

Improved acoustic data generation and classification for fault detection in prognostics and health management

Ruth A. Willet ¹, Karl M. Reichard ²

^{1,2} *Graduate Program in Acoustics, College of Engineering,
The Pennsylvania State University, University Park, PA, 16802, USA*
raw5930@psu.edu
kmr5@psu.edu

ABSTRACT

This research is investigating how acoustic knowledge can be used to improve data classification and generation for acoustic and ultrasonic data. Supervised machine learning methods are first used to study feature selection for a water pipeline acoustic dataset via overall classification accuracy. While there is sufficient fault and no-fault data for binary fault diagnosis, some subclasses of faults have few representative samples and insufficient data. Generative methods are being investigated to supplement existing data for specific fault types based on statistical and acoustic features. This work is also investigating signal generation for weld crack detection and localization in order to extend the use of generative methods to multiple sensors. By understanding the dispersive signals that could be seen at several sensors, the effect of data generation for a broader range of acoustic data can be studied. Incorporating acoustic knowledge in model creation and training to generate time series data can hopefully provide better insights on what generative methods are most useful for system diagnostics. This work aims to quantify results using a variety of methods, including classification accuracy from the supervised machine learning methods and direct comparison of generated data and actual data for usability in actual systems.

1. PROBLEM STATEMENT

Acoustics, or the study of the generation, transmission, and reception of sound and vibration, is a field of study that touches all aspects of modern engineering. Acoustic engineers could be interested in information from speech and voice production, harmonic music synthesis, broadband ocean or water noise measurements, tonal machinery data, propagated sound from sound-structures interaction, or non-linear elastic wave propagation in materials to name only a

few. These signals are measured using accelerometers, microphones, hydrophones, or acoustic emission sensors depending on the frequency range of interest, specific type of signals that will be generated by a system, and physical space available to mount a sensor.

Because of the variety of acoustic data that can be collected and analyzed, acoustics is incredibly useful for prognostics and health management (PHM) problems in monitoring and diagnosing systems (Prajapati, Bechtel, & Ganesan, 2012; Ambhore, Kamble, Chinchani, & Wayal, 2015). This work will be focusing on acoustic data specifically, though other sensors like strain gauges, temperature gauges, and pressure sensors can be used to confirm what is being measured acoustically. Regardless of data domain, however, one of the biggest problems for prognostics and health management is the availability of data. If a mechanical system, for example, is instrumented with sensors to record vibrations at set intervals which will be classified as either "healthy" or "faulty", the likelihood that more of the measured data will be healthy is much higher than the data being faulty. After all, if a system has components that are not working properly, they need to be repaired for safety and overall system usefulness. As a result, many PHM datasets have an uneven distribution of data, with much more data being classified as healthy than faulty. In addition, it can be challenging to have a large amount of data in total if the time needed to gather data on a system conflicts with its daily operations. Therefore, the problem narrows down further to what to do with acoustic data that also has an uneven data distribution.

Signal processing is an important part of the PHM process to make data classification possible in the first place. Machine learning methods have become very common to complete an initial data classification, and this is especially true in acoustics (Bianco et al., 2019). These machine learning methods could be supervised or unsupervised methods, but some of the most common common supervised models that appear in PHM literature include linear discriminant analysis

Ruth Willet 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.

(Ramashini, Abas, Grafe, & De Silva, 2019), support vector machines (Papandrea, Frigieri, Maia, Oliveira, & Paiva, 2020), decision trees (Navada, Ansari, Patil, & Sonkamble, 2011), K-nearest neighbors (Gil, Grochowina, & Leniowska, 2019), and Naive Bayes (Jafari & Dorafshan, 2021).

One method of improving uneven data or increasing the overall amount of data is data augmentation. Classically, this has incorporated methods of oversampling or interpolation, such as bootstrap sampling, synthetic minority oversampling technique, or simply adding noise to existing data (Liu, Zayed, & Xiao, 2024). Generating data from a physics-based model, like finite-element analysis is also very effective. However, generative AI is now a useful tool to do data augmentation with methods like Generative Artificial Networks (Goodfellow et al., 2014), Large Language Models (Naveed et al., 2025), and Diffusion (Zhao, Xu, Song, Shen, & Ma, 2023). Generative methods such as GANs originally focused on creating data based off of input images. This has been transferred to acoustic applications by using spectrograms with time-frequency information as image-like inputs (Yin, Wang, Xu, Yang, & Zhu, 2025). Since acoustic data is commonly visualized as feature distributions or frequency domain energy spectra, this maps well to classification and data generation. However, only analyzing data in the frequency or feature domains neglects important time domain information such as phase. Researchers have now shown that generative methods can be successfully applied to time series data in several fields (Brophy, Wang, She, & Ward, 2023). By using generative methods as opposed to finite element methods, both the signal and its underlying noise could be learned by the algorithms for more realistic data.

To refine the core issue, the gap to be addressed is using acoustics-specific knowledge in tandem with machine learning methods to improve synthetic acoustic data time series generation and fault diagnostics. Acoustic signals have diverse characteristics, such as short time-duration signals, harmonic signals, continuous broadband noise, narrow-band signals, and impulsive signals. Each one requires some acoustic knowledge to extract consistent features that can be used for system diagnostics. By picking analysis methods that align with specific signal types, understanding the acoustics of a system for fault detection and localization, and comparing generated time series signals to other measured signals, acoustics-specific knowledge can be used to improve data generation and classification for fault detection.

Two acoustic datasets that provide opportunities to tackle this problem are being used in this work. First, a water utility pipeline acoustic data is currently being studied because it provides a fairly balanced dataset of labeled leak and no-leak data with a wide variety of frequency information. While splitting the data with a binary leak or no-leak classification is fairly balanced, subsets of data looking at specific leak types

are not balanced when compared to the overall amount of healthy data. This provides opportunities to train classifiers on a variety of subsets of uncorrelated data, improve feature selection, and generate data for specific leak types. To further investigate time series information in data generation with multiple sensors, an acoustic emissions dataset of weld crack detection and localization will be created for this work. Time domain information is incredibly important due to the dispersive nature of the signal, so it may provide additional insights that can't be seen with continuous water pipeline data. Unlike acoustic and vibration data, the labeled class of data is also dependent on the type of material which adds more variability to the analysis.

2. NOVEL CONTRIBUTIONS

The key aim of this work is to contribute to time series data generation for diagnostics and prognostics. Literature has mostly focused on using acoustic signals in the frequency and time-frequency domains for diagnostics, so pushing the understanding of how to incorporate time domain signals is extremely relevant. This work aims to apply generative artificial intelligence models such as generative adversarial networks or diffusion models to a wide spectrum of acoustic signals to train, test, and validate diagnostic algorithms. This work will be focusing on combining acoustic informed features with advancing generative methods to generate data and quantify it using supervised classification algorithms.

This work also aims to improve weld crack localization in materials. Given the nonlinear behavior of acoustic emissions signals that depend on many material properties, there are gaps to be found in feature extraction, specific material and weld type combinations, and time series data generation. By incorporating generative models that could learn key characteristics of available signals, algorithms for system prognostics can be improved.

3. RESEARCH PLAN

There are several stages of research that are required to fully investigate this research problem. First, fault classification and generative algorithms are being made from the Thames Water utility dataset. This dataset has over 37,000, ten second acoustic files that are labeled as no-leak (no-fault) or leak (faulty). When technicians were sent into the field to investigate potential pipe leaks, they only identified the sensor's location when an actual leak was identified. If the acoustic monitoring system had pinged an issue at a boundary box but there was no leak, the data was only classified as no leak. When leaks were actually found, technicians noted where the leak was occurring in over half of the data. This means that a large percentage of the data files were classified as leak or no-leak, but crucial information on what parts were being monitored were not noted.

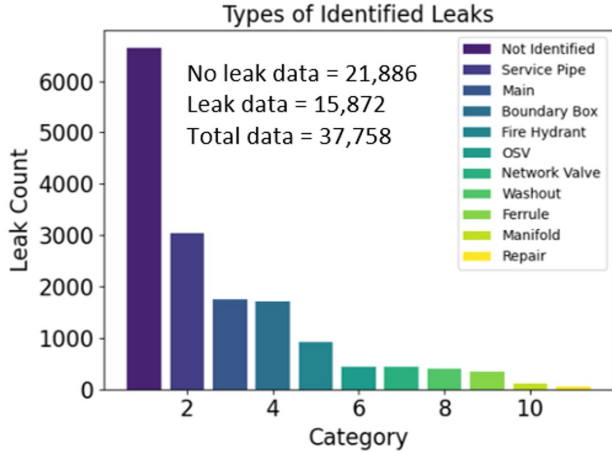


Figure 1. The distribution of leak types for the Thames Water dataset. Since most types of labeled leaks have a small number of files, data generation is needed to increase synthetic data and improve classifier accuracy.

Work to date has focused on using statistical and acoustic data features to classify the water pipeline data as fault or no-fault data using machine learning models. These classifiers have included K-nearest neighbors, Linear Discriminant Analysis, Support Vector Machines, Decision Trees, Random Forest, Naive Bayes, and Convolution Neural Networks. There has not been solid improvement in classification with any specific features, but work is ongoing to understand feature pruning that could indicate what physics-informed features are most helpful before these classifiers are used as one metric for assessing and comparing the generated data.

The biggest challenge of this data comes from the distribution of types of leaks as seen in Figure 1. While the number of files for a binary leak vs no-leak analysis is fairly balanced, that balance drops off as soon as specific leak types, such as service pipes, main waterlines, or boundary boxes, are extracted. Current work aims to extract key features for specific leak types and implement generative models like LSTM-GANs have been shown to work well at generating data (Liu et al., 2024). The goal of this aspect of the research is to verify that convincing acoustic data can be generated for this type of data, equalize the number of files for each type of leak, and improve fault classification.

Once the water pipeline data has been generated, it needs to be assessed relative to real data. One way to validate and test the generated data is to examine the distribution of features for leak types when compared to real data. Figure 2 shows a representative scatter plot of two features for three categories of data, no-leak, one leak type, and any type of leak, to search for feature clusters in the measured data. The blue and red subplots were randomly selected from data that was labeled as either no-leak (blue) or leak (red). The same

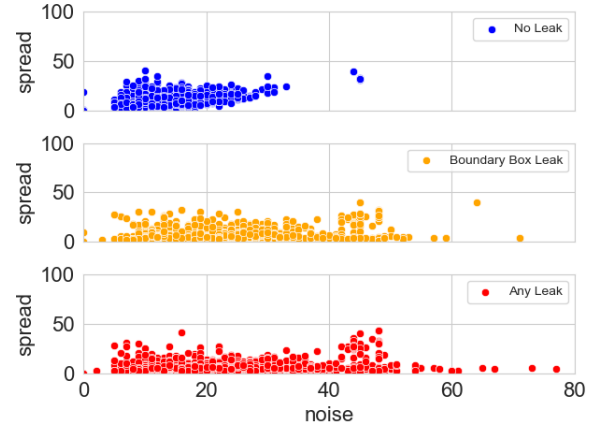


Figure 2. Examining the data scatterplot of noise vs spread for randomly selected no-leak data, all cases of boundary box leaks, and any type of leak data does not indicate clear clustering with these specific features.

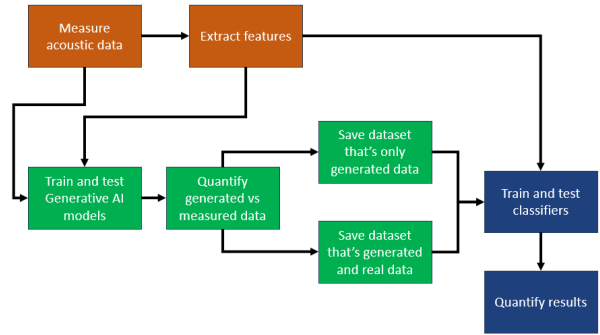


Figure 3. Research plan for implementing generative data into a PHM process. The three color schemes indicate pre-processing work in orange, data generation in green, and data diagnostics and validation in blue.

number of files as the amount of boundary box leaks, in yellow, were chosen to be able to compare them. By plotting them directly next to each other, the hope was that a subset of features with clear clustering for a specific leak type could be found. Preliminary analysis is not showing clear clustering of features, but features will still be examined with scatter plots and other dimensionality reduction techniques for validation and classification. Figure 3 shows the intended workflow for both datasets in this work.

Looking more broadly, a similar process will be taken with the weld-crack detection and data generation over the next two years. A test bed is being designed and built to enable laboratory acoustic emission data collection. Additional data will be collected from welded fluid tanks. Sensor localization is a well-studied use of acoustic emissions, so research will start there to verify that received signals are being analyzed correctly for the specific material(s) and weld types. Once there is confidence in the test bed data collection and analysis,

new generative models will be trained on this dataset to augment the detected cracks data. The supervised machine learning classifiers that are already built will be re-trained with a mix of actual data and generated data to validate and test the effect of adding synthetic data to classification algorithms.

Studying several datasets will hopefully allow for direct comparisons between types of acoustic signals. By studying the results for different time series data generation, feature selection, and classifier accuracy, there is hope that the role of acoustic data generation for accurate diagnostics and prognostics can be further quantified.

ACKNOWLEDGMENT

This work was partially supported by the PIPELINE (Penn State Intern Pipeline Links to Navy Engineering) program, ONR grant #N000142312656. The Penn State PIPELINE Program motivates and connects students and faculty to careers and research opportunities with the Navy technical workforce.

REFERENCES

- Ambhore, N., Kamble, D., Chinchani, S., & Wayal, V. (2015). Tool condition monitoring system: A review. *Materials Today: Proceedings*, 2(4-5), 3419–3428.
- Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., & Deledalle, C.-A. (2019, November). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5), 3590–3628.
- Brophy, E., Wang, Z., She, Q., & Ward, T. (2023, February). Generative adversarial networks in time series: A systematic literature review. *ACM Comput. Surv.*, 55(10). 10.1145/3559540
- Gil, D., Grochowina, M., & Leniowska, L. (2019). Detecting of the rolling bearing state based on acoustic signal and the k-nn classifier. In *2019 signal processing: Algorithms, architectures, arrangements, and applications (spa)* (pp. 246–249).
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Jafari, F., & Dorafshan, S. (2021). Bridge inspection and defect recognition with using impact echo data, probability, and naive bayes classifiers. *Infrastructures*, 6(9), 132.
- Liu, R., Zayed, T., & Xiao, R. (2024). Advanced acoustic leak detection in water distribution networks using integrated generative model. *Water Research*, 254, 121434.
- Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. In *2011 IEEE Control and System Graduate Research Colloquium* (p. 37-42). 10.1109/ICSGRC.2011.5991826
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... Mian, A. (2025, August). A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5). 10.1145/3744746
- Papandrea, P. J., Frigieri, E. P., Maia, P. R., Oliveira, L. G., & Paiva, A. P. (2020). Surface roughness diagnosis in hard turning using acoustic signals and support vector machine: A pca-based approach. *Applied Acoustics*, 159, 107102.
- Prajapati, A., Bechtel, J., & Ganesan, S. (2012). Condition based maintenance: a survey. *Journal of Quality in Maintenance Engineering*, 18(4), 384–400.
- Ramashini, M., Abas, P. E., Grafe, U., & De Silva, L. C. (2019). Bird sounds classification using linear discriminant analysis. In *2019 4th international conference and workshops on recent advances and innovations in engineering (icraie)* (pp. 1–6).
- Yin, J., Wang, Y., Xu, Z., Yang, Y., & Zhu, M. (2025). Masgan: a malfunction sound detection model for industrial machines using the generative adversarial networks. *Expert Systems with Applications*, 293, 128602.
- Zhao, X., Xu, F., Song, D., Shen, J., & Ma, T. (2023). A novel blade crack detection method based on diffusion model with acoustic-vibration fusion. In *2023 IEEE 21st international conference on industrial informatics (indin)* (pp. 1–7).