Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P4996-cd

Condition monitoring approach based on unsupervised anomaly detection for pumps regulating groundwater level under a coastal infrastructure

Arto Niemi, Felipe A. Costa de Oliveira, Georges Anicet Kuete Kouam, Jose Luis Quinones Gonzalez, and Bartosz Skobiej

DLR Institute for the Protection of Maritime Infrastructures, Bremerhaven, Germany.
E-mail: arto.niemi@dlr.de

Habbo Cramer

Free Hanseatic City of Bremen, Office of the Senator for Economic Affairs, Ports and Transformation, Referat 31, Bremen, Germany

Peter Kara and Frieda Schneider

bremenports GmbH & Co. KG, Bremerhaven, Germany

This paper describes a condition monitoring approach for pumps regulating groundwater level under a port infrastructure. We focus on the Bremerhaven container terminal located in northwest Germany at the mouth of the river Weser. Our aim was to construct a strategy to detect potential pump failure indications that could inform conditional maintenance actions. Two signals were available for us: the groundwater level, measured with a radar, and the binary pump on/off operation signal. For this purpose, we tested four unsupervised machine learning-based anomaly detection algorithms, in combination with multiple post-processing methods for anomaly scoring and thresholding. Additionally, we developed a model to simulate the groundwater level signal, enabling the test of failure modes that were not present in measured data. We found that the appropriate selection of model and post-processing method was critical for obtaining satisfactory results in both measured and simulated signals.

Keywords: Condition based monitoring, anomaly detection, deep learning.

1. Introduction

Reliability Centered Maintenance (RCM) revolutionized maintenance in terms of increasing expectations on equipment reliability and availability, Moubray (1997). These advances were mainly a result of condition monitoring, which allowed maintainers to carry out maintenance actions before a failure would occur. Thus, the full consequences of a failure, costs or hazards, could be mitigated with advance maintenance. The approach has been standardized and implemented by many organizations. For example, NASA has a guide on how to apply RCM to facilities and collateral equipment, NASA (2008). The investment in maintenance has resulted in significant savings. The guide also introduces an intuitive approach, which identifies and implements obvious, condition-based tasks with minimal analysis. Our work can be seen to follow this principle.

The challenge is to detect critical degradation of the system before a failure. A typical approach involves a degradation model that would inform the maintenance actions. Building such a model requires an understanding of the degradation processes and having measurements for estimating that degradation. Without these measurements, this approach is impossible. Therefore, we investigate an alternative strategy that uses time-series unsupervised anomaly detection models to inform condition based maintenance actions.

We present a case study scenario for pumps regulating groundwater level (GWL) under a port infrastructure. The focus is on the Bremerhaven container terminal which is located in northwest Germany at the mouth of the Weser River. This location creates a complication as the water level in the Weser River is affected by the North Sea through tides and wind stress. The GWL must be kept in a certain range in accordance to the river

water level. In this port, the GWL is regulated by two pairs of drainage pumps, where the paired pumps are operated asynchronously. This setup allows comparing the pumps against each other, which improves the detectability of degradation. A degrading pump performance can be caused by clogging of pipes with ground sediment, or by mechanical wear of the pump. However, without direct measurements of the drainage pipes condition and the pump electrical current, the detection of those degrading states is difficult.

The drainage system performance is monitored through two signals: the GWL, measured with a radar, and the binary pump on/off operation signal. Both signals are sampled at regular one minute intervals. From the pump operation, we derive the pumping time duration, and from the GWL measurement, we can estimate the pumping rate. GWL measurements along the berth can further be used for detecting the location of clogging. A physics-based model could be constructed for these purposes, but setting the boundary conditions and model parameters would be challenging and time consuming. In this project, we opted for a machine learning (ML) based approach. We had access to real data to train predictive models. We further supplemented the dataset with simulated GWL signals that resembled degradation scenarios that were not present in the real data. In that way, we can further test the capabilities of our models without being limited to the measured data.

2. Background

2.1. *Groundwater management and drainage pump reliability*

Groundwater management has several important applications. In coastal areas, fresh groundwater can be used as drinking water or for irrigation, which can lead to seawater intrusion and may cause land subsidence Hussain et al. (2019). The presence of groundwater affects slope stability for example excessive rainfall can trigger landslides Sapari et al. (2008). Reducing the amount of groundwater is also required for mine dewatering, Emal Qazizada and Pivarčiová (2018). In the port of Bremerhaven, the GWL must kept in a certain range in accordance to the river water level through pumps.

The OREDA handbook detailing reliability data from the offshore oil and gas industries contains information on pumps, SINTEF and NTNU (2015). The handbook describes that the critical faults of a pump include abnormal instrument readings, pump breakdown, failure to start, abnormal output, leakage, noise, overheating, and choking. Several works have addressed condition monitoring for pumps. Ahonen et al. (2012) monitored pump power consumption to estimate motor shaft power and the pump flow rate through linear interpolation and third-order polynomial function approximations. Turkeri and Kiselychnyk (2024) derived pressure and flow rate estimations with artificial neural networks using stator current, estimated input active power, and reference stator voltage frequency as inputs. While Bohn et al. (2019) proposed a ”practical approach” based on monitoring thermodynamic efficiency, vibrations, and dynamic fluid pressure. The lack of these monitoring data lead us to consider ML-based approaches.

2.2. *Anomaly Detection Overview*

The problem of anomaly detection (AD) has been investigated across a wide range of domains, with relevant applications in cybersecurity, finances, telecommunications, computer vision, medicine, astronomy, and others, Ruff et al. (2021). AD is well suited for early fault and damage detection of engineering equipment and structures, Xu and Saleh (2021). It is intrinsically related to sensor data. For industrial applications, these data typically come in a streaming fashion. While AD exists in supervised and semi-supervised mode, it is widely used in the unsupervised mode. The reason is that unlabeled data are often broadly available and labeled data are expensive to obtain and rare. In reliability and safety applications, data are usually available for (labeled) nominal operational conditions only.

The basic idea is to detect the data points, or sequences of points, that deviate considerably from an expected normality. The general strategy follows these steps:

(i) training a model that learns the distribution of the normal scenario;
(ii) make predictions with that model and compare them with the new data;
(iii) if there is a significant deviation between predicted and observed values, those values are flagged as anomalous.

In some cases, that difference is used to derive an anomaly score that represents the degree to which a point, or sequence of points, is anomalous. Various methods exist for computing that score, and that choice can significantly impact the detection performance. The simplest method is the computation of the point-wise absolute error difference between the observed and predicted signal. A smoothing algorithm, such as an exponentially weighted moving average (EWMA), might be employed next to mitigate false alarms caused by noisy signals. Another option, valid for time-series data and used for speech recognition tasks, is dynamic time warping (DTW), Berndt and Clifford (1994). It computes the similarity between two time signals that can vary in speed or be out-of-phase. Next, a method to find suitable threshold values for the anomaly score is used to determine the boundary between normal and anomalous data. Non-parametric dynamic threshold (NDT), Hundman et al. (2018), and peak over threshold (POT), Siffer et al. (2017), are relevant strategies for that task. In this work, we compare the results of each tested model using these different post-processing strategies.

3. Methodology

3.1. *Problem Formulation*

An optimal condition monitoring approach depends on the level of knowledge on the application, Baur et al. (2020). If the application is well known, a rules-based approach or physics model can be implemented. When this is not the case, a data-driven is better. Our challenge was the limited operational information available for monitoring the pump performance. There are two pumping stations which both have a pair of of pumps. In this work, we focus on the signals of one of those stations. The pump activation times from each pump are stored as well as GWL measurements from both stations. This setup allows monitoring pumps individually and pairwise comparison of pumps located in a station. The problem is that these measurements relate to the performance of the pumping system, including the drainage pipes, but not directly to a pump.

The pumping efficiency can be indirectly assessed from the GWL signal in combination with the computed on cycle duration (the timestamp interval that a pump operated continuously). Since the pumps are programmed to start and stop when specific water levels are reached, the duration of a pumping cycle should be proportional to the pumping rate. Fig. 1 depicts the relationship between the pump cycles and the GWL signal. Historical data for the pumps on cycle duration show that there is a normality range for those times, but on some rare occasions, anomalous values exist. Fig. 2 shows a boxplot on the variation of the on cycle duration during several months of the year 2020. Unexpectedly long cycle durations occur in May.

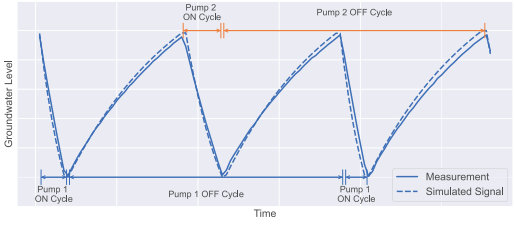

Fig. 1. Influence of pump operations on the GWL measurements. Comparison of a measured signal with a simulated one.

These long cycles can occur due to: a communication failure in the pump activation signal; heavy rainfall; degradation of the pump; or a clogging of the drainage pipes. Using only the monitored signals, it is difficult to identify the cause of such anomalous events. Yet, the automatic detection of these anomalies can initiate failure diagnosis leading to other maintenance actions. In this work, we investigate the feasibility of using unsupervised AD models to initiate this type of maintenance actions, using the GWL signal.

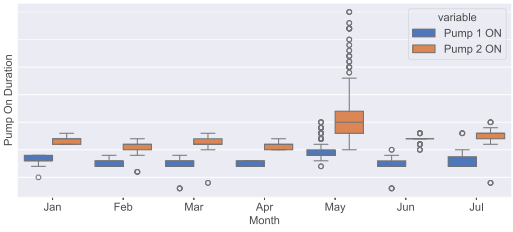

Fig. 2. Boxplot with the monthly aggregation of the observed pump on cycle duration. Values have been obscured due to data privacy reasons.

3.2. *Selected Anomaly Detection Algorithms*

There are many AD models with reported good performance in benchmark datasets. We selected four of these models based on popular neural network architectures that had open source code available. The selection contains a mix of prediction and reconstruction based algorithms. A brief overview of these models is provided next.

Auto-Encoder with Regressor (AER) is based on an encoder-decoder architecture implemented by biLSTM layers, Wong et al. (2022). It combines reconstruction and prediction based errors in its loss function to leverage the advantages of both strategies. Wong et al. (2022) propose a bi-directional anomaly score to overcome the limitation of prediction methods at the start of the timestamp sequences, with a combination of the forward and reverse direction predictions.

LSTM-NDT was proposed by Hundman et al. (2018). It consists of a Long-Short-Term-Memory with non-parametric dynamic thresholding. The recurrent neural network (RNN) architecture has memory-like capabilities enabling the model to learn long-term dependencies in the time-series. Hundman et al. (2018) propose an unsupervised dynamic thresholding method to find the anomalous sequences in the residuals (error) from the model predictions.

TadGAN proposed by Geiger et al. (2020) uses generative adversarial networks (GANs) based model implemented with LSTM layers for the generator and critics module. It introduces a cycle-consistent GAN architecture to enable time-series mapping. Various methods to compute the reconstruction error in combination with the critics output are proposed along with a benchmarking system for AD evaluation.

TranAD is a transformer based AD model capable of handling multivariate time-series inputs, Tuli et al. (2022). It uses focus score based self-conditioning to enable robust multi-modal feature extraction and adversarial training to gain stability.

3.3. *Groundwater Level Signal Simulation*

Due to the lack of representative examples of anomalies in the real data on a slow deterioration of the drainage system, we will simulate added signals to mimic that behavior. Accordingly, we can test the selected AD models on different types and intensities of disturbed signals. Thus, evaluating their effectiveness and sensitivity to specific conditions.

The GWL was modeled as a quasi-triangular periodic signal with alternating curves for the up and downward slopes. As depicted in Fig. 1, the GWL rise time is given by the pump off cycle duration, and the down slope is affected by the pump on cycle duration. Since the pumps in a station always operate in alternation, the consecutive off or on related slopes can differ. Although the two pumps at each station are from the same manufacturer and model, in the collected measurements spanning eight years, the pumping times are always different. The historical data was used to model the normality behavior for each pump considering that difference. The parameters for the simulation model were:

- The on and off pumping cycle duration for each pump ($t_{P1_{\text{on}}}, t_{P1_{\text{off}}}, t_{P2_{\text{on}}}, t_{P2_{\text{off}}}$);
- the upper and lower GWL limits, modeled as Gaussian with small variance ($L_{\text{high}} \sim \mathcal{N}(\mu_{\text{high}}, \sigma^2)$, $L_{\text{low}} \sim \mathcal{N}(\mu_{\text{low}}, \sigma^2)$);
- the shape of the downward and upward slopes, given by a scale parameter in the related exponential curve (s_{off} and s_{on}).

In this way, the GWL was simulated as a time-series signal according to the following equations:

$$G_{\text{WL}}(t) = \begin{cases} L_{\text{low}} + R_{\text{WL}} f_{\text{off}}(t), & \text{off cycle} \\ L_{\text{high}} - R_{\text{WL}} f_{\text{on}}(t), & \text{on cycle} \end{cases}$$

where,

$$R_{\text{WL}} = L_{\text{high}} - L_{\text{low}}, \tag{1}$$

$$f_{\text{off}}(t) = 1 - S(e^{\frac{-s_{\text{off}}}{t_{Pi_{\text{off}}}} t}), \tag{2}$$

$$f_{\text{on}}(t) = 1 - S(e^{\frac{-s_{\text{on}}}{t_{Pi_{\text{on}}}} t}), \tag{3}$$

and $S()$ is a min-max scaler to ensure that the exponential term is within the 0 to 1 range.

The data analysis indicates that the pump 2 operation time has significantly higher discrepancies than the pump 1. There are multiple instances of step-like increases on the pump 2 on cycle duration, with hardly any change in the pump 1. In some cases, the cycle time for both pumps increases, but the increase is disproportionately higher in the pump 2. There are also instances in which the increased on or off cycle duration results in lower and higher limits for the water level. That might be due to a fault communication in the control signal for activating or deactivating the pumps. This type of event will be simulated as well.

The historical data were used as a reference for setting the simulation parameters for the normality case. For the anomalous samples, the extreme values of the pump operation times were used as references to determine the ranges for low, moderate, and severe anomalies. For each intensity, two types of disturbances were simulated: step-like, with sudden change to the operation time parameters; and ramp-like, with a progressive change. In all cases, the signal returns to normality a certain time after the peak anomalous values are reached. Additionally, samples with and without changes in the GWL high and low limits were created. Four samples with slightly different values were generated for each configuration, totaling 96 test samples. Fig. 3 shows two examples of these configurations.

4. Results

We selected a 50k points sample of the measured GWL signal for training the AD models. The sample is representative of normal operation, without any anomalies. Pre-processing steps such as normalization and conversion to time-series windows of fixed size (100 points) were conducted to condition the data according to the input specification of the models. The window size was chosen as a reasonable time interval that could provide enough context for the detection of anomalies. For reference, the GWL signal in Fig. 1 has 120 points. Most of the other hyperparameters were kept as their model-specific default values, or with the values reported in their respective papers. The exception was the number of epochs, which was set as ten for all models, with an early stop in case of validation loss increase. No grid-search or hyperparameter optimization was done, as we were interested in evaluating the ease of implementation of this data-driven approach.

The trained models were first tested in the nominal case (anomaly free data), checking if the predicted (or reconstructed) signal could match the original measurements. As presented in Fig. 4, all four models were able to reproduce the GWL signal under the normality setting. In some cases, the predicted signal would get out-of-phase with the original, resulting in spikes in the point-wise absolute difference between them. Those could be falsely flagged as anomalous sequences. This issue stressed the importance of the appropriate selection and fine-tuning of the post-processing steps. We observed that an additional smoothing step in the error computation greatly reduced the amount of these false positives.

The models were then tested with 14 selected anomalous samples from the real measurements, and 96 simulated samples. Various metrics exist for evaluating an AD model performance[a], and the results can be highly dependent on that metric choice. Our objective is the identification of anomalous measurements to inform maintenance actions. If the predicted anomalous sequence has

[a] We recommend the survey Correia et al. (2024).

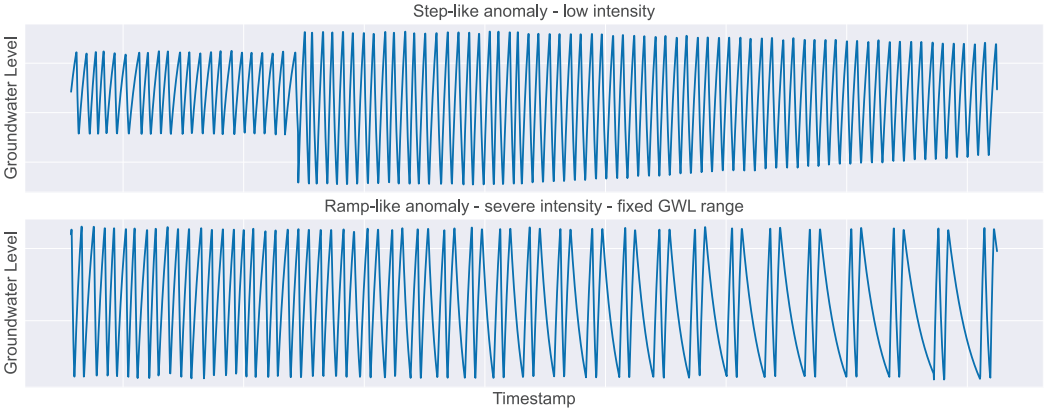

Fig. 3. Examples of simulated test samples with different anomaly configurations.

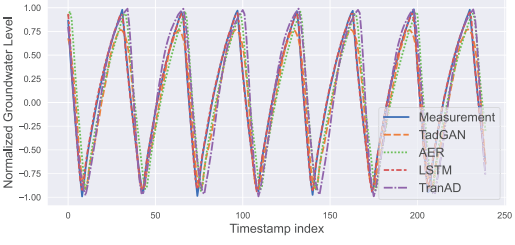

Fig. 4. Sample of reconstructed GWL signals, in comparison with the original measurements.

any overlap with the known ground truth anomalous sequence, it is computed as a true positive. If the prediction does not overlap the ground truth at any point, it is counted as a false positive. Finally, if it fails to detect any part of the real anomalous sequence, a false negative is computed. In this way, the unweighted contextual F1 score (Wong et al. (2022), Hundman et al. (2018)) was computed for each model, in each configuration and for each test sample. The overall results are summarized in Table 1.

For our dataset, the LSTM had the best overall performance. Surprisingly, the non-parametric dynamic threshold method, the recommended post-processing step in that model paper, was outperformed by the peak over threshold. POT yielded better results for all tested models and configurations. The only exception was in the results for

Table 1. Summary of the overall performance of the different models and configurations in test samples.

Model	Anomaly Score Method	Threshold Method	Avg. F1 Score
AER	Abs. error	NDT	0.54
		POT	0.69
	DTW	NDT	0.65
		POT	**0.98**
LSTM	**Abs. error**	NDT	0.60
		POT	**0.99**
	DTW	NDT	0.64
		POT	0.98
TranAD	Abs. error	NDT	0.51
		POT	0.66
	DTW	NDT	0.61
		POT	**0.72**
TadGAN	**Abs. error**	NDT	0.52
		POT	**0.81**
	DTW	NDT	0.57
		POT	0.67

the 14 real data test samples, for which the NDT method for the TadGAN model performed better than POT (with a score of 0.93 against 0.86). The overall performance with the real data was slightly worse than with the simulated samples. However, the combination of LSTM with a smoothed absolute error based anomaly score and POT defined threshold, also achieved the best F1 score (of 0.95,

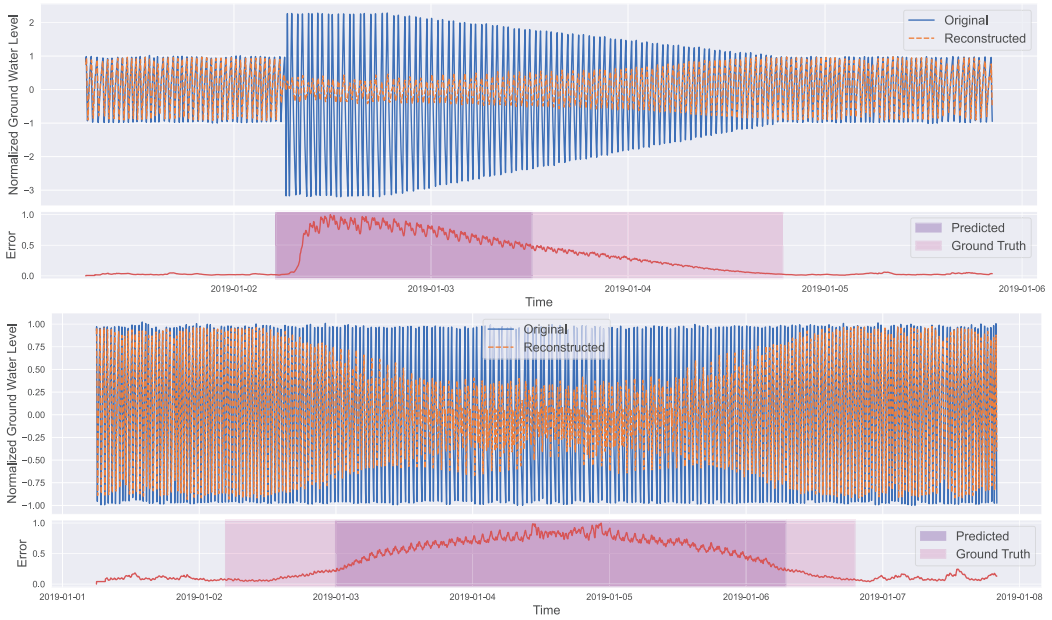

Fig. 5. On the top, simulated moderate step anomaly, and on the bottom, simulated moderate ramp anomaly. Both, signals have been reconstructed with an AER model. The difference between original and reconstructed signals causes these incidents to be flagged as anomalies.

against 0.93, 0.92 and 0.73 for TadGAN, AER-DTW-POT and TranAD-DTW-POT respectively).

With the selection of the best post-processing steps for each model, the contextual F1 score for the simulated samples was practically perfect. Both AER and LSTM models were able to detect all step and ramp type anomalies. Therefore, there was no observable difference in the detectability of these two types of anomalies across the three severity categories. The plots in Fig. 5 show successful results of the AER model in simulated samples.

5. Discussion and Conclusions

We have shown that with the appropriate data conditioning, selection of models and post-processing steps, an unsupervised ML-based AD strategy can yield good results. However, it is important to recognize the limitations of such a strategy. Multiple combinations of models and post-processing methods were tested before reaching a satisfactory result for our dataset. That process can be time consuming and computationally expensive. Even after fine-tuning a viable model and pipeline, it might not be robust enough to withstand the changing conditions in a dynamic environment. Further tests are required to make that assessment.

Another important consideration is the choice of the evaluation metric for AD results. Using an unweighted contextual F1 score can give overly optimistic results. If the application requires accurate temporal identification of the anomalous sequence range, that metric is not valid. As shown in Fig. 5, the predicted anomaly range overlaps with the ground truth label, resulting, for that particular sample, in a perfect contextual F1 score of 1. However, when considering a metric that penalizes the difference between predicted and ground truth ranges, the results are much worst[b].

For our purposes, the AD results provide a first step towards an informed condition based maintenance plan for the pumps and drainage pipes. This plan depends on the timely identification of

[b]The F1 score accounting for that difference, for the best performing model configuration, drops from 0.99 to 0.81.

system degradation. The obtained results show its feasibility. This approach can be further supported by a physics, or knowledge-based rules informed models; or by a supervised ML approach, in which the simulated GWL signal can be used for training. We will conduct these investigations in future works.

References

Ahonen, T. et al. (2012). Centrifugal pump operation monitoring with motor phase current measurement. *Int. J. Electr. Power Energy Syst. 42*, 188–195.

Baur, M. et al. (2020). A review of prognostics and health management of machine tools. *Int. J. Adv. Manuf. Technol. 107*, 2843–2863.

Berndt, D. J. and J. Clifford (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pp. 359–370. AAAI Press.

Bohn, B. et al. (2019). Sensing concept for practical performance-monitoring of centrifugal pumps. In *2019 IEEE Sensors*.

Correia, L. et al. (2024). Online model-based anomaly detection in multivariate time series: Taxonomy, survey, research challenges and future directions. *Eng. Appl. Artif. Intell. 138*, 109323.

Emal Qazizada, M. and E. Pivarčiová (2018). Reliability of parallel and serial centrifugal pumps for dewatering in mining process. *Acta Montan. Slovaca 23*, 141–152.

Geiger, A. et al. (2020). TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks . In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 33–43. IEEE Computer Society.

Hundman, K. et al. (2018). Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 387–395. Association for Computing Machinery.

Hussain, M. S. et al. (2019). Management of seawater intrusion in coastal aquifers: A review. *Water 11*, 2467.

Moubray, J. (1997). *Reliability-centered Maintenance* (Second ed.). Industrial Press Inc.

NASA (2008). *Reliability-centered maintenance guide for facilities and collateral equipment*. National Aeronautics and Space Administration.

Ruff, L. et al. (2021). A unifying review of deep and shallow anomaly detection. *Proc. IEEE 109*(5), 756–795.

Sapari, N. et al. (2008). The influence of rising groundwater on slope stability and engineering properties of soil. In *Proceedings of EnCon2008*.

Siffer, A. et al. (2017). Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1067–1075. Association for Computing Machinery.

SINTEF and NTNU (2015). *Offshore and onshore reliability data* (6th ed.), Volume 1 - Topside equipment. OREDA Participants.

Tuli, S. et al. (2022). TranAD: deep transformer networks for anomaly detection in multivariate time series data. *Proc. VLDB Endow. 15*, 1201–1214.

Turkeri, C. and O. Kiselychnyk (2024). Improved flow rate and pressure ANN estimators for a centrifugal fan with an induction motor drive. *J. Energy Syst. 8*, 130–142.

Wong, L. et al. (2022). AER: Auto-encoder with regression for time series anomaly detection. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1152–1161. IEEE Computer Society.

Xu, Z. and J. H. Saleh (2021). Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliab. Eng. Syst. Safe. 211*, 107530.