2023 conference proceedings:

1. *nivic*: Gear Pitting Fault Diagnosis using Domain Generalizations and Specialization Techniques (Chu *et al*., 2023)

2. *Thumper*: Interpolate and Extrapolate Machine Learning Models using An Unsupervised Method (Liu, 2023)

3. *KUL*: Predicting pitting severity in gearboxes under unseen operating conditions and fault severities using convolutional neural networks with power spectral density inputs. (Vaerenberg *et al*., 2023)

4. *Amitory*: Anomaly Detection and Fault Classification in Multivariate Time Series Using Multimodal Deep Models. (Ryu *et al*., 2023)

5. *zwang1916*: Gearbox Degradation Prediction through Deep CNN and Bayesian Optimization.(Shen *et al*., 2023)

At the time when the data challenge was closed, the highest testing score was 463.5, and the highest validation score was 472. A further analysis on the validation score of 472/812 reveals that this score corresponds following performance: for machine learning metric **precision at k**, precision at 1 = **66.38%**, precision at 2 = **86.70%**, and precision at 3 = **98.15%**. That means the top performing team has correctly predicted **98.15%** of the sample with a label within the error distance of 2, which is a very impressive result.

Table 5. PHM 2023 Conference Data Challenge Final Scores

Teams	Validation Score	Test Score
nivic	472.0	463.5
CUMTIIPT	418.5	404.5
KUL	282.2	213.3
thumper	227.5	249.0
SDML	192.0	191.9
Aimtory	187.5	200.5
TJ	178.0	186.5
JSEG	158.9	212.0
FlyTogether	143.0	232.0
zwang1916	130.1	67.43
jbarriga	125.5	190.8
fjamil	122.5	150.6
ILIKEPHM	115.4	121.0
TeamSSS	94.44	145.8
beking	38.0	195.5
polimeca	24.5	160.2
neaolei1	-10.5	213.0
wk_team	-168.8	-197.3
S5	-265.0	-177.5
teamoslo	-2.495e+03	-2.378e+03

ACKNOWLEDGEMENT

Y. Qu acknowledges Wuhan University of Technology for providing the testing facility and the equipment, and National Science Foundation of China (51505353), 2016 – 2018, for partial funding support for data collection.

REFERENCES

Jialin Li, Renxiang Chen, Xianzhen Huang, Yongzhi Qu, (2022), Development of Deep Residual Neural Networks for Gear Pitting Fault Diagnosis Using Bayesian Optimization, *IEEE Transactions on Instrumentation & Measurement*. Vol. 71.

Alex Nichifor, Yongzhi Qu. (2021), Koopman Operator Based Fault Diagnostic Methods for Mechanical Systems, *Proceedings of International workshop on structural health monitoring (SHM)*, Stanford University, Palo Alto, CA, 2021

Xueyi Li, Jialin Li, Chengying Zhao, Yongzhi Qu, David He, (2020), Gear pitting fault diagnosis with mixed operating conditions based on adaptive 1D separable convolution with residual connection, *Mechanical Systems and Signal Processing*, Vol. 142.

Jialin Li, Xueyi Li, David He, Yongzhi Qu, (2020) Unsupervised rotating machinery fault diagnosis method based on integrated SAEDBN and a binary processor, *Journal of Intelligent Manufacturing*, Vol. 31.

Jialin Li, Xueyi Li, David He, Yongzhi Qu, (2020), A domain adaptation model for early gear pitting fault diagnosis based on deep transfer learning network, *Journal of Risk and Reliability*, Vol. 234, No. 1, pp. 168-182.

Yongzhi Qu, Yue Zhang, Miao He, David He, Chen Jiao, and Zude Zhou, (2019), Gear pitting fault diagnosis using disentangled features from unsupervised deep learning, *Journal of Risk and Reliability*, Vol. 233, No. 5, pp. 719-730.

Xueyi Li, Jialin Li, Yongzhi, Qu, David He, (2019), Gear Pitting Fault Diagnosis Using Integrated CNN and GRU Network with Both Vibration and Acoustic Emission Signals, *Applied Sciences*, Vol. 9, No. 4 (768).

Jialin Li, Xueyi Li, David He, Yongzhi Qu, (2019), A Novel Method for Early Gear Pitting Fault Diagnosis Using Stacked SAE and GBRBM, *Sensors*, Vol. 19, No. 4 (758).

Yongzhi Qu, Yue Zhang, David He, Miao He, Zude Zhou, (2019), A Regularized Deep Clustering Method for Fault Trend Analysis, *Proceedings of the Annual Conference of the PHM Society*, Vol. 11, No. 1, Scottsdale, AZ, USA, September 21-26.

Xueyi Li, Zhendong Liu, Yongzhi Qu, and David He, (2018), Unsupervised Gear Fault Diagnosis Using Raw Vibration Signal Based on Deep Learning, *IEEE Conference on Prognostics and System Health Monitoring*, Chongqing, Oct. 26-28, Chongqing.

Fan Chu, Lei Wang, Yuxuan Zhou, Ronghuan Zhao, Jiangzhao Wu, and Meng Lei, (2023), Gear Pitting Fault Diagnosis Using Domain Generalizations and Specialization Techniques, *Proceedings of the Annual Conference of the PHM Society*, Vol. 11, No. 1, Salt Lake City, UT, USA, Oct. 28- Nov.2.

Peng Liu, (2023), Interpolate and Extrapolate Machine Learning Models using An Unsupervised Method, *Proceedings of the Annual Conference of the PHM Society*, Vol. 11, No. 1, Salt Lake City, UT, USA, Oct. 28- Nov.2.

Rik Vaerenberg, Douw Marx, Seyed Ali Hosseinli, Fabrizio De Fabritiis, Hao Wen, Rui Zhu and Konstantinos Gryllias. (2023), Predicting pitting severity in gearboxes under unseen operating conditions and fault severities using convolutional neural networks with power spectral density inputs. *Proceedings of the Annual Conference of the PHM Society*, Vol. 11, No. 1, Salt Lake City, UT, USA, Oct. 28- Nov.2.

Gunwoo Ryu and Nohyoon Seong, (2023), Anomaly Detection and Fault Classification in Multivariate Time Series Using Multimodal Deep Models, *Proceedings of the Annual Conference of the PHM Society*, Vol. 11, No. 1, Salt Lake City, UT, USA, Oct. 28- Nov.2.

Kai (Kevin) Shen, Haoyu Wang, Yuwei Liao, and Dev Kakde, (2023), Gearbox Degradation Prediction through Deep CNN and Bayesian Optimization, *Proceedings of the Annual Conference of the PHM Society*, Vol. 11, No. 1, Salt Lake City, UT, USA, Oct. 28- Nov.2.