Autoencoder based anomaly detector for gear tooth bending fatigue cracks

Adrian Hood[1], Christopher Valant[2], Patrick Horney[3], Allen Jones[4], Jared S. Lantner[5], Josiah Martuscello[6], Nenad Nenadic[7]

[1] *DEVCOM Army Research Laboratory, Aberdeen Proving Ground, MD, 21005, USA*
adrian.a.hood.civ@mail.mil
[2,5,6,7] *Rochester Institute of Technology, Rochester, NY, 14623, USA*
cxvgis@rit.edu
jqmgis@rit.edu
jslgis@rit.edu
nxnasp@rit.edu
[3] *NAVAIR, Cherry Point, NC, 28533, USA*
patrick.horney@navy.mil
[4] *NAWCAD, Patuxent River, MD, 20670, USA*
allen.jones2@navy.mil

Abstract

This article reports on anomaly detection performance of data-driven models based on a few selected autoencoder topologies and compares them to the performance of a set of popular classical vibration-based condition indicators. The evaluation of these models employed data that consisted of baseline gearbox runs and the associated runs with seeded bending cracks in the root of the gear teeth for eight different gear pairings. The analyses showed that the data-driven models, trained on a subset of baseline data, outperformed classical condition indicators as anomaly detectors and may show promise for damage assessment.

1. Background

Condition monitoring of gearboxes aims to increase component life, vehicle readiness, and reduce operation and maintenance costs. Vibration-based *Conditional Indicators* (CIs) that reliably track damage severity are sought, allowing, not only detection, but life predictions. There are several excellent comprehensive reviews of vibration-based CIs (Lebold, McClintic, Campbell, Byington, & Maynard, 2000; Samuel & Pines, 2005; G. Jinks, 2016; Sharma & Parey, 2016b). Of particular interest to this study are NP4 is the normalized kurtosis of the signal power computed from Wigner-Ville transform (Polyshchuk, Choy, & Braun, 2002); NA4, a kurtosis of the residual signal (Zakrajsek, Townsend, & Decker, 1993), FM4 (Stewart, 1977), M6A/M8A (Martin, 1989), and Energy Ratio and Crest Factor (Swansson, 1980).

The focus of this paper is the gear tooth root crack failure mode. Gear tooth root cracks manifest as changes in gear mesh stiffness which changes the gearbox's vibration characteristics. Both analytical models (Chaari, Fakhfakh, & Haddar, 2009; Chen & Shao, 2011; Liang, Zuo, & Hoseini, 2015) and numerical models (Cooley, Hood, & Wang, 2021) have been developed to better understand the relationship between crack size and the resulting acceleration. (Nenadic, Wodenscheck, Thurston, & Lewicki, 2011) conducted a series of experiments to develop a database of seeded fault experiments that carefully tracked crack size and gearbox housing acceleration. This data serves the purpose of model validation and diagnostic algorithm development. While analytical models have suggested a monotonic change in classical CIs with crack growth, this was not consistently observed in our experiments across multiple test gears (Nenadic et al., 2013). This inconsistency has also been observed by others. For example, (Sharma & Parey, 2016a) calculated condition indicators for three spur gears tests with different crack sizes. Wire Electrical Discharge Machine (EDM) was used to introduced different sized flaws into two of the gears and results were provided for different speed fluctuations. They found that the classical CIs did not perform well with increasing damage for all speed fluctuations. They introduced two new CIs, PS-I and PS-II that were able to track the test cases and showed promise, however only one gear at each damage level was tested. (Bechhoefer & Butterworth, 2019) also found many CIs performed poorly on their own when analyzing three undamaged gearboxes and one with a cracked spiral bevel gear tooth. They created several health indexes (HIs), derived from different combinations of 88 CIs. They found that using the CIs with the largest statistical separability increases the sensitivity of the HI to the crack.

This works attempts to improve on classical CI performance using machine learning tools, in particular, the autoencoder topology, to develop data driven condition indicators used to detect fatigue induced gear tooth cracks in spur gears across multiple baseline and damage cases.

The principal *prognostic health management* (PHM) capabilities are, in increasing order: anomaly detection , diagnostics, and prognostics (Vachtsevanos, Lewis, Roemer, Hess, & Wu, 2006; Goebel et al., 2017). The lowest level of PHM capability, anomaly detection is very important in its own right, and can, over-time, be used to attain higher levels of PHM capability (Sikorska, Hodkiewicz, & Ma, 2011).

A successful implementation and deployment of an autoencoder predates the emergence of *deep learning* (Japkowicz, Myers, & Gluck, 1995). More recently, autoencoder-based anomaly detectors have been shown to have considerable promise because, unlike classical classifiers that demand balanced datasets, their training can be based on data associated with normal operation, which comes in abundance, as opposed to data associated with failures, which is difficult to come by (Eklund, 2018; Yan & Yu, 2015). The performance of autoencoders was compared to classical CIs.

2. DATA GENERATION

Data used for the modeling and evaluation was custom-generated over a sequence of gearbox runs. The block diagram in Figure 1 depicts the test process at a high level. The process consisted of the following steps: 1) break-in (low-torque, low-speed) eight hour run, 2) baseline (nominal torque, nominal speed) two hour run 3) crack initiation, 4) crack verification, 5) installation of *crack-propagation* (CP) sensors, and 6) crack propagation until failure. Cracks were initiated using a fatigue tester, using the previously-developed process described in (Nenadic et al., 2011).

The main dataset consists of eight tests, each denoted by the label of the gear with a cracked tooth, viz. Gear 207 for gear pair 207/208 and Gear 209 for pair 209/210, etc. All the gears were the same new NASA-designed spur gears (NASA, 1994). The number of gears employed in the experiment was limited by the time required to successfully conduct the labor-intensive experiments that require multiple gearbox assemblies following a detailed checklist of measurements, crack initiation, and crack verification (see Figure 1). Acceleration data, along with the tachometer and CP sensor data, was captured at 100 kHz sampling rate and saved in separate files representing one-second of data. The accelorometer placement was based on a previous study that employed the same gearbox (Nenadic et al., 2013).

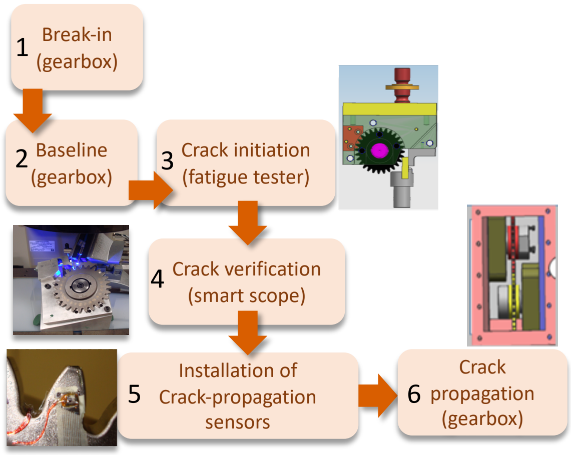
1 Break-in (gearbox)
2 Baseline (gearbox)
3 Crack initiation (fatigue tester)
4 Crack verification (smart scope)
5 Installation of Crack-propagation sensors
6 Crack propagation (gearbox)

Figure 1. Block diagram of the test process.

Figure 2 depicts the operating conditions associated with a two-hour baseline test: the torque and speed are held constant but the temperature exhibits a transient as no oil preheating was employed.

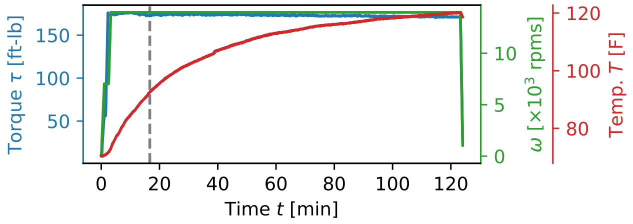

Figure 2. Operating conditions during a typical baseline run: torque, speed, and temperature.

Figure 3 shows the acceleration and tachometer data associated with the dashed line in Figure 2. The data was sampled at 100 kHz for 1 second.

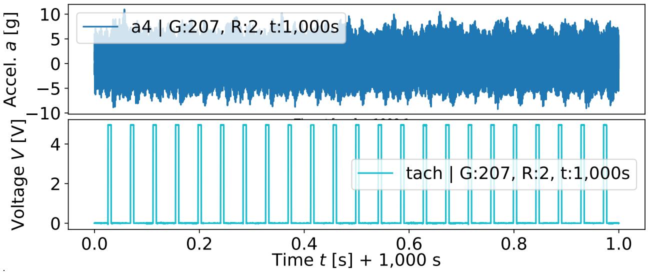

Figure 3. Acceleration and tachometer sampled at 100 kHz.

The fixture is equipped with multiple accelerometers, as shown in Figure 4, but only accelerometer 4 was used in this

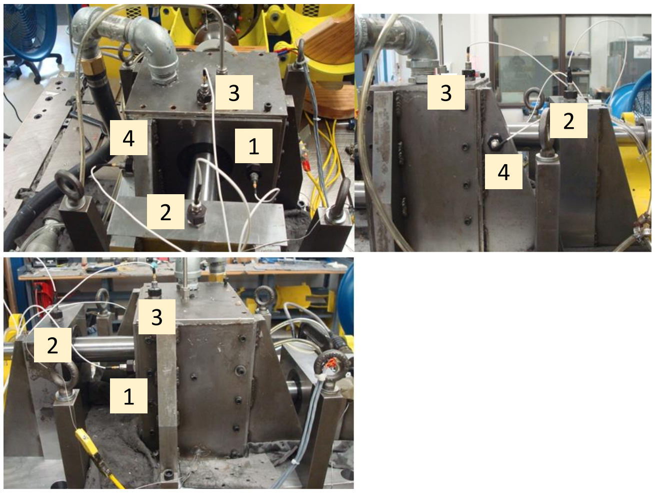

Figure 4. Gearbox with accelerometer locations. Only accelerometer 4 was used in the study.

study because it has been previously shown that this location is the most sensitive to crack detection on the test stand (Nenadic et al., 2013).

The propagation test employed the same operating conditions as the baseline test, however, the duration varied due to different failure times. Failure is defined as when all strands are broken on both CP sensors.

An example of a propagation test that lasted 54 minutes is given in Figure 5. Also shown are both CP sensor outputs that were processed and interpreted as *damage levels*, DL_1 and DL_2.

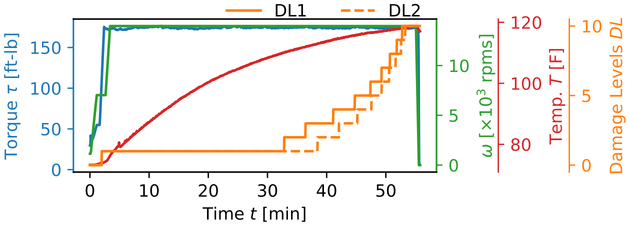

Figure 5. An example of a crack-propagation test.

3. Modeling

The vibration data was processed using *time synchronous averaging* (TSA), which used a tachometer to average over multiple shaft rotations, effectively converting time-domain data into angle domain data in the range $0 \leq \theta < 2\pi$. An accelerometer vector, $\boldsymbol{a}(t)$, with length of 100,000, associated with one second of operation, was compressed into the TSA vector $\boldsymbol{x}_{TSA}(\theta)$ with a length of 4,096 points. Twenty-four revolutions were used for the average.

$$\boldsymbol{a}(t) \rightarrow \boldsymbol{x}_{TSA}(\theta) \tag{1}$$

TSA compresses and smooths raw accelerometer data and is a preprocessing step employed by many common vibration CIs (e.g. FM0, NA4, FM4, M6A, NP4, etc.) (Lebold et al., 2000) and it was employed as the input of autoencoders in this study.

To facilitate data-driven development, the data associated with each baseline and propagation test was organized in HDF5 files that contained a matrix of TSA data along with contextual data of torque, rotational speed, temperature, DL_1 and DL_2.

Two main type of autoencoders were explored; those employing *fully-connected* (FC) layers and those employing convolutional layers – convolution and max-pooling, often referred to as *Convolutional Neural Networks* (CNNs). In both cases the activation functions employed by hidden layers were ReLU and the output activation was linear, as typically used in regression problems. Exploration of modeling topologies also included global symmetric and asymmetric autoencoder structures. Exploration of modeling topologies also included global symmetric and asymmetric autoencoder structures. These two topologies were selected because they were found to be effective for anomaly detection (see e.g. (Eklund, 2018)). The autoencoders were then trained to *encode* the TSA vectors into progressively shorter vectors and to *decode* them back into a TSA vector estimate. It is important to note that these are not the only topologies of interest. For example, given the similarity between vibration data and speech, and because of their successes in speech applications, *Recurrent Neural Networks* (RNNs) - specifically *Long Short-Term Neural networks* (LSTMs) (Hochreiter & Schmidhuber, 1997) or *gradient recurrent units* (GRUs) (Cho et al., 2014) - are also good candidate topologies for gearbox analyses. Another type of neural networks of interest are transformers (Vaswani et al., 2017). However, experimentations with RNNs and transformers were not a part of the present study.

The performance metric employed was the *Mean-Squared Error* MSE, which was computed once per TSA acquisition, corresponding to 1 second of operation.

$$\begin{aligned} MSE &= ||\boldsymbol{x}_{TSA} - \hat{\boldsymbol{x}}_{TSA}||^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}\left(x_{TSA}[k] - \hat{x}_{TSA}[k]\right)^2, \end{aligned} \tag{2}$$

where $\hat{\boldsymbol{x}}_{TSA}$ is the autoencoder estimate and N=4,096 is the number of points in the TSA signal. Figure 6 shows a typical autoencoder output compared to the input.

Section 4.1 describes the analyses of these errors in regards to the ability to distinguish between baseline and propagation.

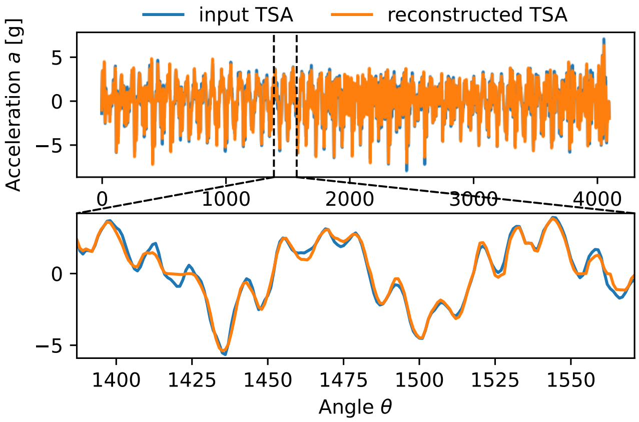

Figure 6. The figure shows reconstructed TSA data (orange) plotted over the input (blue). The bottom subplot shows a zoomed in view of the data.

4. EVALUATION

The performance of autoencoders as anomaly detection was evaluated and compared to the related performance of classical CIs. The evaluation of the reference classical CI performances, autoencoders, and their comparison are presented in the next three sections.

4.1. Condition Indicator Performance

Commonly used, classical, vibration-based CIs served as the reference for performance evaluation.

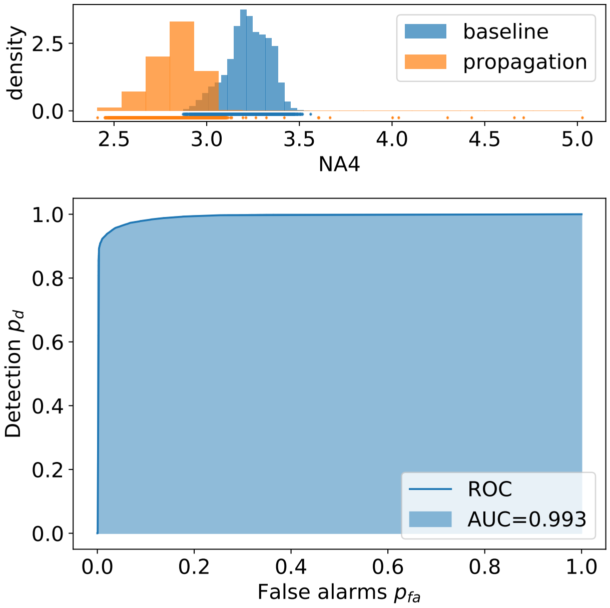

Figure 7. Example of AUC computation, where CI is NA4, evaluated on Gear 209.

Figure 7 shows a comparison of normalized histograms between data taken from the two hour baseline run (blue) and that from the subsequent crack propagation run (orange). In this specific example, the selected CI was NA4. An absence of overlap would indicate excellent anomaly detection capability. The bottom plot is the corresponding *Receiver Operating Characteristic* (ROC) curve (Fan, Upadhye, & Worster, 2006). The *Area Under the Curve* (AUC) of the (ROC) is used for the single valued performance metric, consistent with an earlier study (Nenadic et al., 2013). Broadly, ROC and AUC are popular for comparing classifiers in machine learning and pattern recognition. AUC was selected because anomaly detection process can be seen as a binary classifier that distinguished the healthy from a degraded state of a gear.

Figure 8 shows the same information for ten CIs, organized as columns, evaluated for eight baselines and eight associated crack-propagation tests, organized in rows. The AUC values are also given.

It is interesting to note that several CIs exhibited great performance on some but not all tests. For example, RMS distinguished propagation from baseline on Gear 211 and Gear 209, but not on the others; whereas kurtosis performed the best of all CIs on Gear 217, but had a considerable overlap for Gear 209.

4.2. Autoencoder Performance

The autoencoders were trained on baseline data only. During our analyses, we experimented with multiple topologies of fully-connected, and convolutional layers, and hybrids (networks that employ both fully-connected and convolutional layers) for autoencoders. The performances of these variations were very similar: they all seem to outperform classical CIs. The results of one hybrid topology of autoencoder-based anomaly detector are displayed in Figure 9. The encoder consisted of 7 fully connected layers with ReLU activations, followed by a single convolutional layer and max pooling. The output of the encoder was 32 features each of length 16. The decoder consisted of concatenating the features into a single 512 length feature vector, and passing it through a single linear layer to reconstruct the 4096 input points. The model was referred to as a asymmetric FC/CNN model with a linear decoder. An Adam optimizer was used with a learning rate of $\eta = 10^{-3}$. The model was trained to 500 training steps, but the best model was achieved around 150 training steps. No regularization was used in this computational experiment.

The figure shows a series of models with one model per column and the gear it was evaluated on in the rows. The histograms represent autoencoder *Mean-Squared Error* (MSE) associated with baseline and the same error associated with propagation.

Each model was trained on progressively more baseline sub-

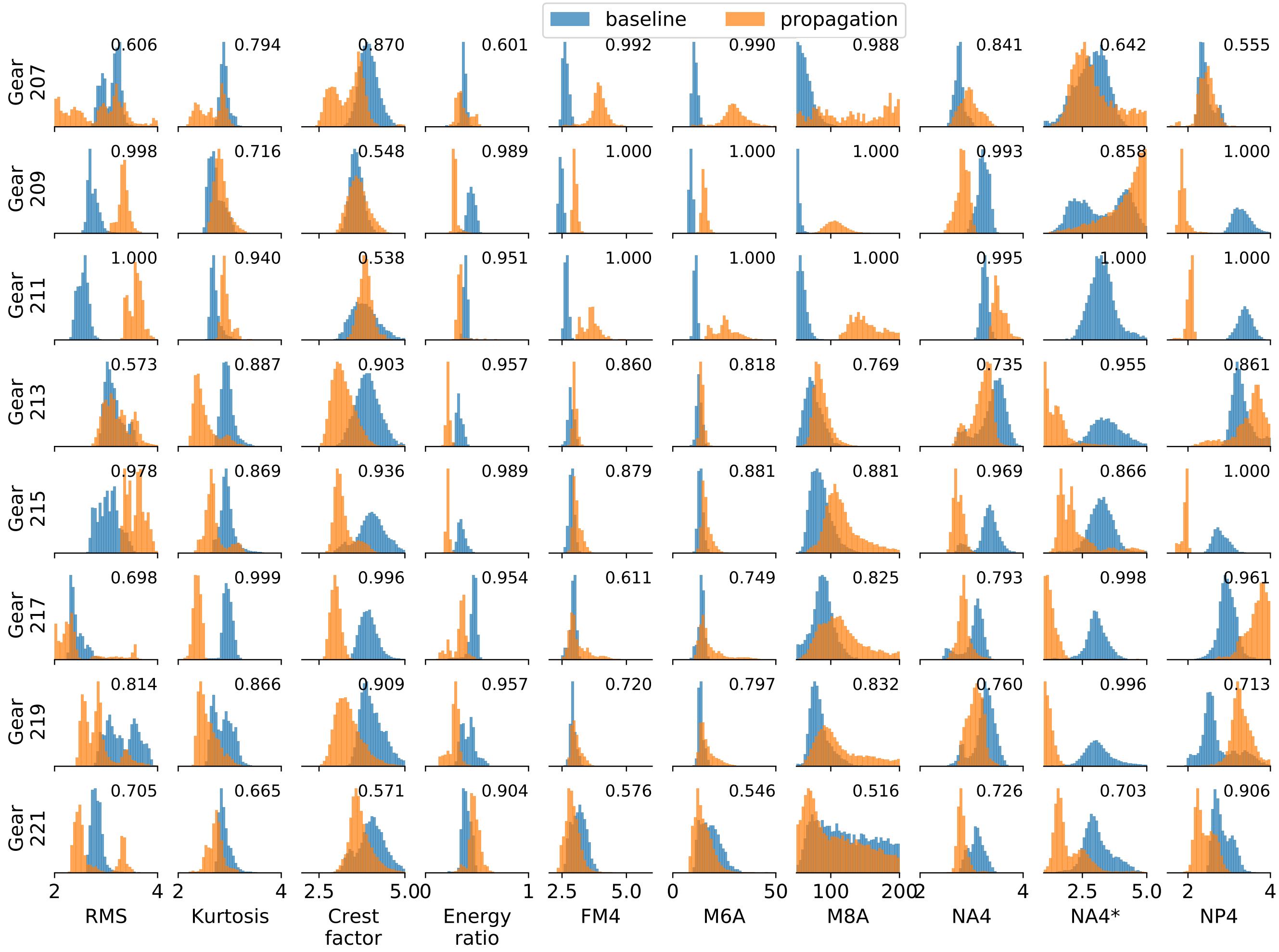

Figure 8. Performance of commonly used classical CIs as anomaly detectors on eight gears. The numbers in the subplots indicate AUC.

sets: the first model ”207” was trained on the subset of baseline data from the test with Gear 207, the second model ”207 + 209” was trained on data subsets coming from these two baselines, etc.

Going across from left to right for a given test gear, the general trend is that performance improves as more baseline cases from other test gears are used in the training. However, going down each column, performance decreases for the gears whose baseline data was excluded from the training. This would suggest that the models do not generalize well from one gear to the next. It is also observed that a model trained on an individual gear and evaluated on that same gear had a wider separation between error distributions associated with baseline and crack-propagation data, as shown in the top-left histogram (Model ”207” evaluated on Gear 207).

4.3. Classical CIs vs Autoencoder

A concise comparison of different classical condition indicators and two autoencoder-based models (one based on fully-connected layers, and the other on convolutional layers) is depicted in Figure 10. We plotted the metric AUC produced by two autoencoders and 11 CIs. Each symbol in the graph represents one gearbox experiment. The mean of the CIs $\langle AUC \rangle_{gear}$ over the gear experiments and the corresponding median $\mathrm{Median}(AUC)$ were also indicated. The autoencoder models are based on training that included all 8 baseline cases. They showed much better ability to distinguish between baselines and crack propagations.

5. MULTIPLE BASELINE TESTS

5.1. Anomaly Detection

To ensure that the autoencoder anomaly detector did not learn some spurious data associated with a specific run or gearbox

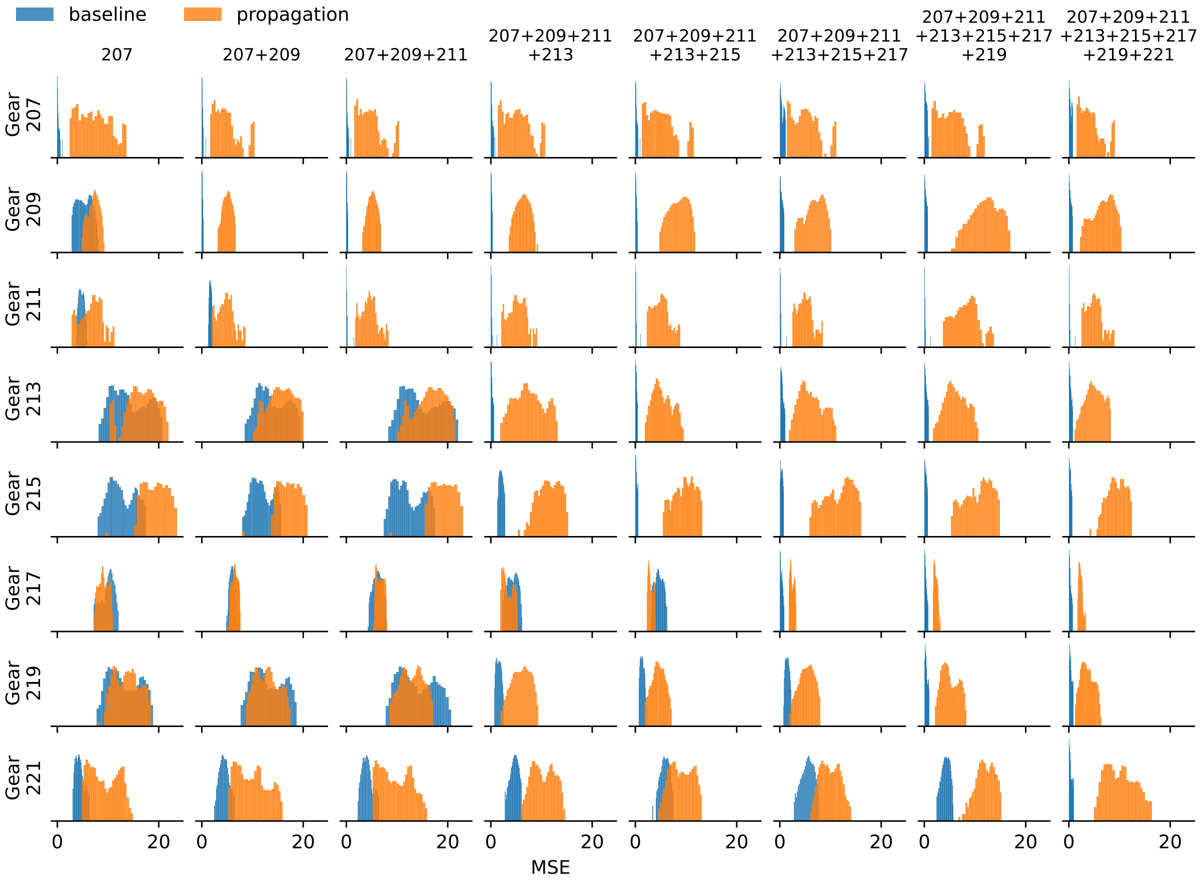

Figure 9. Autoencoders (one per column), trained on one or more baselines, as indicated on the top of columns, evaluated over eight gear baseline and crack-propagation tests. Log-scales were used in the y-axes to better show the distribution tails.

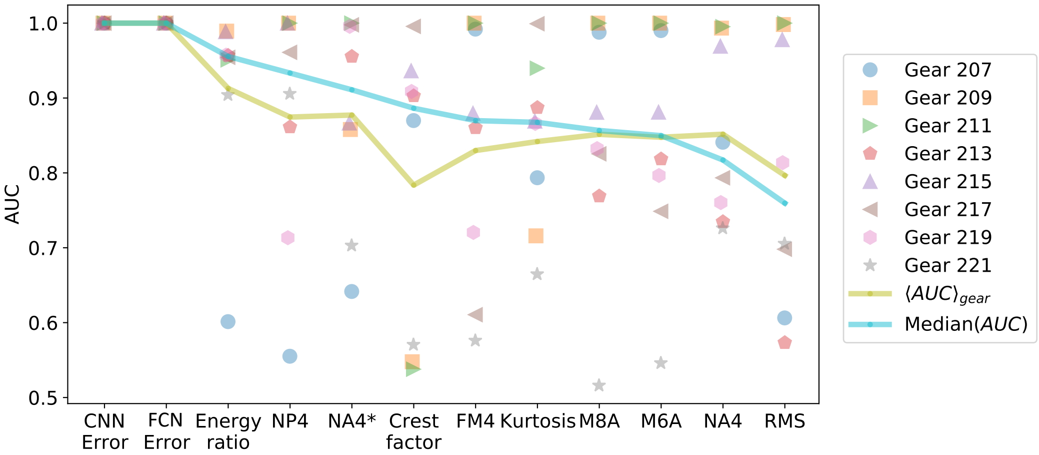

Figure 10. Comparison of AUCs across different models.

build, the modeling approach was further evaluated by designing a 9th test featuring multiple baseline runs of a new gear pairing.

Instead of one baseline run before a crack was seeded into one of the gears, a total of 8 baselines were run for the new gear pairing before propagation across multiple start/stop cycles and re-assemblies. These datasets were labelled $B1$ to $B8$. Baselines $B7$ and $B8$ involve partial gearbox disassembly for which the top gear was removed and reinstalled. The number of baselines was somewhat arbitrary (and the fact that it coincides with the number of crack propagation was just a coincidence): the objective was to collect data on a number of baselines, but at the same time to avoid unintended crack propagation (due to $\simeq$30% fatigue bending overload) before gear tooth is equipped with a crack-propagation sensor.

Several different model variations using both fully-connected and CNN layers were used. The models were trained on different subsets of baselines and many of those cases showed very good performance. One such performance is illustrated in Figure 11.

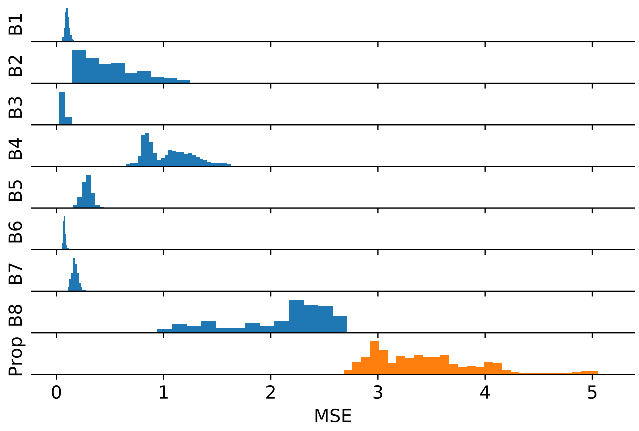

Figure 11. Distributions of MSE error for the eight baselines and the propagation.

Figure 11 depicts the autoencoder MSE errors associated with the eight baselines and single propagation datasets. This specific asymmetric autoencoder employed only five fully connected layers (associated neurons are 4096-256-64-16-1024-4096) and ReLU activation function for the hidden layers, was trained on baselines 1, 3, and 6, and evaluated on all baselines and the propagation. Note that the topology of the autoencoder is asymmetric: the encoding sub-network, defined by 4096-256-64-16, has three layers of weights, while the decoding sub-network, defined by 16-1024-4096 has two layer of weights. This topology was found through experimentation and was selected by its ability to create error that tracks damage level, as further described at the end of this section. The best performance was attained using Adam as the optimizer, zero dropout (although values up to 20% were experimented with), learning rate of $\eta = 10^{-4}$, and 100 epochs. The plots show that the error associated with baselines not involved in training is large than those that were used in training, but the propagation error is still larger.

Figure 12 shows the concatenated error distributions of all baselines B1-B8 in the same axes with the MSE distribution associated with the crack propagation. The AUC is very close to 1.

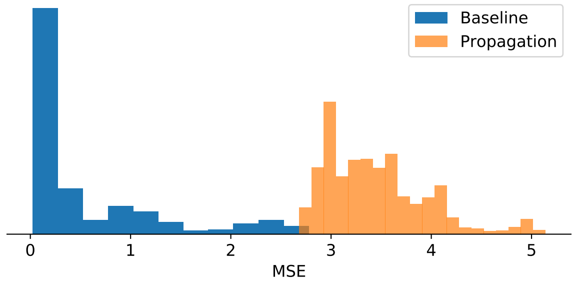

Figure 12. Distributions of the cumulative baseline and propagation with ROC.

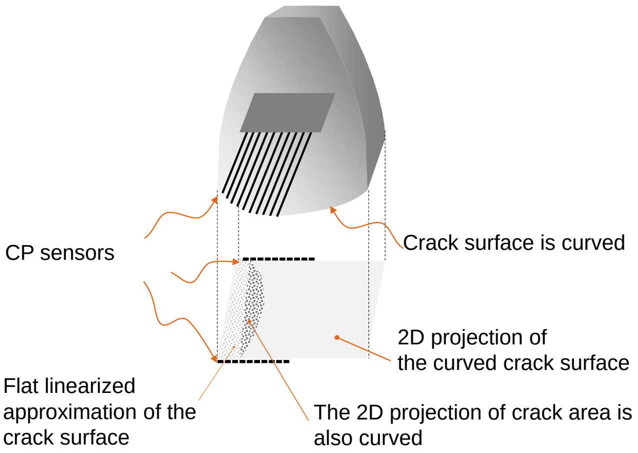

Figure 13. Distributions of the cumulative baseline and propagation with ROC.

5.2. Damage Assessment

Figure 13 shows one of the two CP sensors used to measure the surface crack length on each side of the gear near the root. These values were used to calculate metrics referred to as damage level 1 (DL_1) for one side and damage level 2 (DL_2) for the other. The damage level equals the total number of broken strands (Nenadic, Ardis, Hood, Thurston, & Lewicki, 2015). An example of a typical CP output was given in Figure 5 for which it took 54 minutes to break the 20 strands. The crack was relatively symmetric, as evidenced by similar progression as estimated by DL_1 and DL_2, with DL_2 being slightly delayed relative to DL_1.

We observed that some of the models trained on various baseline subsets (B1-B3-B5, B2-B4-B5, B1-B2-B3-B4-B5) showed not only good anomaly detection, but also damage assessment capability expressed by the surprisingly high Pearson correlation coefficient between the autoendoer's MSE during propagation and the estimated damage level. Figure 14 illustrates this correlation.

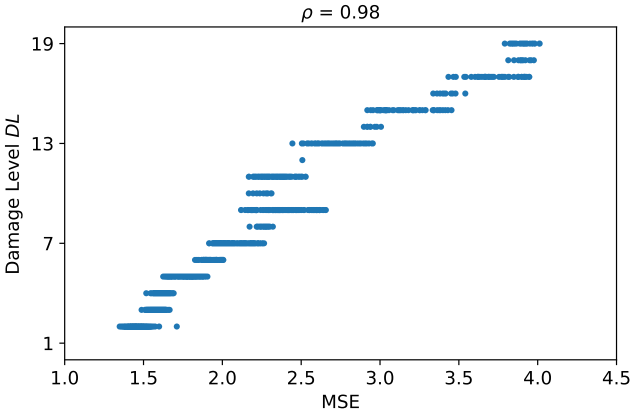

Figure 14. An example correlation between damage levels and Autoecoders MSE.

This behavior was not observed on any of the previous 8 tests using only one uninterrupted baseline case. However, when we trained autoencoders on only one of the eight baselines of the multi-baseline experiment, we had success in training an autoencoder with an MSE that highly correlated with DL ($\rho \geq$80%). Many autoencoders had MSE with greater than 80% correlation to DL when trained on two or more baselines.

For illustrative purposes, we fitted linear models of DL vs. MSE of the five model variations with highest Person correlation coefficients, as shown in Figure 15. The dots correspond to the DL vs. MSE scatters, the lines correspond to the associated linear fits, and the shaded area to range of the linear models. For example, the dashed line in the plot suggest that for MSE = 2, the five models approximately indicate the damage level in the range between 5 and 9, that is $DL \in [5, 9]|MSE = 2$. The purpose of the linear fits was not to propose damage level models, but just to show that the autoencoder error of some models track damage fairly linearly. It is important to emphasize that no information of fault and damage progression was used during training and only two parameters were used to fit the error to the damage level, slope and intercept. These results are preliminary: while at this time it is not clear what are the conditions that give rise to this *automatic* damage tracking, there seems to be sufficient evidence that indicate that these correlations are not spurious, or incidental. To be able to potentially use MSE as a damage estimator, the conditions that give rise to this phenomena must be clearly understood and assured.

6. CONCLUSION AND FUTURE WORK

Autoencoder-based, data-driven models showed improved and more consistent performance than the classical CIs for the first level of PHM capability, anomaly detection. To obtain more reliable performance and reduce Type II errors (false alarms), the autoencoder should be trained on multiple runs of the asset across all different operating and environmental conditions of interest.

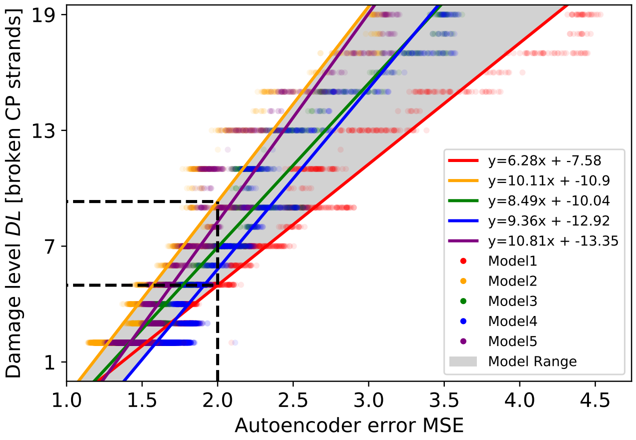

Figure 15. Five best linear models of damage level DL as a function of autoencoder's MSE. The shaded area signifies the uncertainty or ”disagreement” among the model variations.

High Pearson correlation coefficients between autoencoder's MSEs and estimated damage levels during crack propagation were observed across multiple models, suggesting that that a higher level PHM capability can sometime spontaneously arise.

After demonstrating the potential of autoencoder-based anomaly detectors the next step in the PHM development will be to examine their potential for the next level of capability, damage assessment. In addition to understanding the conditions that give rise to spontaneous damage assessment, the pre-trained autoencoders will be fine-tuned, using *transfer learning* to learn damage level and crack-propagation sensors as the ground truth of damage progression. In addition, alternative models using RNNs (LSTMs or GRUs) or transformers will be explored for predicting damage, by using a subset of damage progression for training and the rest for validation. The team is also working to prepare the dataset to be shared with the gear research community.

ACKNOWLEDGMENT

This material is based upon work sponsored by the Office of Naval Research under Award No. N00014-18-1-2339.

DISCLAIMER

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

Biographies

Adrian Hood received his B.S. and M.S. in Mechanical Engineering and PhD in Aerospace Engineering from the University of Maryland. He is currently a Research Mechanical Engineer at the US Army Research Lab in Aberdeen Proving Ground, MD. working in the Power and Propulsion Branch of the Weapons and Materials Research Directorate. His research is focused on Health and Usage Monitoring of Helicopter Transmissions, gear dynamics, tribology, and machine learning.

Christopher J. Valant received his B.S. in Physics and M.S. in Data Science from Rochester Institute of Technology (RIT). He joined the Golisano Institute of Sustainability (GIS) in 2015. His research interests include applied machine learning, in particular, artificial intelligence with computer vision and decision theory, as well as evolutionary algorithms, optimization, and programming.

Patrick R. Horney received his B.S. in Aerospace Engineering from Purdue University in 2008. Since 2012 he has worked for NAVAIR in the areas of Diagnostics, Prognostics and Condidition Based Maintenance (CBM). He has established CBM/CBM+ capability for the US Navy on two aviation plaforms and serves as an advisor in the areas of PHM and CBM+.

Allen Jones received his B.S in Mechanical Engineering from the University of Maryland and his M.S. in Engineering Management from the George Washington University. He is a mechanical engineer at the Naval Air Warfare Center Aviation Division in Patuxent River, MD, working in the Air Systems Group, Propulsion and Power Engineering, Drives and Mechanical Systems Branch. His research interests include machine learning and advanced algorithms for improved diagnostics and prognostics using Health and Usage Monitoring System data.

Jared S. Lantner is currently working towards a B.S. in Computer Science from Rochester Institute of Technology (RIT). He has been working at the Golisano Institute for Sustainability (GIS) since January 2021. His interests include machine learning, algorithms, and the development of software.

Josiah Martuscello is a graduate from Rochester Institute of Technology with a BS in Mechanical Engineering. He intends to complete more applied research with in neuroscience and machine learning.

Nenad G. Nenadic received his B.S. in Electrical Engineering from University of Novi Sad (Novi Sad, Serbia) in 1996 and his MS and Ph.D. in Electrical and Computer Engineering from University of Rochester (Rochester, NY, USA) in 1998 and 2001, respectively. He joined Kionix Inc. in 2001, where he worked on development of microelectromechanical inertial sensors. Since 2005, he has been with Center for Integrated Manufacturing Studies (CIMS) at Rochester Institute of Technology, where he is currently a Research Associate Professor. His research interest include design, analysis, and monitoring of electromechanical devices and systems. He has two patents in electromechanical design. He co-authored a textbook *Electromechanics and MEMS* and is a member of IEEE.

References

Bechhoefer, E., & Butterworth, B. (2019). A comprehensive analysis of the performance of gear fault detection algorithms. *Annual Conference of the PHM Society, 11-1*.

Chaari, F., Fakhfakh, T., & Haddar, M. (2009). Analytical modelling of spur gear tooth crack and influence on gearmesh stiffness. *European Journal of Mechanics - A/Solids*, *28*(3), 461-468. doi: https://doi.org/10.1016/j.euromechsol.2008.07.007

Chen, Z., & Shao, Y. (2011). Dynamic simulation of spur gear with tooth root crack propagating along tooth width and crack depth. *Engineering Failure Analysis*, *18*(8), 2149-2164. doi: https://doi.org/10.1016/j.engfailanal.2011.07.006

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cooley, C., Hood, A., & Wang, Y. (2021). Tooth mesh modeling of spur gears with tooth root crack damage using a finite element/contact mechanics approach. *SAE Technical Paper 2021-01-0699*, *48*, 185-200. Retrieved from `https://doi.org/10.4271/2021-01-0699` doi: https://doi.org/10.1016/j.engfailanal.2014.11.015

Eklund, N. (2018). Anomaly detection tutorial. In *Proceedings of the annual conference of the prognostics and health management society.*

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, *8*(1), 19–20.

G. Jinks, M. B., J. Langhout. (2016). *ADS-79E-HDBK condition based maintenance system for us army aircraft*. US Army Research, Development and Engineering Command, Aviation and Missile Research, Development and Engineering Center, Huntsville, AL.

Goebel, K., Daigle, M. J., Saxena, A., Roychoudhury, I., Sankararaman, S., & Celaya, J. R. (2017). *Prognostics: The science of making predictions*. CreateSpace Independent Publishing Platform.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Japkowicz, N., Myers, C., & Gluck, M. (1995). A novelty detection approach to classification. In *Ijcai* (Vol. 1, pp. 518–523).

Lebold, M., McClintic, K., Campbell, R., Byington, C., & Maynard, K. (2000). Review of vibration analysis methods for gearbox diagnostics and prognostics. In *Proceedings of the 54th meeting of the society for machinery failure prevention technology* (Vol. 634, p. 16).

Liang, X., Zuo, M. J., & Hoseini, M. R. (2015). Vibration signal modeling of a planetary gear set for tooth crack detection. *Engineering Failure Analysis*, *48*, 185-200. doi: https://doi.org/10.1016/j.engfailanal.2014.11.015

Martin, H. (1989). Statistical moment analysis as a means of surface damage detection. In *Proceedings of the 7th international modal analysis conference, society for experimental mechanics, schenectady, ny.*

NASA. (1994). Specifications for test gears, specification no 3-548642.

Nenadic, N. G., Ardis, P. A., Hood, A. A., Thurston, M. G., Ghoshal, A., & Lewicki, D. G. (2013). Comparative study of vibration condition indicators for detecting cracks in spur gears. In *presented at the american helicopter society 69th annual forum and technology display, phoenix, az.*

Nenadic, N. G., Ardis, P. A., Hood, A. A., Thurston, M. G., & Lewicki, D. G. (2015). Processing and interpretation of crack-propagation sensors. In M. J. Daigle & B. Anibal (Eds.), *Proc of annual conference of the prognostics and health management society* (p. 560-568). PHM Society.

Nenadic, N. G., Wodenscheck, J. A., Thurston, M. G., & Lewicki, D. G. (2011). Seeding cracks using a fatigue tester for accelerated gear tooth breaking. In T. Proulx (Ed.), *Rotating machinery, structural health monitoring, shock and vibration* (Vol. 8, pp. 349–357). Springer.

Polyshchuk, V., Choy, F., & Braun, M. (2002). Gear fault detection with time-frequency based parameter np4. *International Journal of Rotating Machinery*, *8*(1), 57–70.

Samuel, P. D., & Pines, D. J. (2005). A review of vibration-based techniques for helicopter transmission diagnostics. *Journal of sound and vibration*, *282*(1), 475–508.

Sharma, V., & Parey, A. (2016a). Gear crack detection using modified tsa and proposed fault indicators for fluctuating speed conditions. *Measurement*, *90*, 560-575. doi: https://doi.org/10.1016/j.measurement.2016.04.076

Sharma, V., & Parey, A. (2016b). A review of gear fault diagnosis using various condition indicators. *Procedia Engineering*, *144*, 253–263.

Sikorska, J. Z., Hodkiewicz, M., & Ma, L. (2011). Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, *25*(5), 1803-1836.

Stewart, R. (1977). Some useful data analysis techniques for gearbox diagnostics. In *Machine health monitoring group, institute of sound and vibration research, university of southampton, report mhm/r/10/77.*

Swansson, N. (1980). Application of vibration signal analysis techniques to signal monitoring. In *Conference on friction and wear in engineering 1980. institution of engineers, australia, barton, australia.*

Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., & Wu, B. (2006). Intelligent fault diagnosis and prognosis for engineering systems. In *1st ed. hoboken.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Yan, W., & Yu, L. (2015). On accurate and reliable anomaly detection for gas turbine combustors: A deep learning approach. In *Proceedings of the annual conference of the prognostics and health management society.*

Zakrajsek, J. J., Townsend, D. P., & Decker, H. J. (1993). *An analysis of gear fault detection methods as applied to pitting fatigue failure data* (Tech. Rep.). NATIONAL AERONAUTICS AND SPACE ADMINISTRATION CLEVELAND OH LEWIS RESEARCH CENTER.