Datasets for data science investigation of fused EM/RF and vibroacoustic equipment monitoring

Tom Grimes
PNNL

Lynn Wood
PNNL

Karl Pitts
PNNL

Nathaniel Smith
PNNL

Jihee Yang
PNNL

Eva Brayfindley
PNNL

Elisabeth Moore
PNNL

Jan Irvahn
PNNL

Jeff Miller
PNNL

Jack Dermigny
PNNL

ABSTRACT

We present a dataset for enabling the use of deep learning for understanding the state of operating machinery. This dataset includes voltage and vibroacoustic measurements taken in a phase-locked fashion on a variety of common office and lab equipment. All included equipment connects to the wall via standard electrical cords with a strong emphasis on small hand tools. The distribution package includes the raw voltage and vibroacoustic data, the metadata, a Jupyter notebook that trains and evaluates a baseline deep learning model for performing useful sample tasks (e.g., building a classifier to determine if a single piece of equipment is on or off), a modifiable PyTorch dataloader, and a file to build an appropriate python environment. Using these tools, the next generation of data science practitioners can work toward newer and better approaches specifically for analyzing and understanding signals from operating machinery.

INTRODUCTION

Recent advances in artificial intelligence are truly impressive, especially for text and image data modalities. These advances are due in part to publicly available open-source text and image datasets [1]. Harnessing the full power of artificial intelligence, machine learning, and deep learning for more accurate, useful, and nuanced safeguards capabilities will require carefully curated datasets for technique development. Unfortunately, to date publicly available datasets specific to safeguards have been severely lacking or entirely absent. Here we present a new dataset constructed to help fill this gap.

We construct and distribute this dataset with two goals in mind. First, the overall meta-goal is to generate a methodology that produces datasets for new modalities relevant to the non-proliferation mission that are optimized for usefulness with data science techniques. To make progress toward this meta-goal, the more proximate goal is to develop and demonstrate this methodology by generating, procuring, compiling, cleaning, modeling, and distributing datasets incorporating carefully chosen modalities of interest to safeguards. Once assembled, this dataset needs to be useful for algorithmic development, algorithmic comparison, and transfer learning. We take these requirements into consideration as well as ensuring that the final dataset will be relevant to safeguards when choosing data collection schemes and modalities.

Algorithmic Development: Designing a dataset to be relevant for machine learning model development places a heavy burden on the selection of data. Traditional methods for building a dataset start by identifying a relevant problem and then collecting the data that would be required to solve that problem. Generally, doing so has one of two outcomes: (1) the chosen problem is unexpectedly extremely easy such that all classifiers correctly identify the solution, or (2) the chosen problem is unexpectedly extremely difficult such that no classifiers correctly identify the solution. These scenarios are depicted in Figures 1a and 1b. The left-hand column shows datasets arranged by difficulty of separating the classes from one another (in this case using signal-to-noise ratio to denote difficulty). The right-hand column shows models arranged by capacity and capabilities. By starting with a single dataset and taking all the data at that difficulty, the results are suboptimal in both cases. In the 'easy' case, a very simple model can separate the data and more advanced models that would have led to improved capabilities are not leveraged against the problem because there is no motivation to do so or means to discriminate between the model selected and more refined models. In the 'hard' case, the data is so difficult to model that it requires multiple leaps beyond the state of the art. Because all classifiers perform very poorly, it is extremely difficult to make progress toward a performant classifier. It is likely that performant classifiers will not be found and therefore there will be no gains in capabilities.

There is, however, another possible paradigm for taking data that doesn't suffer from these difficulties. Naïve data taking started with a selected problem and the simplest possible approach because the problem itself was interesting and proceeded to work upwards in model complexity. Disciplined, capability-maximizing data taking instead starts with a model selected because it is interesting, well-benchmarked, matched to the hardware, and matched to the application and a very simple version of the problem to take data against. As the model shows capacity to extract meaningful features and do meaningful classification the difficulty of the problem is increased until the model can no longer succeed. At this point, there is a phase of alternation between increasing the complexity of the model and increasing the difficulty of the dataset until no model can succeed. Thus, the dataset produced at a slightly higher difficulty than can be solved by the most advanced model provides the most interesting problem to develop new models against. In addition, all the datasets produced along the way are extremely useful for testing new models or for transfer learning to the difficult problem. Disciplined, capability-maximizing data taking is depicted in Figure 2. Put succinctly: You don't know what data is most valuable to take until you start analyzing it!

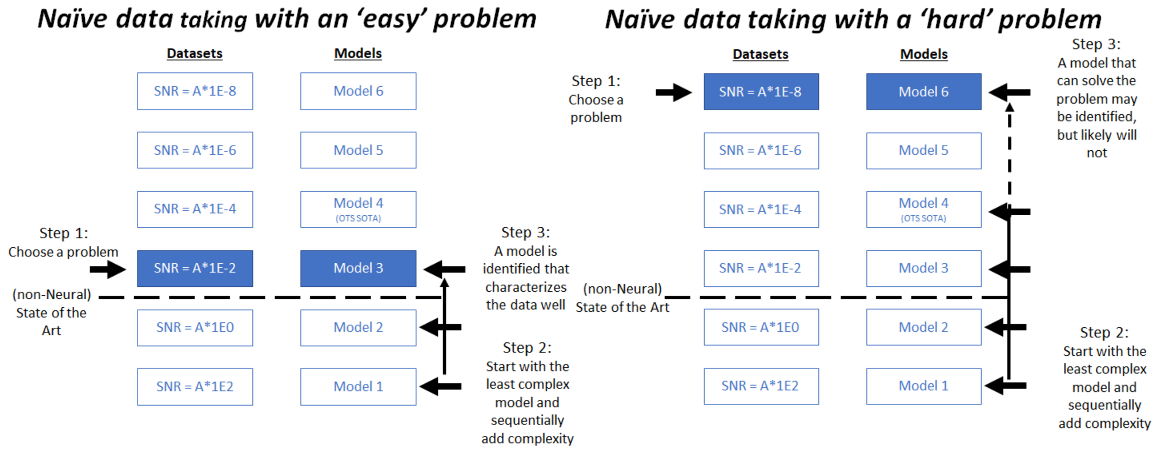

Figure 1 a,b. Naïve data taking paradigm applied to (a) a problem that is too easy (too rich in signal) and (b) too hard (too little signal)

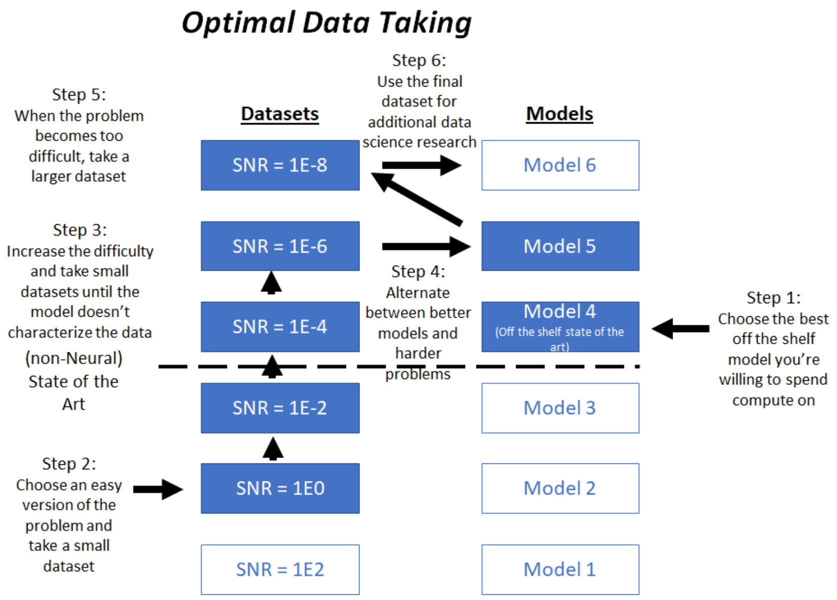
Optimal Data Taking
Step 6:
Use the final dataset for additional data science research
Datasets
Models
Step 5:
When the problem becomes too difficult, take a larger dataset
SNR = 1E-8
Model 6
Step 3:
Increase the difficulty and take small datasets until the model doesn't characterize the data
SNR = 1E-6
Model 5
Step 4:
Alternate between better models and harder problems
SNR = 1E-4
Model 4
(Off the shelf state of the art)
Step 1:
Choose the best off the shelf model you're willing to spend compute on
(non-Neural)
State of the Art
SNR = 1E-2
Model 3
Step 2:
Choose an easy version of the problem and take a small dataset
SNR = 1E0
Model 2
SNR = 1E2
Model 1

Figure 2. Optimal data taking paradigm where a series of data collects are taken and validated by a model until eventually dataset complexity increases drive model development

Algorithmic Comparison: A high-quality shared dataset is a powerful instrument for community building. By virtue of being publicly available, high-quality, and of an appropriate difficulty, the dataset draws together the community of researchers to coordinate around a common problem. Well-known examples of this behavior include the widespread usage of the MNIST [2] handwriting dataset and the CIFAR [3] and ImageNet [4] image datasets. One of the major advantages to working on such a dataset is that the data cleaning and data engineering tasks (often regarded as most of the work on data science problems) can be performed before the data is distributed and thus eliminated for the researchers. Furthermore, because all researchers are working with data in the same structure and format, code-sharing and collaboration becomes significantly easier. This also significantly reduces barriers to reproducibility (a major issue with data science in many hard science fields). Finally, having a common task and shared evaluation metric makes it possible to compare approaches in an apples-to-apples objective fashion. This makes it much simpler to identify promising lines of inquiry, both for individual researchers and funding organizations.

Transfer Learning: One of the most powerful results in the machine learning field has been that pre-training on large, diverse datasets yields a meaningful and sometimes tremendous improvement for models trained on smaller datasets for a related, but usually more bespoke task. If shared datasets are large, rich in features, and well aligned with interesting problems, we believe it will be possible to use them for pretraining to obtain better classification accuracy on the interesting problems. In the safeguards space there are often limitations to how similar a distributed surrogate dataset can be to interesting problems and thus, for many modalities transfer learning will not be feasible; however, for many modalities, it is possible to produce data on a surrogate problem not at all related to controlled information but nonetheless spanning a wide enough set of features that pretraining using the data is useful.

BACKGROUND

The modality selected for initial investigation was a fused dataset consisting of EM/RF collected by measuring voltage and vibroacoustic data collected with a laser vibrometer. These are complimentary modalities because both the acoustic waves and the EM/RF signatures originate from common sources in the circuitry. Thus, in collecting the modalities together the user is provided with two individual data sets for the modalities which are useful on their own in addition to the fused dataset. Because there is mutual information in the data streams, clever algorithms should be able to use them together to achieve accuracy and robustness that a single modality alone would be unable to achieve. Thus, the community is incentivized to create and compare algorithms for data fusion which may be applicable to a wide array of other useful tasks.

There are many different safeguards motivations for wanting to monitor the state of equipment. Preventative maintenance applications make it possible to identify equipment operating outside of a normal envelope and act to make repairs before the equipment fails and induces down-time. Non-intrusive load monitoring applications makes it possible to measure and understand the state history of a piece of equipment. This is useful for process controls, maintenance, and for cooperative safeguards matching true usage to declared usage for that equipment.

TECHNICAL APPROACH

The data was gathered in 5-minute segments. Each device was allowed to obtain steady state and then data was captured for 5 minutes. Following that, as soon as feasible 5 minutes of data was collected with the device off. This was done to minimize on/off differences in background in adjacent segments. 3 rounds of on/off segments were captured for each device with the intent of two segments being used for training/validation and a third segment to serve as test data. All data was gathered using a Saleae Logic Pro 16 [5]. The sample rate of the Saleae was 1.56 MS/s/channel. The EM/RF data was gathered by taking power from the wall socket and using a voltage divider box to generate a hot-neutral, hot-ground, and neutral-ground channel and connecting them all to the Saleae DAQ. The vibroacoustic data was gathered using a Polytec VibroFlex Xtra [6]. Data channels for the displacement, velocity, and acceleration were routed from the laser vibrometer into the Saleae DAQ. Because the signal magnitudes were so different, the laser vibrometer settings had to be recalibrated for each device. The same settings were used for all 'On' and 'Off' captures for each device. The point for the vibroacoustic laser was marked in sharpie on each device so that it was consistent across captures. The devices were affixed to the work bench such that the operation orientation could be consistent.

Item Number	Item Name	Item Description
1	Fan_36CFO	Fan
2	HighPace400_Turbo_cart	Turbo pump
3	Milwaukee_Magnum	Drill
4	Dremel_3000	Drill
5	Eraser_RT2S	Wire stripper
6	Air_King	Fan
7	Branson_1800	ultrasound
8	HighCube_Turbostation	Turbo pump
9	KSL_1100X_Central	Muffle Furnace
10	DeWalt_D26960	Heat Gun
11	Pfeiffer_Hipace_80	Turbo pump
12	Agilent_IDP7	Rough Pump
13	Agilent_IDP3	Rough Pump
14	Null	Null

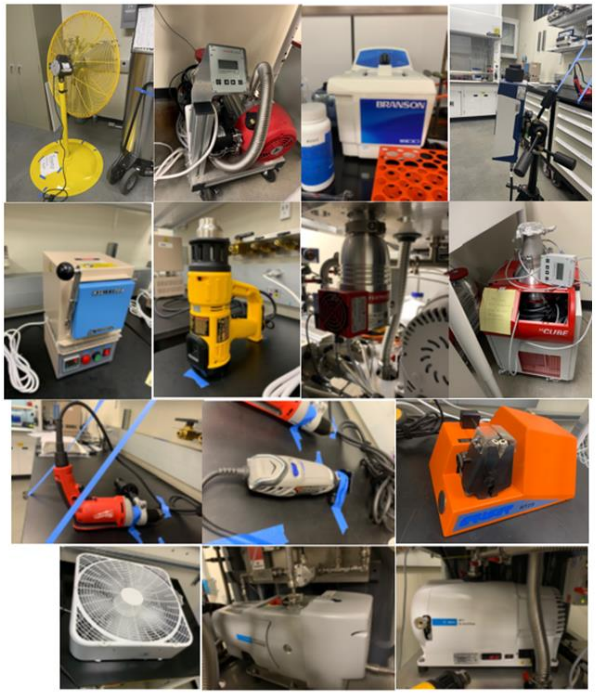

Figure 3. List and images of the equipment used in the dataset

RESULTS

The data and code for training and evaluating a baseline model from our first data collection campaign contains 5 main components/component types:

Ds2eda_environment.yml – This contains a list of dependencies for the python environment. Conda can automatically generate a suitable environment for running the rest of the code using this file.

Dataloaderotf.py – This contains the code for the dataloader which makes spectrograms from raw waveforms. The data is stored in waveforms and the code generates spectrograms for use by a machine learning model as inputs to the network on the fly. As such, the code is configured to use multiple CPUs to build batches fast enough to keep pace with the GPU.

Measurements_Summary_Metadata.pdf – This contains all the equipment settings and all the metadata associated with the data capture. Metadata of note includes item name, serial number, model, brand, description, time of the collections, and a long list of relevant laser vibrometer settings.

Equipment_State_Date_Time.hdf5.zip - Data files for all equipment in all states in all repetitions. Data is stored in the hdf5 format and contains 6 channels – 3 EM/RF followed by 3 Vibroacoustic. For details, see the associated notebook.

SpectrogramResNet.ipynb – This contains a Jupyter notebook that includes tutorials, comments, and code illustrating the process of building and analyzing a baseline classifier. Raw data is loaded and a tutorial describing how the raw multi-channel signals are converted into spectrograms is shown, including flexibility to adjust the input parameters to the spectrograms. Some example raw data and associated spectrograms are visualized to give the user a better intuition about the problem. The notebook then walks through setting up and training a ResNet50 (neural network) for predicting instrument type in PyTorch using a custom dataloader. Then, we provide some additional code illustrating how to evaluate the trained model on a test set, including comparing to baseline accuracy and displaying a confusion matrix so the user can gain a better understanding of the trained model's behavior. Finally, the notebook concludes with some suggestions and exercises for the user as to additional ways to explore the baseline model behavior and how the baseline model might be further improved. The baseline model is intentionally left with room for improvement as an educational tool. Depending on the available computing hardware, running the entire notebook end-to-end takes on the order of a few hours.

Baseline Model Performance

The ResNet50 baseline model provided along with our dataset performs decently well on the suggested task but does have room for improvement. This gap is intentional so that the code we provide can be used as an educational tool for the user to familiarize themself not only with the data but also with data science techniques. We modify the pre-trained ResNet50 to accept a 5-channel input (one of the calculated vibroacoustic channels is found to exhibit numerical instability). The baseline model uses spectrograms generated using nperseg of 2000, nfft of 5625, and an FFT width of 224*2000 so that the spectrogram image sizes match the image size expected by the pretrained ResNet50. We fine-tune the ResNet50 over 10 epochs with a batch size of 10, 5,000 samples in the training set, 1,000 samples in the validation set, and 1,000 samples in the test set. The test and train/val sets do not overlap; the test data comes from the first portion of the data collect (1st repetition), which the train/val data is sampled from the later data collects (2nd and 3rd repetitions). The baseline model is trained to classify which instrument is turned on or if all instruments are off. We find that this model achieves an average training accuracy of 99%, an average validation accuracy of 96%, and an average training accuracy of 83%. Figure 4 depicts the basic model processes. Figure 5 shows one set of time-correlated spectrograms for the EM/RF and Vibroacoustic measurements.

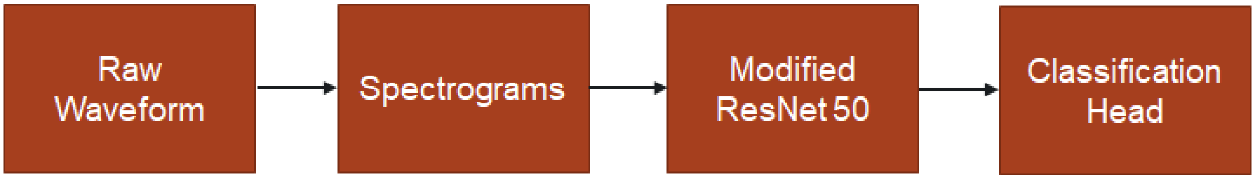

Figure 4. Block diagram of the data to classification pipeline

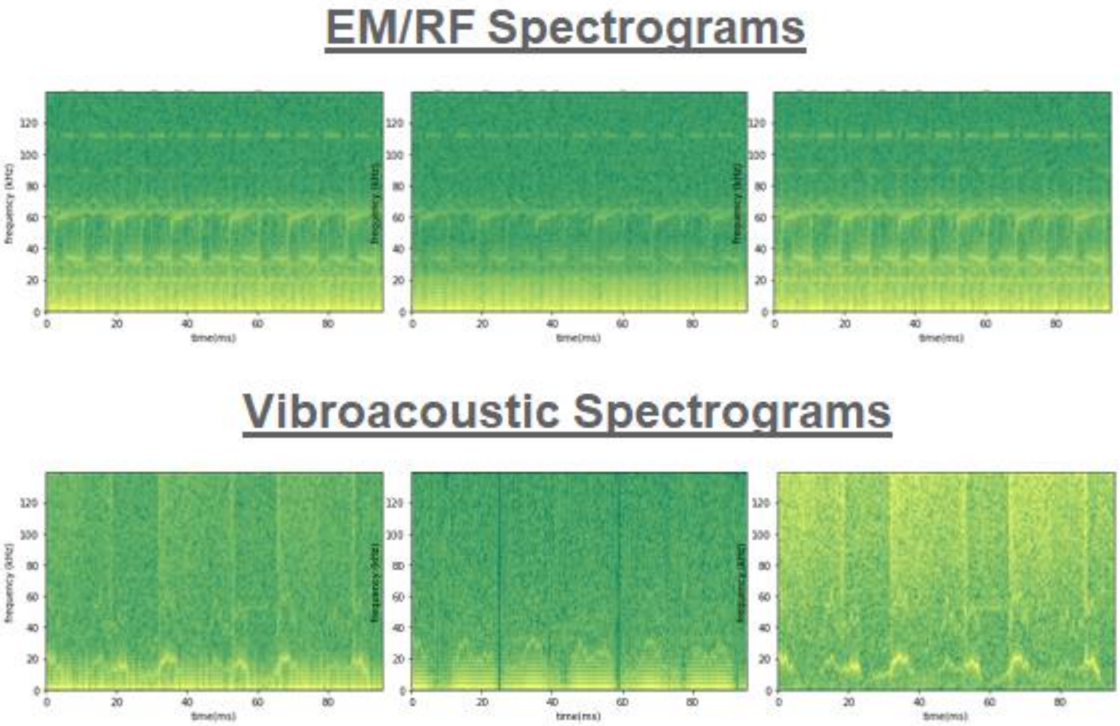

Figure 5. Sample data from roughing pump Agilent IDP7

NEXT STEPS

Following the generation, compilation, cleaning, modeling, and distribution of this initial dataset, subsequent efforts will focus on obtaining new data with appropriate modeling complexity for continued model development. Because of the relatively accurate performance of unsophisticated models in classifying the data, subsequent data will aim to reduce the quality of the signal being emitted by the devices. To this end, the project has constructed tunable circuits aimed at power conditioning to blunt the EM/RF signal and has identified new locations for gathering the vibroacoustic signal from which the expected signatures will be reduced.

REFERENCES

[1] Langlotz, Curtis & Allen, Bibb & Erickson, Bradley & Kalpathy-Cramer, Jayashree & Bigelow, Keith & Cook, Tessa & Flanders, Adam & Lungren, Matthew & Mendelson, David & Rudie, Jeffrey & Wang, Ge & Kandarpa, Krishna. (2019). A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology. 291. 190613. 10.1148/radiol.2019190613

[2] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6), 141–142.

[3] Krizhevsky, Alex, and Geoffrey Hinton. "Learning multiple layers of features from tiny images." (2009): 7.

[4] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009.

[5] Saleae "Saleae User Guide 1.1.15" http://downloads.saleae.com/Logic+Guide.pdf accessed 4-23-23

[6] Polytec "VibroFlex" https://www.polytec.com/us/vibrometry/products/single-point-vibrometers/vibroflex accessed 4-23-23