MARCH 2023

COMPARISON OF ANOMALY DETECTION MODELS IN AN INDUSTRIAL CONTEXT

APPLICATION TO NON-CYCLIC DATASETS

Thibaut Le Magueresse, Ph. D., Data Scientist
Sébastien Le Gall, Chief Technical Officer
Amiral Technologies

Abstract

Anomaly detection task applied to industrial time series is still a hot topic in the scientific community. The sensor data contains the nominal functioning information of the monitored equipment, generally free from defect. The goal is then to set up algorithms designed to learn the normal behavior of the equipment and to distinguish defects from the normality. In that context, this study highlights the difficulty to reach genericity in terms of performance based on state-of-the-art algorithms. This genericity is a necessary characteristic to be able to scale a Machine Learning software addressed to different clients and industries. To address this need, Amiral Technologies edits such a software called DiagFit®. It includes a combination of models that exploits both mono-variated and multi-variated information to efficiently detect a wide variety of defects while providing explicability. The present publication aims at evaluating generic Machine Learning algorithms applied on ten multivariate datasets. It assesses the performances of DiagFit compared to classical approaches based on Matthews Correlation Coefficient (MCC) and the Area Under Curve criteria (AUC), as well as equivalent carbon dioxide emissions over a selection of ten datasets representing a variety of industrial usages and sensor characteristics.

Introduction

Several types of Machine Learning algorithms have been developed to treat the detection anomaly problem in multivariate time series. Among them, supervised approaches consist in classifying healthy and non-healthy data based on fully labelled data. This approach is difficult to apply in the industry because of the non-exhaustivity and the scarcity of the failure types. In contrast, the one-class classification approaches are less constraining because they only need healthy data. The model is then built using only the healthy instances to infer a normality space containing the inherent structure of the healthy data. The anomaly presence is characterized by its distance with this normality space. Then we fall into novelty detection paradigm where the purpose is to detect new behavior of the equipment interpreted as anomaly in the sense that it has never been seen in the training phase. This new behavior indicates generally an early warning sign used to predict equipment failure.

This paper aims at completing our first benchmark focused on the performance of failure prediction models applied on cyclic datasets [1]. The proposed methodology follows the same as the previous study applied this time to non-cyclic datasets, which do not present any cyclic pattern in the time series. Ten industrial datasets containing various anomalies have been selected and split into train and test subsets. Each model provides an anomaly score over time which should be high in presence of anomalies. One may then extract comparison criteria to assess the performance of each model.

ML models

For tackling the problem of anomaly detection, two main approaches are considered in the present paper. First, the features-based method aims at splitting the time series into snapshots, extracting some significant features from them, and finally building a normality space encapsulating these features. Second, the cross-correlation based approach aims at focusing on invariant relationships that occur between sensors. The DiagFit model encompasses the two approaches into one global model, by extracting features and take care of the cross-correlation relationship between sensors. Below we present the list of the benchmarked models:

- Local Outlier Factor (LOF) (2): a density-based approach which considers that an outlier is placed in a region of low density of points. LOF computes the ratio between the density of the current point and the average of densities of the neighbors.
- Isolation Forest (IF) (3) (4): a tree-based method consisting in building an ensemble of trees to isolate sequences of data. The number of built branches for isolating the actual point gives the anomaly score.
- One Class Support Vector Machine (OCSVM) (5): domain-based approach consisting in building a boundary that separates the normal and the abnormal data point. This boundary may be linear in the data space or in another non-linear (feature) space after applying the "kernel-trick": the margin optimization problem is resolved by writing the dual problem where a mathematical trick allows to calculate the boundary without explicitly calculating the projection of the data in the feature space.
- Auto Encoder (6): a neural network which is trained to reproduce the input data. The idea is to create a "bottleneck" on the center of the network which plays the role of data compression. The left part of the network is the encoder, and the right part is the decoder. This bottleneck enforces the algorithm to represent the data with a small number of parameters. After being trained on healthy data, the test data is given to the auto-encoder. The difference between the prediction and the real observation gives a score of anomalies.
- Multivariate Time-series Anomaly Detection via Graph Attention Network (MTADVAE) (7): a neural network based on graph attention layers which learns the signal information in both sensors and temporal dimensions.
- DiagFit®: a commercial software developed by Amiral Technologies learning the normality space including mono-variated features of time series and multivariate relationships between sensors.

The three methods LOF, IF and OCSVM have all been combined with a features extraction module as pre-processing step. For this paper, the TS Fresh module (8) which computes common mono-variated time series characteristics has been benchmarked.

Dataset presentation

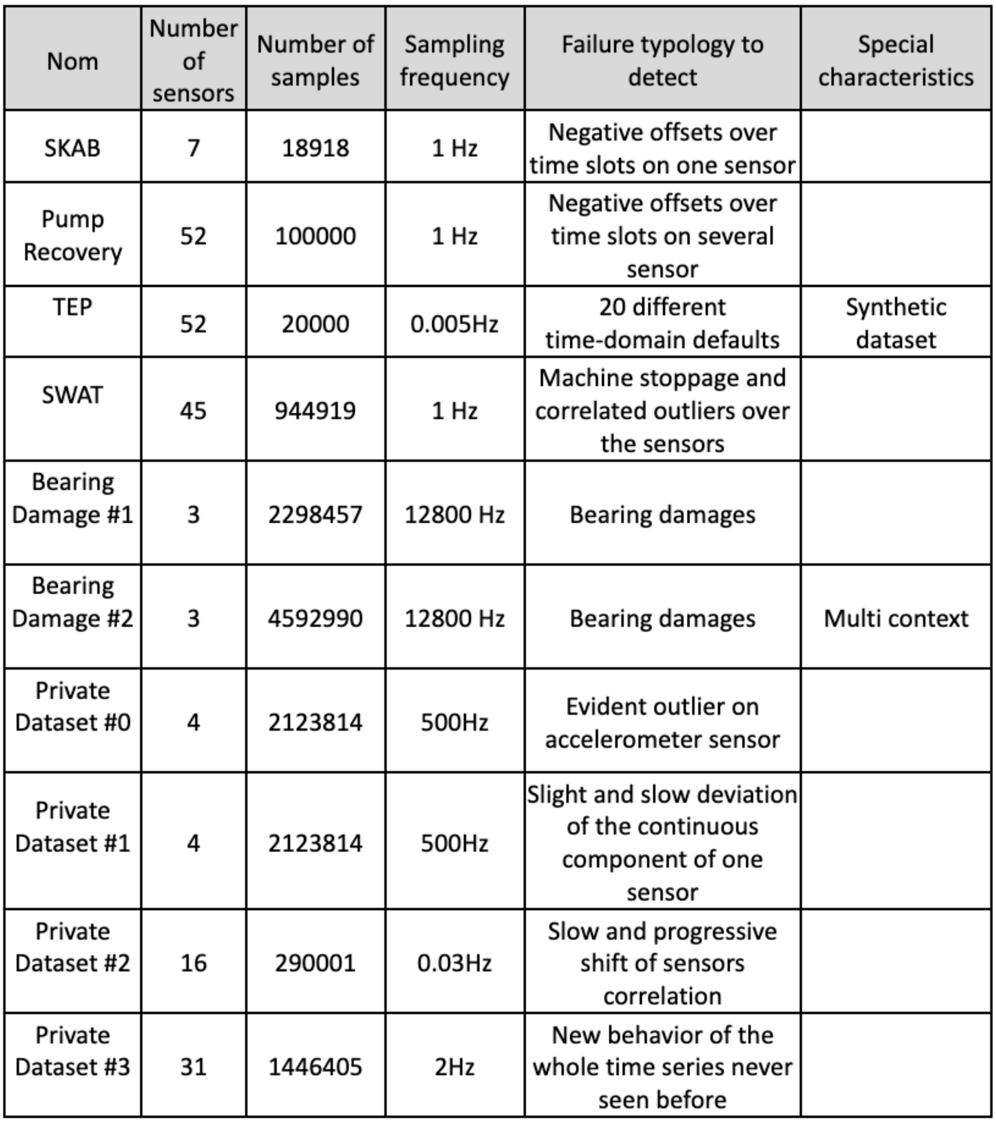
Ten different datasets have been selected for evaluating the algorithm performances. The selection criteria include:

- Absence of significant periodicities in the signal,
- Presence of several sensors,
- Presence of failures for the evaluation of the algorithms,
- Measurements from an industrial process.

The selected datasets are the following:

- Skoltech Anomaly Benchmark (SKAB) (9) : dataset representing a multivariate time series collected from the sensors installed on a water circulation system-based testbed.
- Pump Recovery (10): Water pump system dataset affected by seven system failures.
- Tennessee Eastman Process Simulation Dataset (TEP) (11) : dataset containing simulations of Tennessee Eastman Process, an industrial process composed by a two-phase reactor, a separator, a stripper, a compressor, and a mixer.

- Secure water treatment (SWAT) [12]: dataset containing sensors and actuators taken from a management system of hydraulic system.
- Bearing Damage [13]: dataset containing motor currents and vibration measurements of rolling bearings
- Private dataset #0: dataset containing high frequency sampling measurements recorded in a flying vehicle perturbed by an evident fault.
- Private dataset #1: same source as private dataset #0, showing a slow sensor deviation.
- Private dataset #2: dataset containing low frequency sampling and sensors-correlated measurements provided by an underwater industry player.
- Private dataset #3: dataset containing new behavior in the test part never seen in the train data extracted from a military vehicle.
- The following table depicts the main characteristics of the selected datasets. One may observe the variety in terms of sampling frequencies ranging from 0.005 Hz to 12800Hz. Each of them holds a specific anomaly. In regards with this variety, this benchmark evaluates the capacity of the algorithms to be generic and agnostic.



Nom	Number of sensors	Number of samples	Sampling frequency	Failure typology to detect	Special characteristics
SKAB	7	18918	1 Hz	Negative offsets over time slots on one sensor	
Pump Recovery	52	100000	1 Hz	Negative offsets over time slots on several sensor	
TEP	52	20000	0.005Hz	20 different time-domain defaults	Synthetic dataset
SWAT	45	944919	1 Hz	Machine stoppage and correlated outliers over the sensors	
Bearing Damage #1	3	2298457	12800 Hz	Bearing damages	
Bearing Damage #2	3	4592990	12800 Hz	Bearing damages	Multi context
Private Dataset #0	4	2123814	500Hz	Evident outlier on accelerometer sensor	
Private Dataset #1	4	2123814	500Hz	Slight and slow deviation of the continuous component of one sensor	
Private Dataset #2	16	290001	0.03Hz	Slow and progressive shift of sensors correlation	
Private Dataset #3	31	1446405	2Hz	New behavior of the whole time series never seen before	

Figure 1: Summary table of dataset characteristics

Time series extracted from the datasets

Some examples of public time series are shown in the table below. For each dataset, one randomly chosen sensor has been displayed over time. The diversity of signals over the datasets is notable.

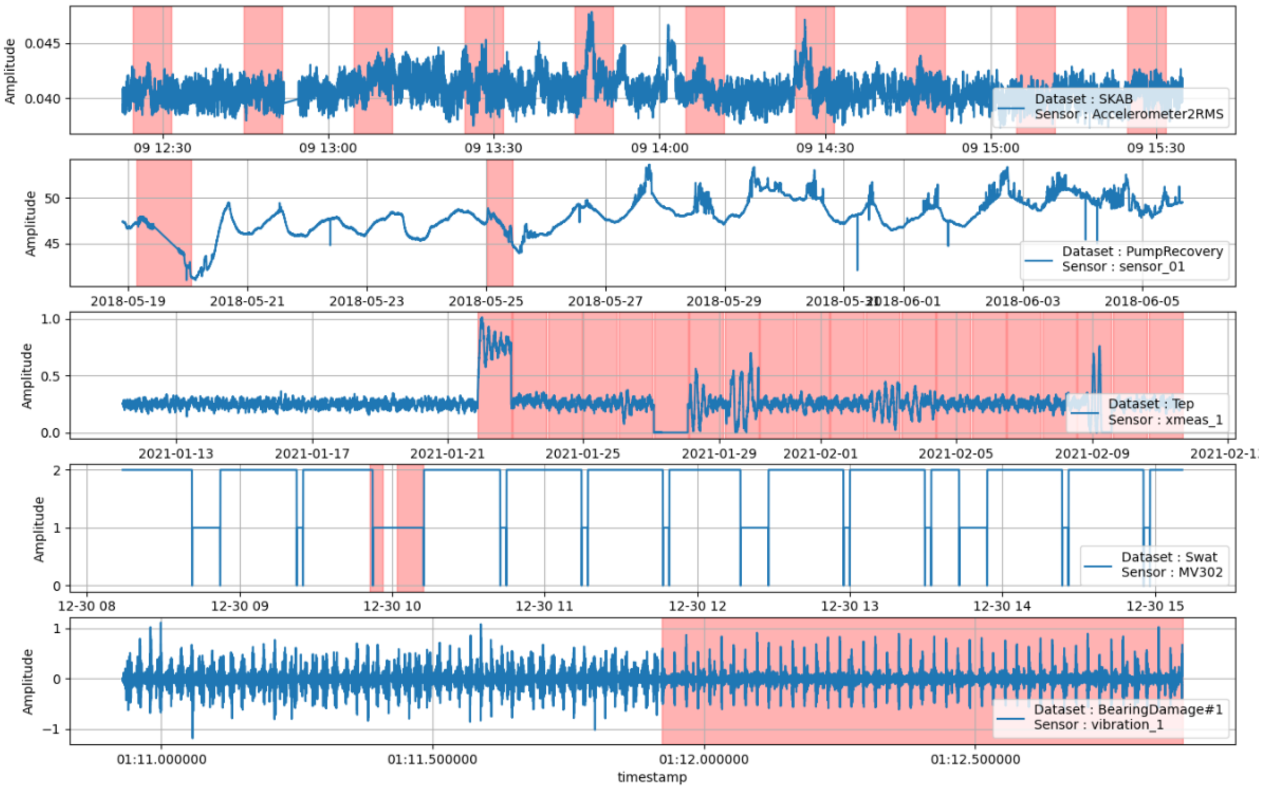


Figure 2: example of time series extracted from public datasets.
Each graph shows 25000 samples of one sensor (except for SKAB dataset which contains around 18000 samples in total).
The red area indicates the anomaly presence.

Results

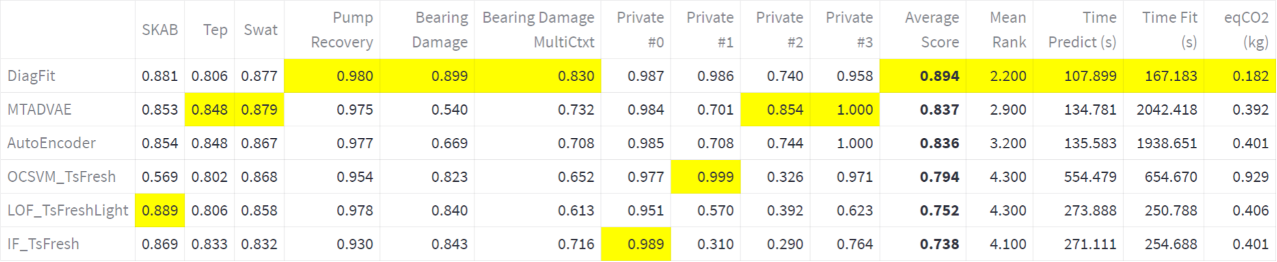
For evaluating the algorithms, two performance metrics have been analyzed: the Matthews Correlation Coefficient (MCC) and the Area Under Curve (AUC). The relevance of the first metric has been demonstrated by Zmitri [14] while the second is widely used in binary classification.

Moreover, the time cost of each model has been monitored. An estimation of the amount of equivalent carbon dioxide (CO2) has been done using the CodeCarbon package [15]. This quantity is the product of the carbon intensity of the electricity consumed (depends on the type of energy source) and the power consumed by the hardware for the computation (depends on the execution time and the efficiency of the computer). Both model training and model inference have been accounted for in the eqCO2 estimation. The learning cost has been directly estimated by the package as a linear relationship with the computation time. The inference cost has been extrapolated by considering that the model processes the equivalent of one day of data. The presented eqCO2 is the sum of the two parts (learning and extrapolated inference times).

Figure 3 and Figure 4 show a summary table of the three criteria of interest described above. In addition to MCC, AUC and eqCO2 metrics, the tables show three additional metrics:

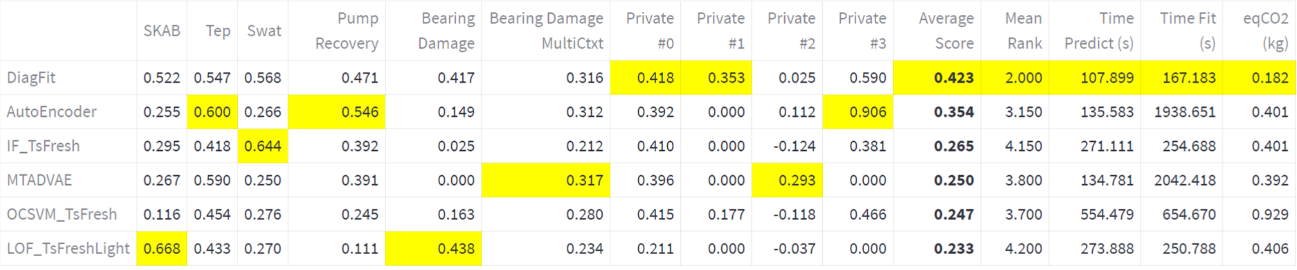
- Average Score: Average score obtained across all databases. This criterion is a measure of the genericity of the models.
- Mean rank: Average rank of the model across all databases. This criterion is a measure of the relative performance of the models.
- Time (s): Cumulative learning and prediction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

The different algorithms are sorted descending by Average Score to clearly highlight the best generic algorithm.



	SKAB	Tep	Swat	Pump Recovery	Bearing Damage	Bearing Damage MultiCtxt	Private #0	Private #1	Private #2	Private #3	Average Score	Mean Rank	Time Predict (s)	Time Fit (s)	eqCO2 (kg)
DiagFit	0.881	0.806	0.877	0.980	0.899	0.830	0.987	0.986	0.740	0.958	**0.894**	2.200	107.899	167.183	0.182
MTADVAE	0.853	0.848	0.879	0.975	0.540	0.732	0.984	0.701	0.854	1.000	**0.837**	2.900	134.781	2042.418	0.392
AutoEncoder	0.854	0.848	0.867	0.977	0.669	0.708	0.985	0.708	0.744	1.000	**0.836**	3.200	135.583	1938.651	0.401
OCSVM_TsFresh	0.569	0.802	0.868	0.954	0.823	0.652	0.977	0.999	0.326	0.971	**0.794**	4.300	554.479	654.670	0.929
LOF_TsFreshLight	0.889	0.806	0.858	0.978	0.840	0.613	0.951	0.570	0.392	0.623	**0.752**	4.300	273.888	250.788	0.406
IF_TsFresh	0.869	0.833	0.832	0.930	0.843	0.716	0.989	0.310	0.290	0.764	**0.738**	4.100	271.111	254.688	0.401

Figure 3: Summary table of results based on AUC. In yellow, the best value for each column.



	SKAB	Tep	Swat	Pump Recovery	Bearing Damage	Bearing Damage MultiCtxt	Private #0	Private #1	Private #2	Private #3	Average Score	Mean Rank	Time Predict (s)	Time Fit (s)	eqCO2 (kg)
DiagFit	0.522	0.547	0.568	0.471	0.417	0.316	0.418	0.353	0.025	0.590	**0.423**	2.000	107.899	167.183	0.182
AutoEncoder	0.255	0.600	0.266	0.546	0.149	0.312	0.392	0.000	0.112	0.906	**0.354**	3.150	135.583	1938.651	0.401
IF_TsFresh	0.295	0.418	0.644	0.392	0.025	0.212	0.410	0.000	-0.124	0.381	**0.265**	4.150	271.111	254.688	0.401
MTADVAE	0.267	0.590	0.250	0.391	0.000	0.317	0.396	0.000	0.293	0.000	**0.250**	3.800	134.781	2042.418	0.392
OCSVM_TsFresh	0.116	0.454	0.276	0.245	0.163	0.280	0.415	0.177	-0.118	0.466	**0.247**	3.700	554.479	654.670	0.929
LOF_TsFreshLight	0.668	0.433	0.270	0.111	0.438	0.234	0.211	0.000	-0.037	0.000	**0.233**	4.200	273.888	250.788	0.406

Figure 4: Summary table of results based on MCC. In yellow, the best value for each column.

According to the two tables above, the general observations are:

- DiagFit® model offers the best generic performances for the MCC and AUC metrics and the lowest computation time (and hence de lowest eqCO2).
- Deep Learning models (Auto-Encoder and MTAVAE) offer interesting results on Private #2 and #3 datasets, for an expensive time cost.
- The mono-variated models (IF, OCSVM, LOF) are clearly less performant than the others.
- Even though the deep learning methods are the most time-consuming during the training phase, they are not necessarily the largest emitter of CO2 since they are much more efficient in the inference phase, once the model is learnt.

Conclusion

Despite the no free lunch theorem [16], the industrial context requires genericity, efficiency, and explicability. DiagFit® combines these three advantages compared to state-of-the art algorithms.
A perspective would be optimizing the computation speed by using specific hardware such as GPU.

References

(1) S. Le Gall and T. Le Magueresse, "Comparison of failure prediction models – Application to cyclic datasets," White paper, 2021.

(2) M. Breuniq, "LOF: identifying density-based local outliers," ACM sigmod record 29, p. 93–104, 2000.

(3) Z. Ding and M. Fei, "An Anomaly Detection Approach Based on Isolation Forest Algorithm for Streaming Data using Sliding Window," IFAC Proc., p. 12–17, 2013.

(4) F. Liu, K. Ting and Z. Zhou, "Isolation forest," In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, p. 413–422, 2008.

(5) B. Schölkopf, "Support vector method for novelty detection," In Proceedings of the 12th International Conference on Neural Information Processing Systems, p. 582–588, 1999.

(6) D. Park, "A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-based Variational Autoencoder," IEEE Robotics and Automation, 2018.

(7) H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, B. Xu, J. Bai, J. Tong and Q. Zhang, "Multivariate Time-series Anomaly Detection via Graph Attention Network," arXiv, 2020.

(8) M. Christ, N. Braun, J. Neuffer and K.-L. A.W., "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)," Neurocomputing 307, pp. 72-77, 2018.

(9) u. Kozitsin, D. Katser and V. O., "Skoltech Anomaly Benchmark (SKAB).," Kaggle, 2020.

(10) "Kaggle," (Online). Available: https://www.kaggle.com/datasets/nphantawee/pump-sensor-data.

(11) C. Rieth and e. al., "Issues and Advances in Anomaly Detection Evaluation for Joint Human-Automated Systems," Applied Human Factors and Ergonomics, 2017.

(12) "Itrust," (Online). Available: https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/.

(13) C. Lessmeier, J. K. Kimotho, D. Zimmer and W. Sextro, "Condition Monitoring of Bearing Damage in Electromechanical Drive," European conference of the pronistics and health management society, 2016
.

(14) M. Zmitri, "Evaluation des performances d'un détecteur de pannes," White paper, 2022.

(15) "Code Carbon package," V.2.1.4, (Online). Available: https://codecarbon.io. (Accessed February 2023).

(16) D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization," IEEE Transactions on evolutionary computation, Vols. Vol. 1, Issue 1, 1997.