AMIRAL TECHNOLOGIES

BENCHMARK

Comparison of failure prediction models

04/2021

Sébastien Le Gall
CTO
Thibaut Le Magueresse
Data Scientist

Summary

The prediction of industrial equipment failures connected in blind mode is based on algorithms for detecting anomalies or new features on raw or transformed time series data. A number of algorithms have been compared on a set of time series databases that exhibit cyclicity [1]. The performance evaluation relies on the ROC [2] and Precision-Recall [3] curves to calculate a score for the genericity and relative performance of the models. It emerges that the problem of fault prediction is not an easy subject but that the association of Amiral Technologies algorithms, namely the features generator DiagFeatures (DF) and the anomaly detector BlindFaultDetector (BFD) seems to be the best genericity-performance-speed compromise.
A discussion on the relevance of these scores during the deployment of models in production, leads us to consider a scoring system which takes into account the operational constraints on which we have also noted the models.

Introduction

Predicting failures in connected industrial equipment involves anticipating a failure before it happens or detecting it as early as possible. The most recent approaches to solve this problem rely on signal processing and machine learning methods by creating learned models on historical data. The types of data generated by industrial equipment can be of different natures but this document focuses on a subset nevertheless representing large families of equipment, namely cyclic time series.

To create predictive models two approaches can be considered:

- the supervised approach of using healthy data (without failures) and unhealthy data (with failures) to create a model that will best classify a so-called "healthy" or "unhealthy" time series. This approach is difficult to implement today because the number of industrial breakdowns is often relatively low, and therefore insufficient to allow this kind of learning.
- the unsupervised approach of creating a model only from healthy data by creating a space of normality that represents the proper functioning of the equipment. When new data comes out of this space of normality, it will be considered unhealthy. This approach does not require a failure history to build the model (although a few occurrences are needed to validate it), so it is more easily applicable to the field of industrial failure prediction.

Machine learning algorithms for creating models in unsupervised mode are part of the class of anomaly detection or novelty detection algorithms. The best known in the literature are Isolation Forest [4] and OneClassSVM [5]. These methods apply to data windows of fixed size and can be used directly on the raw data or on data generated by preprocessing algorithms such as the Fourier transform [6] to name just one. More recently, deep learning methods have emerged. The best known use recurrent neural networks with an autoencoder architecture [7]. The principle is to compress the signal in a space of reduced dimension then to reconstruct it; the difference between the input signal and the reconstructed signal is used as an indicator of normality.

The objective of this document is to compare several anomaly detection algorithms on a set of relevant databases. The remainder of the document presents the implemented methodology, the studied algorithms and the datasets used before presenting the results.

Methodology

Comparing algorithms requires algorithms, but also training and test datasets, and finally a method of comparison. Regarding the algorithms, in addition to the algorithms of Amiral Technologies, a set of known algorithms was selected for the study. The goal was not to set up every algorithm on every dataset in order to get the best results but rather to use these algorithms with hyperparameters by default, which corresponds to the general case of blind learning. Five public databases were chosen according to certain criteria that we will develop later as well as a private database. The comparison criteria used are the area under the curve (AUC) [2] and the mean precision (AP) [3]. These two criteria are based respectively on the ROC and Precision-Recall curves.

In order to classify the algorithms studied, we opted for two criteria reflecting genericity and relative performance.

- Genericity (average score): This criterion is calculated by averaging the scores of the algorithm on all the datasets used.
- Relative performance (mean rank): This criterion is calculated by taking the average rank of the algorithm relative to the other algorithms studied, over all the datasets used.

Finally, we compared the time required to produce the model on a single core (except for neural networks which required graphics card (GPU) resources). To automate the comparison work, an automatic comparison framework algorithms have been developed. The databases selected generally present data from several sensors. As some algorithms are not suitable for multivariate processing, a model per variable was built, producing one prediction by sensor. Then, they were merged into a single global prediction by using a simple method similar to a weighted average by their prediction height.
The learning follows a cross-validation protocol to avoid any overfitting.

Models / Algorithms

In the rest of this document, a model (Figure 1) is defined as a sequence of algorithms comprising a feature generation algorithm and a novelty / anomaly detection algorithm. In the case of neural methods, the features extraction step does not exist.

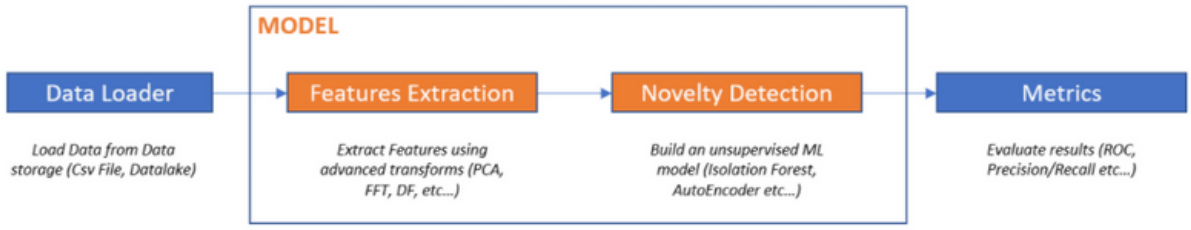

Figure 1 : Diagram of a predictive model

In this study, three feature generation algorithms and five anomaly detection algorithms were tested.

Features generation

- FFT [6]: The Fourier transform is a function which transforms a time series into a series of information describing its frequency spectrum.

- Tsfresh [8]: Characteristic generator based on signal processing methods in different domains (temporal, frequency, spectral).

- DF: **DiagFeatures is the set of characteristic generators owned by Amiral Technologies.** It is adapted to industrial cyclic time series, comes from the field of research in signal processing and automation.

Detection of anomalies

- IF [4]: Isolation Forest calculates an anomaly score for each observation in the database. To calculate this score, the algorithm isolates the data recursively, it randomly chooses a variable and a cutoff threshold at random, and assesses whether this isolates a particular observation.

- LOF [9]: This algorithm makes it possible to find anomalies by measuring local deviations of a point with respect to its neighbors.

- OCSVM [5]: The One Class SVM algorithm is based on the machine support vector method but suitable for the detection of anomalies.

- AE [10]: The LSTM-AutoEncoder is based on neural networks. The idea is to encode the time series in a lower dimensional space and then try to reconstruct the input signal. The algorithm is trained on healthy data. The difference between the reconstructed signal and the original signal is seen as the distance from the space of normality. The network architecture is based on LSTM cells.

- FORE [11]: LSTM-Forecasting has the same approach as LSTM-Autoencoder but instead of reconstructing the input signal, the algorithm tries to predict future data. The observed deviation between the prediction and the actual data is the distance from the space of normality. The network architecture is based on LSTM cells.

- BFD: **The Blind Fault Detector is an algorithm developed by Amiral Technologies** operating on the principle of anomaly detection in the same way as Isolation Forest or OneClassSVM. It is fast in computing speed and was designed for use with DiagFeatures.

Databases

For this study, it was necessary to select a set of relevant databases with the following criteria:

- Containing cyclic time series
- Specific to the detection of anomalies on industrial equipment
- Ideally multi-sensors
- Having occurrences of failures in order to validate the model
- Having a sampling frequency greater than or equal to 1Hz

Five public databases were selected as well as a proprietary database:

- **Zema [12]:** The data deal with the evaluation of the state of a hydraulic test bench on measurements coming from several sensors. The database has several occurrences of failures with different levels of severity.

- **Hai [13]:** This database was developed for research in the detection of anomalies in cyber-physical systems such as railways, water treatment and power plants.

- **BearingVibration [14]:** The data in this database contains vibration signals collected from bearing systems in different health states, and under conditions of rotational speed varying over time.

- **CentrifugalPump [15]:** The vibration data is collected on a self-priming centrifugal pump data acquisition system. Data is collected under normal and fault conditions, including bearing roller wear, inner ring wear and outer ring wear fault conditions, as well as condition of wheel wear defect.

- **BatteryAging [16]:** This dataset was collected from a custom battery prognosis test bench at the NASA Ames Center of Excellence for Prediction (PCoE). Li-ion batteries have undergone 3 different operational profiles (charge, discharge and electrochemical impedance spectroscopy) at different temperatures. The discharges were performed at different current load levels until the battery voltage fell to the preset voltage thresholds. Some of these thresholds were lower than that recommended by the manufacturer (2.7 V) in order to induce aging effects by deep discharge. Repeated charging and discharging cycles lead to accelerated aging of the batteries.

- **DigitalTwin:** This is a synthetic database created by Amiral Technologies. Although it is synthetic, it nevertheless represents a real system, more precisely a double pendulum system, namely a 6-state nonlinear system. From this "digital twin", a database of 10 sensors was generated. The base contains 290 cycles of size 200.

For public databases, data formatting as well as cleaning operations (replacement of Nan values, linearization of timestamps, division into cycles, resampling, selection of sensors, adaptation of labels) had to be applied so that the whole models can work and our comparison tool can operate.

Figure 2 summarizes the characteristics of formatted test databases.

Name	Number of healthy cycles	Number of unhealthy cycles	Number of sensors	Sampling frequency	Cycles' size
ZEMA	10	119	7	100 Hz	6000
HAI	5303	25	2	1 Hz	100
BEARINGVIBRATION	198	216	1	5000 Hz	5000
CENTIFUGALPUMP	81	80	5	100 Hz	256
BATTERYAGING	438	206	6	10 Hz	1000
DIGITALTWIN	200	90	10	N/A	200

Figure 2: Summary table of formatted databases

Results

The naming of the models in the remainder of this document is the concatenation of the novelty detection algorithm with the name of the characteristic generator separated by the "_" character. For example, for a model using Isolation Forest (IF) and DiagFeatures (DF), its name is “IF_DF”. If there is no features generator, the term None is used, except for neural networks.

Figure 3 shows a summary table of the areas under the curve (AUC). As a reminder, the classification criteria used are:

- Average Score: Average score obtained across all databases. This criterion is a measure of the genericity of the models.
- Mean rank: Average rank of the model across all databases. This criterion is a measure of the relative performance of the models.
- Time (s): Cumulative learning and prediction time in seconds on a single core. Only methods based on neural networks use GPU resources.

	ZEMA	BearingVibration	HAI	CentrifugalPump	DigitalTwin	BatteryAging	Average Score	Mean Rank	Time (s)
BFD_DF	0.881033	0.995669	1.000000	0.998211	0.968254	0.877758	0.953487	4.083333	87.392006
LOF_DF	0.841190	0.999184	0.996781	0.984797	0.755767	0.928937	0.917763	7.666667	88.944392
BFD_None	0.937149	0.855884	0.891344	0.996752	0.991323	0.790977	0.910571	8.333333	13.807184
OCSVM_DF	0.824916	1.000000	0.999790	0.999574	0.738677	0.815495	0.896409	5.916667	74.165996
IF_None	0.959596	0.789576	0.722406	0.997557	0.998095	0.836769	0.884000	7.833333	39.481851
OCSVM_Tsfresh	0.805836	1.000000	0.999824	0.998345	0.681164	0.775511	0.876780	7.416667	24218.436123
IF_Tsfresh	0.598348	0.989546	0.869537	0.998727	0.969418	0.838949	0.874754	9.166667	26614.977597
OCSVM_None	0.926487	0.910096	1.000000	0.998911	0.560317	0.834653	0.871744	6.416667	7.974381
IF_DF	0.662177	0.994698	0.768736	0.993675	0.979471	0.793858	0.865436	8.666667	103.598173
BFD_Tsfresh	0.597082	0.989098	0.998474	0.998363	0.747937	0.805319	0.856045	8.166667	26048.705819
LOF_Tsfresh	0.656566	0.936231	0.998038	0.981921	0.895873	0.659883	0.854752	10.833333	25599.985191
FORE	0.659371	0.872312	0.999891	0.883511	0.975026	0.709556	0.849945	10.000000	1056.018176
AE	0.944444	0.895759	0.991197	0.992311	0.821481	0.306461	0.825276	9.833333	1849.135605
LOF_FFT	0.554433	0.775314	0.932257	0.977148	0.921958	0.650278	0.801898	14.166667	19.529669
OCSVM_FFT	0.543771	0.874627	0.997233	0.999899	0.611323	0.779295	0.801025	10.166667	28.978763
LOF_None	0.864759	0.848865	0.988556	0.988589	0.461693	0.621294	0.795613	12.666667	16.724266
IF_FFT	0.448934	0.787186	0.954039	0.987464	0.936508	0.244445	0.726429	14.166667	38.111975
BFD_FFT	0.355219	0.793832	0.894856	0.987581	0.538794	0.413764	0.662674	15.500000	67.101667

Figure 3 : Summary table of results based on AUC

Figure 4 presents the same table as Figure 3 but this time based on the mean precision (AP) criterion.

	ZEMA	BearingVibration	HAI	CentrifugalPump	DigitalTwin	BatteryAging	Average Score	Mean Rank	Time (s)
BFD_DF	0.792909	0.998572	1.000000	0.999398	0.976737	0.756046	0.920610	3.750000	87.392006
OCSVM_DF	0.736075	1.000000	0.925926	0.999852	0.743661	0.627378	0.838815	5.416667	74.165996
OCSVM_None	0.871220	0.976937	1.000000	0.999638	0.499089	0.658957	0.834307	6.083333	7.974381
FORE	0.471509	0.965421	0.980036	0.947441	0.979015	0.555920	0.816557	10.000000	1056.018176
OCSVM_Tsfresh	0.711766	1.000000	0.937778	0.999443	0.644025	0.528742	0.803626	7.250000	24218.436123
LOF_Tsfresh	0.589459	0.931487	0.590439	0.993128	0.983417	0.559739	0.761278	12.000000	25599.985191
LOF_DF	0.683307	0.999745	0.478295	0.994815	0.613524	0.791022	0.760118	8.500000	88.944392
IF_None	0.911428	0.928551	0.018725	0.999162	0.998075	0.689193	0.757522	8.000000	39.481851
BFD_Tsfresh	0.439287	0.996209	0.727323	0.999450	0.784581	0.644903	0.751959	8.166667	26048.705819
BFD_None	0.833622	0.949929	0.111158	0.998891	0.993268	0.680963	0.747972	7.833333	13.807184
OCSVM_FFT	0.423945	0.966262	0.717141	0.999966	0.582139	0.598535	0.714665	9.666667	28.978763
IF_DF	0.513834	0.998347	0.049807	0.997892	0.985131	0.624633	0.694941	8.666667	183.598173
AE	0.787161	0.969058	0.214047	0.997647	0.871615	0.271312	0.685148	10.500000	1849.135605
IF_Tsfresh	0.463724	0.996985	0.033348	0.996547	0.952959	0.661861	0.684237	9.500000	26614.977597
LOF_FFT	0.403434	0.935494	0.086716	0.992186	0.926153	0.430430	0.629069	14.333333	19.529669
LOF_None	0.679582	0.949138	0.252884	0.995655	0.433708	0.460374	0.628544	13.166667	16.724266
BFD_FFT	0.297965	0.943292	0.592499	0.995855	0.535633	0.288794	0.609086	14.000000	67.101667
IF_FFT	0.335304	0.941683	0.094885	0.995800	0.925588	0.239865	0.588854	14.166667	38.111975

Figure 4 : Summary table of results based on AP

According to the 2 tables above and whatever the classification methods used, the conclusions are as follows:

- The Amiral Technologies characteristic generator upstream of a novelty detection algorithm improves the performance of the model.
- Neural networks have good results on some bases but very bad results on others. These methods seem less generic by applying a common setting to all the databases.
- There is no model that surpasses the others, thus highlighting the difficulty of deploying a single model.

By combining all the metrics, the BFD_DF model emerges as the best model on these comparison datasets. In the majority of applications, failure prediction must be done in real time on several devices at the same time. Calculation time is therefore a criterion that should not be neglected.

Figure 5 shows the Performance (AUC) -Speed tradeoff on a 2D graph. On this criterion, the BFD_DF model from Amiral Technologies stands out from other models.

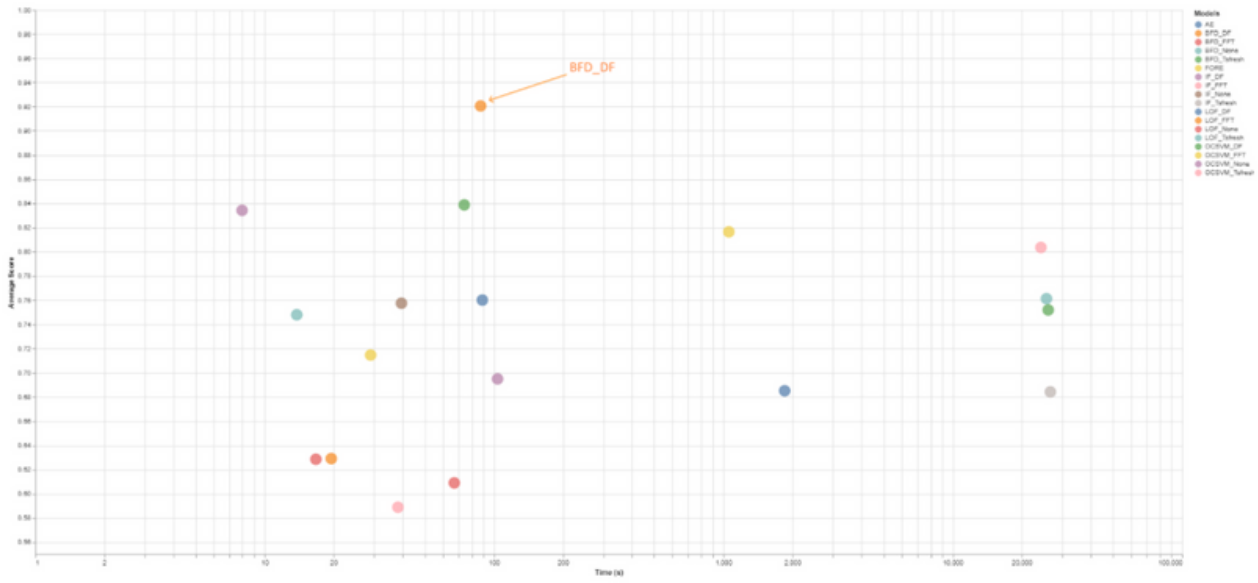

Figure 5 : Compromise Speed-Performance

Discussion

The document presents a first comparison of novelty detection algorithms on cyclic time series data. Metrics that are based on ROC and Precision-Recall curves are subject to discussion. First, a model that predicts a failure long before the ground truth has labeled it as such is penalized. Second, the calculation of the curve is based on the predictions of the algorithms individually without taking into account the temporal notion of anomalies. Work should be undertaken on modifying the calculation of the prediction curves. On the other hand, the model that is deployed in production to predict whether a cycle is healthy or not must necessarily have a decision threshold. Thus the model only corresponds to one point of the ROC (or Precision-Recall) curve. A bad selection of this threshold can have big consequences on the performance of the model.
To better illustrate the problem, an automatic threshold calculation based on the desired false acceptance rate (set at 1% in this study) was applied to all models.

Figure 6 shows the comparison results of these models based on the F1-Score [3].

	ZEMA	BearingVibration	HAI	CentrifugalPump	DigitalTwin	BatteryAging	Average Score	Mean Rank	Time (s)
BFD_DF	0.771429	1.000000	0.991332	0.993423	0.700000	0.784867	0.873375	4.750000	87.392086
OCSVM_DF	0.792793	0.975207	0.983271	0.887095	0.695652	0.781316	0.852556	8.666667	74.165996
IF_DF	0.771429	0.925373	0.851549	0.978286	0.700000	0.784867	0.835117	7.583333	103.598173
LOF_DF	0.771429	0.851852	0.809482	0.990359	0.700000	0.786540	0.818277	6.583333	88.944392
BFD_FFT	0.765957	0.540881	0.996115	0.920979	0.700000	0.764643	0.781429	8.000000	67.101667
AE	0.650000	0.571429	0.988016	0.920913	0.700000	0.781415	0.768629	9.083333	1849.135685
IF_Tsfresh	0.765957	0.202899	0.885305	0.976736	0.700000	0.784867	0.784827	9.916667	26614.977597
LOF_FFT	0.727273	0.121212	0.888863	0.979783	0.700000	0.786540	0.699279	9.166667	19.529669
LOF_Tsfresh	0.765957	0.092308	0.933855	0.919053	0.700000	0.782700	0.698979	10.000000	25599.985191
IF_FFT	0.757143	0.342342	0.961869	0.726175	0.700000	0.651163	0.689782	11.250000	38.111975
LOF_None	0.739130	0.031746	0.916913	0.983393	0.700000	0.785263	0.679408	10.750000	16.724266
BFD_Tsfresh	0.765957	0.389610	0.919855	0.388066	0.700000	0.786540	0.645005	9.416667	26048.705819
FORE	0.739130	0.171429	0.987049	0.555032	0.700000	0.697059	0.641616	11.666667	1056.018176
BFD_None	0.771429	0.031250	0.854504	0.469114	0.700000	0.785263	0.601927	10.750000	13.807104
OCSVM_None	0.888889	0.000000	0.922903	0.212984	0.700000	0.777542	0.583720	11.500000	7.974381
OCSVM_FFT	0.000000	0.390805	0.990692	0.611504	0.702509	0.786540	0.580342	7.416667	28.970763
IF_None	0.776978	0.000000	0.804652	0.354053	0.700000	0.784867	0.569959	11.916667	39.481851
OCSVM_Tsfresh	0.607595	0.062500	0.992503	0.052845	0.700000	0.776246	0.531948	12.583333	24218.436123

Figure 6 : Comparison of F1-Score

Taking the OCSVM_None model as a reference, although it has an honorable ROC curve on the BearingVibration database, the F1-Score is 0.0. This model is therefore not operational at all . A good model is therefore a model having a good curve with a good decision threshold. Here the models using DiagFeatures are the most robust.

The improvement of the method of calculating the performance of a model as well as the research and development of a better algorithm for the automatic detection of decision thresholds will be the subject of a forthcoming scientific publication.

Finally, the comparison does not include the implementation of an Auto-ML process [17] on the algorithms, which would be interesting to test to complete the comparison work.

Conclusion

This article is a first version that provides a quick overview of the behavior of failure prediction algorithms in unsupervised mode. It compares a set of algorithms from both signal processing, automation, machine learning and deep learning. The comparison was made on a set of databases related to the field of industrial failure prediction. It shows that the BFD_DF model from Amiral Technologies presents the best genericity-performance-speed compromise on all the test databases. In the case which interests us, namely the prediction of failures in blind mode, these criteria are decisive. In addition, when we approach the actual implementation with definition of the decision threshold, the DiagFeatures generator emerges as the one that gives the best results.

References

[1] V.Heurtin, "Data Cyclicity," January 2021. [Online]. Available: https://www.amiraltechnologies.com/actualite/2020/12/data-cyclicity/.
[2] A. Géron, Machine Learning avec Scikit-Learn, mise en oeuvre et cas concret", 2nd édition, p. 90.
[3] A. Géron, Machine Learning avec Scikit-Learn, mise en oeuvre et cas concret", 2nd édition, p. 94.
[4] F. T. Liu, K. M. Ting and Z.-H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining: 413–422. doi:10.1109/ICDM.2008.17. ISBN 978-0-7695-3502-9. S2CID 6505449, 2008.
[5] S. S. Khan and M. G. Madden, "A Survey of Recent Trends in One Class Classification," Artificial Intelligence and Cognitive Science. Lecture Notes in Computer Science. Springer Berlin Heidelberg. 6206: 188–197. doi:10.1007/978-3-642-17080-5_21. hdl:10379/1472. ISBN 9783642170805, 2010.
[6] J. W. a. J. W. T. Cooley, "An algorithm for the machine calculation of complex Fourier series," Math. Comput. 19: 297-301, 1965.
[7] Wikipedia contributors, "Autoencoder," [Online]. Available: https://en.wikipedia.org/w/index.php?title=Autoencoder&oldid=1008370584.
[8] M. K.-L. A. a. F. M. Christ, "Distributed and parallel time series feature extraction for industrial big data applications," 2016.
[9] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: Identifying Density-based Local Outliers," Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD. pp. 93–104. doi:10.1145/335191.335388. ISBN 1-58113-217-4, 2000.
[10] Machine Learning Mastery, "A Gentle Introduction to LSTM Autoencoders," [Online]. Available: https://machinelearningmastery.com/lstm-autoencoders/.
[11] Publici sapiens, "Long Short-Term Memory (LSTM) Networks for Time Series Forecasting," October 2020. [Online]. Available: https://blog.engineering.publicissapient.fr/2020/09/23/long-short-term-memory-lstm-networks-for-time-series-forecasting/ .
[12] E. P. A. S. Nikolai Helwig, "Condition Monitoring of a Complex Hydraulic System Using Multivariate Statistics," Proc. I2MTC-2015 - 2015 IEEE International Instrumentation and Measurement Technology Conference, paper PPS1-39, Pisa, Italy, May 11-14, 2015, doi: 10.1109/I2MTC.2015.7151267.
[13] W. L. J.-H. Y. a. H. K. Hyeok-Ki Shin, "HAI 1.0: HIL-based Augmented ICS Security Dataset," 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET 20), Santa Clara, CA, 2020.
[14] N. B. M. L. H. Huang, "Bearing fault diagnosis under unknown time-varying rotational speed conditions via multiple time-frequency curve extraction," J. Sound Vib., 414 (2018).
[15] Y. W. R. C. Chen Lu, "Fault Diagnosis for Rotating Machinery: A Method based on Image Processing," 2016.
[16] DAWN MCINTOSH, "Li-ion Battery Aging Datasets," [Online]. Available: https://c3.nasa.gov/dashlink/resources/133/.
[17] Wikipedia contributors, "Automated machine learning," [Online]. Available: https://en.wikipedia.org/w/index.php?title=Automated_machine_learning&oldid=1002595487.

AMIRAL TECHNOLOGIES

AMIRAL TECHNOLOGIES

CONTACT US

contact@amiraltechnologies.com

Amiral Technologies
31 rue Gustave Eiffel
38000 Grenoble
France