Health Assessment of Pump Stations using Time Series Anomaly Detection: Deploying AI on the Industrial Edge

Abhishek Murthy[1], Babak Afshin-Pour, Willem Malloy and Vasileios Geroulas
Schneider Electric
[1]*abhishek.murthy@se.com*

Supervisory Control and Data Acquisition (SCADA) is widely used to manually monitor and manage distributed physical assets. Supporting infrastructure was designed and optimized for that need. Specialized communication protocols are utilized for applications which span large geographical deployments. These protocols ensure data robustness and consistency in variable-quality network environments. However, the resulting data, while forming enterprise data pipelines, lacks granularity and has irregular time spacing, making it unsuitable for machine-learning-based predictive maintenance applications.

We present a hybrid cloud-to-edge health monitoring solution for assets connected to SCADA or other legacy control systems. Our solution uses a modbus-based polling system on the edge, to collect data at a much higher granularity than the adjacent SCADA system, letting us detect even subtle and acute patterns in the data. Note that no new sensors are needed, as we connect to the same registers as the existing SCADA system. The high granularity data is assessed at the edge for anomalies, using time series anomaly detection algorithms. We then synthesize the predictions into a health index that quantifies the recency and the frequency of the detected anomalies for the asset. The health index is then transmitted to a web-based application, where the user can configure thresholds for generating alerts based on the criticality of the asset.

We demonstrate our solution in a case study, where the application was deployed using Schneider Electric's Customer First Digital Hub, to monitor a sewage pump station for blockages and other subtle deviations in operating patterns.

1. INTRODUCTION

Wastewater management involves a network of interconnected assets whose purpose is to reliably move wastewater from homes and workplaces back to centralized processing facilities. Failures in these operational assets must be very rapidly addressed to ensure the continuity of critical services to our homes and cities. There are several failure modes in the complex industrial assets. Blockages in pipes may lead to overflows, as there is no redundancy in the infrastructure. As a critical active component of wastewater networks, motors and pumps keep wastewater flowing. Pumps may be installed as redundant pairs, or as single units. Their failure can lead to significant financial costs related to repair and restoration. The regulatory climate for compliance and safety has also motivated the various stakeholders to maintain wastewater treatment facilities with the latest technology and overall effectiveness. The community has conventionally relied on scheduled or surveillance-based maintenance. Such traditional approaches are often more expensive and may not even enable timely interventions. To this end, *the wastewater treatment industry has untapped potential to benefit significantly from machine-learning-based predictive maintenance.*

Wastewater is hazardous prior to treatment. Any untreated discharges, caused by failure events that are not addresses in time, can harm the environment and the public health. On a human and cost front, failures can occur unpredictably, making it difficult for organizations to plan effectively. On a human level, people must respond at any hour of the day. They are often recompensed via overtime, which ultimately the users of the system pay for.

Predicting failure generally relies on proactive identification of leading indicators. In complex systems, those leading indicators are not easily identifiable either to human operators or to the simplistic SCADA alarming mechanisms. Each asset produces a constant stream of data from interdependent variables read from multiple sensors. This inherent complexity does not lend itself to automated machine learning-based modeling and processing.

Moreover, wastewater treatment facilities present some unique challenges for predictive maintenance. The makeup of the waste water changes constantly with time. As the assets age, the types of failures also change. The existing SCADA systems that are used for monitoring them are woefully ill-equipped for supporting advanced predictive maintenance

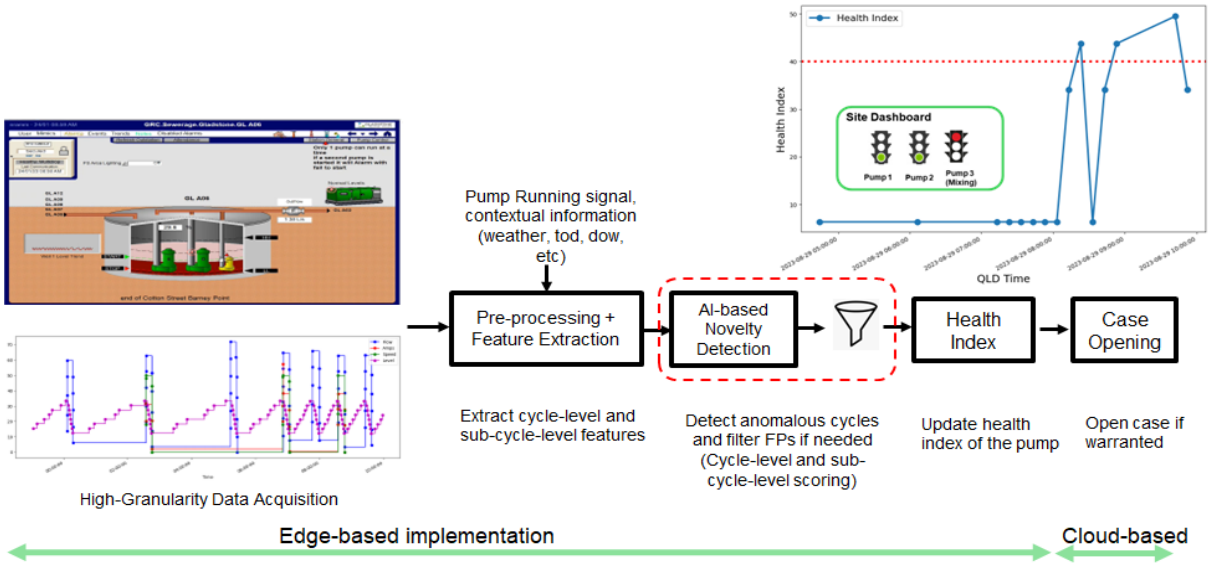


Figure 1. The proposed hybrid cloud-to-edge pump monitoring system.

algorithms. Scalability is also important: the solution must be widely applicable and support diverse modes of operation.

Given the limitations of the existing SCADA systems in enabling predictive maintenance of wastewater systems, there is a need to go beyond the current monitoring and control infrastructure. To this end, *we present a hybrid cloud-to-edge monitoring system to enable predictive maintenance of industrial assets by processing high-frequency multidimensional sensor data at the edge.* Our system trains machine learning models on the cloud and executes them on the edge to produce an actionable health index. A threshold can be set based on the criticality of the asset to open a case when the health index crosses it.

Our system has two noteworthy features: 1) *high-speed polling* over modbus to collect multivariate sensor data and 2) *time series anomaly detection and health index calculation*. The sensors that are used by the existing SCADA system are polled over modbus to get high-frequency data onto an Industrial PC (IPC). ML models are trained on such data in the cloud and deployed on the IPC at the edge. The ML models detect anomalous patterns in the working on the assets and convert the predictions to a health index. The health index quantifies the recency and frequency of the anomalous behavior.

We demonstrated our system at a wastewater treatment facility and showed that we can accurately monitor the health of the pumps that are used in the tanks. Blockage in these pumps is a common problem; a controlled experiment was conducted at the wastewater treatment plant to mimic blockages. The health index and the alarming logic performed accurately and no false alarms were raised.

The rest of the paper is organized as follows. We compare our work with relevant literature in Sec. 2 and provide background for understanding wastewater treatment facilities in Sec. 3. The architecture of the proposed system is presented in Sec. 4. The time series anomaly detection and the health index calculations are detailed in Sec. 5. We present the controlled blockage experiment and other results in Sec. 6. We conclude with directions for future work in Sec. 7.

2. RELATED WORK

The value of SCADA and related data for predictive maintenance of industrial assets is well-recognized by the PHM community. In this section, we focus on wastewater treatment facilities and SCADA and ML-based approaches and provide context for our work.

The authors recognize the opportunities of using SCADA data for machine learning applications in (Šenk, Tegeltija, & Tarjan, 2024). Applications like anomaly detection, predictive maintenance, and system performance optimization are explored. Additionally, the authors acknowledge data quality, security, interpretability as potential challenges. Our work presents a specific realization of these ideas for wastewater treatment facilities.

In (Trstenjak, Palasek, & Trstenjak, 2019), the authors introduce a novel decision support system to forecast wastewater pumping station failures using a Case Based Reasoning (CBR) classification method with a continuous learning algorithm. Our proposed system goes beyond decision support and analyzes high-granularity time series data at the edge to provide near real-time insights about ongoing issues in the pumps. In (Moreno-Rodenas, Duinmeijer, & Clemens,

2021), a computer-vision-based approach is presented to detect the accumulation of fat, oil and grease in the sumps of waster water pumping station. Such approaches do not scale very well and also need the installation of dedicated cameras in potentially hazardous conditions. Our system, on the other hand, uses sensors that the adjacent SCADA system already uses for its operation.

In (Mosallam, Medjaher, & Zerhouni, 2013), the authors present a non-parametric trend modeling approach based on multidimensional sensor data for PHM of industrial assets. Their approach entails novelty detection models trained on nominal data. Our approach is similar, but is focused on wastewater systems and combines contextual parameters, such as the weather and time of day in the anomaly detection algorithm.

Machine learning-based approaches have found success for monitoring the health of pumps in other domains, such as oil and gas and pipelines. In (Concetti, Mazzuto, Ciarapica, & Bevilacqua, 2023), the authors proposed an unsupervised anomaly detection for oil and gas sector. They used a self-organizing map algorithm to identify anomaly from different modes of normal operation. Using an unsupervised method based on a Gaussian mixture model, authors in (Giro, Bernasconi, Giunta, & Cesari, 2021) tracked the normal condition of a centrifugal pump using pressure measured on remote points along the pipeline. Using a unlabeled data, the authors were able to identify pump failure, thereby extending the lifetimes of the pipelines.

3. BACKGROUND

A wastewater treatment facility consists of a network of tanks through which the water is circulated. Each tank consists one or more pumps that work together to circulate the water that is sent from upstream tanks into the downstream stations. Fig. 1 shows one such tank on the top-left. The pumps begin pumping the water when it reaches a high watermark. As the pumps kick-in, the water level gradually decreases. They stop when the water reaches a low watermark.

Conventionally, two or more pumps work in duty cycle fashion, where they take turns in pumping the water out. *Each iteration of the pump turning on, reducing the tank level, and turning off is called a pump cycle.*

In the context of pump stations within the water and wastewater industry, SCADA systems facilitate a lot of critical functions aimed at optimizing operational efficiency and ensuring regulatory compliance. These functions encompass real-time monitoring of water levels, flow rates, pressure levels, pump statuses, and other relevant parameters. By continuously collecting and analyzing data from various sensors and instruments installed throughout the pump stations , SCADA systems enable operators to gain insights into system performance, identify potential issues, and initiate corrective actions proactively. Moreover, SCADA systems facilitate remote control capabilities, allowing operators to adjust pump settings, alter flow rates, and implement emergency shutdown procedures from centralized control centers or mobile devices, thereby minimizing downtime, reducing energy consumption, and mitigating risks associated with equipment failures or process disruptions. A typical sequence of pump cycle data collected by the SCADA system is shown in Fig .2. We show the flowrate (blue), motor amps (red), motor speed

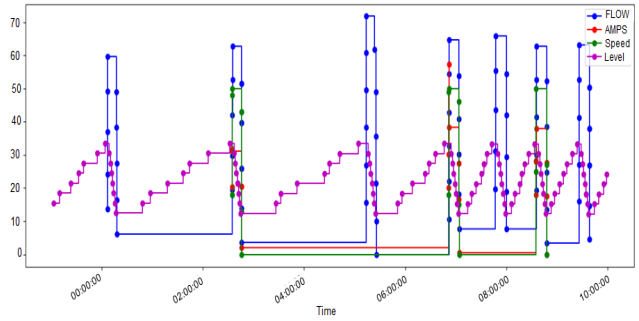


Figure 2. Typical sequence of pump cycle data collected by SCADA.

(green) and the water level (purple) collected by the SCADA system across multiple cycles. Each cycle consists of the water level decreasing as the motor amps and the motor speed attain their nominal values. The flowrate out of the tank also attains its maximum value during the cycle.

SCADA systems originated to enable remote monitoring and control of industrial processes in sectors like energy, water, and manufacturing. Designed for real-time data acquisition over often extensive networks, SCADA systems can handle vast numbers of data points, or "tags" which can reach into the millions in large-scale applications. At the time of their inception, traditional server-based computing was the norm. Performance and capacity were limited by the computing power at their disposal and as a result, trade-offs were made in data polling frequency, system architecture, and prioritization by asset operational criticality to optimize performance within bandwidth and resource constraints.

As industrial operations expanded geographically, traditional scan-based SCADA protocols struggled with latency, reliability, and security over long distances, at the time often underpinned by technologies including serial, modem, microwave, or radio transmission. This prompted the development of specialized communication protocols for wide-area applications. Notably, the Distributed Network Protocol version 3 (DNP3) was created to address these challenges with new features like report-by-exception, timestamping at source and adding data buffering on-device. These features made it well-suited for large-scale, distributed systems. It subsequently enjoyed broad uptake in particular segments and regions which had

less reliable and performant network communication, while needing to continuously manage vast amounts of data.

While SCADA systems are well-evolved for their primary purpose, the nature of the data they collect is often of questionable quality for modern organizational data-pipelines upon which higher level AI capabilities can be built. SCADA data is very coarse, as depicted in Fig. 2. The low granularity is a result of the deadbands configured in the sampling of the sensors. Our proposed system, which is described in the next section overcomes the granularity issue by collecting and processing the data on the edge.

4. SYSTEM ARCHITECTURE

The solution is developed and deployed on Schneider Electrics Customer First Digital Hub framework, within the Industrial Automation Services. This is an open, integrated infrastructure based on industry standards, leveraging the capabilities of Cloud and Edge computing combined with Artificial Intelligence. Key elements of the solution include:

- Edge Gateways Deployment: Edge gateways are installed in the field, interfacing with physical assets and control systems like PLCs. These gateways use standard industrial protocols such as Modbus and OPC UA to facilitate seamless bidirectional communication.
- Machine Learning Integration: Data collected from the field is used to train Machine Learning (ML) models on the cloud, harnessing extensive computational resources. These models are then deployed on the edge, optimizing for latency, connectivity, bandwidth, and storage efficiency.
- Centralized Management: The edge gateways are connected to the internet, allowing centralized management of all edge nodes and ML models. A web-based user interface provides comprehensive oversight, enabling users to monitor real-time inference based on high granularity data, view alarms, and track low granularity data uploaded to the cloud. Users can also provide feedback on ML model accuracy, fostering continuous learning and improvement.

This methodology allows seamless integration with existing control infrastructures, such as legacy SCADA systems, enhancing their capabilities with advanced AI models. Essentially, it is like deploying an "engineer in a box" to valuable assets, enabling the system to: Detect and alarm on abnormal operating events. Capture and automate operator expertise and knowledge through Machine Learning. Deploy workflows to mitigate or prevent issues.

Fig. 3 illustrates the high-level architecture of this solution.

5. TIME SERIES ANOMALY DETECTION AND HEALTH INDEXING

In this section, we describe the anomaly detection pipeline for the sewage pumps. We begin by framing the anomaly detection problem and outlining alarming logic.

As described in Section 3, the pumps in the sewage station operate in cycles. Additionally, the sewage treatment facility experiences different demand patterns through the day. These patterns are related to the periodic activities of the residents of the area. Mornings and evenings, when most people tend to be at home, result in heavier inflow into the sewage stations compared to the afternoon and late night hours. Fig. 4 shows the distribution of length of the cycles for each hour of the day.

In this context, we define the anomaly detection problem for sewage pumps as follows. *During the peak hours, following the onset of a cycle, classify windows of predefined length in time as either anomalous or nominal. During off-peak hours, at the end of a cycle, classify it as either anomalous or nominal.* Note that predictions are made within a cycle during peak hours, whereas for off-peak hours, the entire cycle is labeled as nominal or anomalous.

The alarming logic uses a health index that quantifies the recency and frequency of the anomalies. A pre-defined threshold is created for each asset. In Fig. 1, a threshold of 40 has been chosen. When the health index crosses this threshold, a notification is issued. The health index works as follows.

1. The health index is nominally close to 0. A low non-0 value is used as the floor. This indicates that no anomalies were seen recently, and the asset is working nominally.
2. During the high inflow periods, each 10-minute window of a pump cycle is scored. If the window is deemed nominal, then the health index does not change if it was already at the floor. If it was elevated, the health index is decreased at a rate proportional to its level. If the health index was very high, then the decrease is steep and if the health index was relatively low (but still elevated from the floor), the decline is relatively gentle.
3. If the window is deemed anomalous, then the health index increases. The increase is again proportional to the current value. If the health index was relatively low, the increase is steep. If the health index is already elevated, then the increase is relatively gentle.
4. During normal hours, the health index is updated at the end of the pump cycle using the same logic as steps 3 and 4 above.

We describe our solution to the anomaly detection problem and detail the corresponding alerting logic in the following

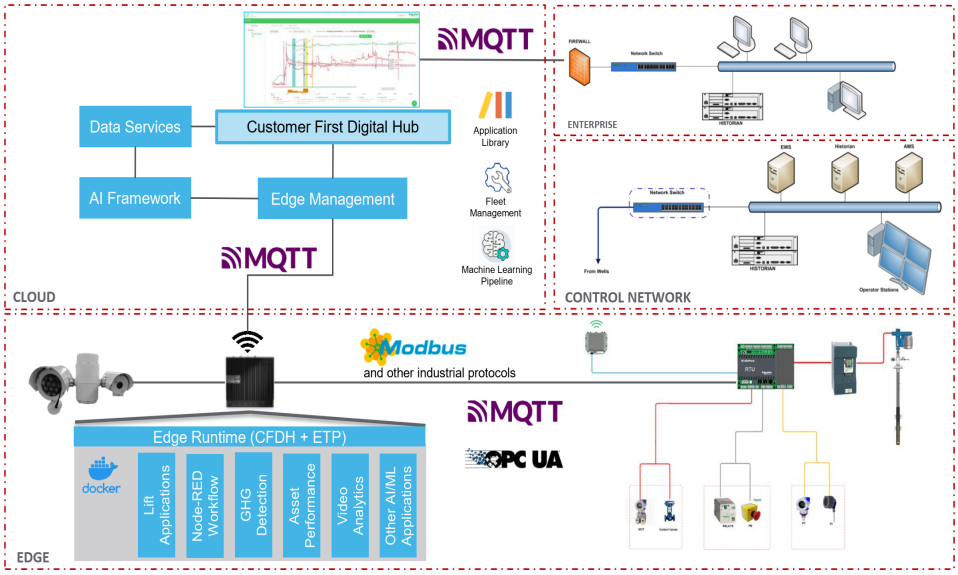


Figure 3. Architecture of the hybrid cloud-to-edge health monitoring system: CFDH.

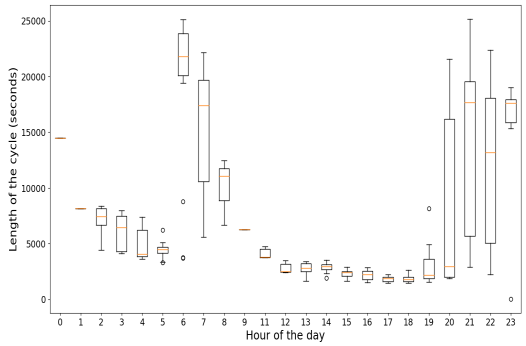


Figure 4. Cycles are longer in the mornings and night times.

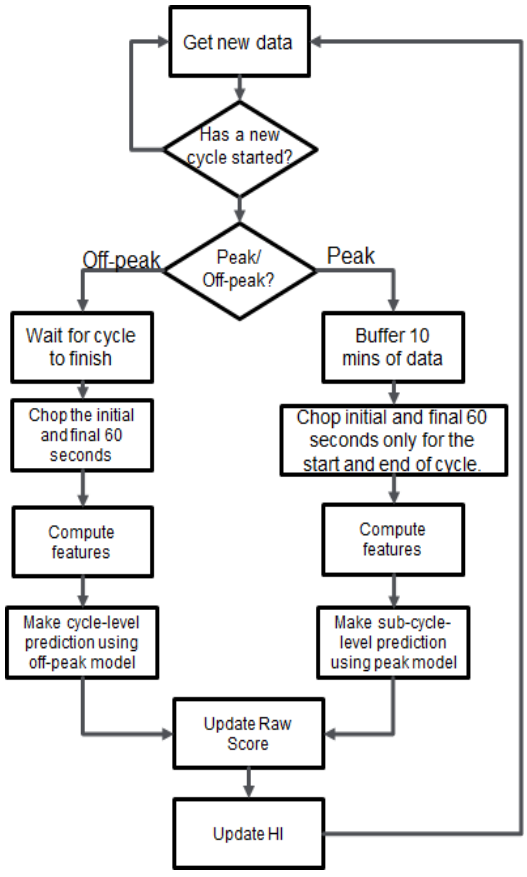


Figure 5. TSAD flowchart.

subsections. We summarize the working of the anomaly detection algorithm in the flowchart below.

5.1. Pre-processing and Feature Engineering

Every pump cycle begins with an inrush of current into the motor. This results in a transient spike in the motor current at the beginning of the cycle. A similar transient behavior is seen at the end of the pump cycle, where the motor current ramps down over a period of few seconds.

We ignore the transient periods at the beginning (inrush) and the end (ramp-down) of the pump cycle. Both these periods are fixed to be 60 seconds in our case.

The pump operates in a steady state during the cycle and we focus our feature engineering logic on this period. The features are intended to capture the temporal patterns across the four condition indicators collected by the system: motor amps, flowrate, water level and the water pressure. The features include both statistical aspects of these signals, as well as morphological changes, like step-changes. We detail the features, along with the motivations below.

1. The *range, mean, median, skew, and kurtosis of the motor current and the water level* in the tank. Motor current is an important condition indicator and any changes in the statistical distribution in time is often associated with wear and tear. Statistical features like the range, mean median, skew and kurtosis can be used to characterize the signal. The nominal distribution is learned during the training phase and any deviations during operations are detected using these features.
2. The number of changepoints in the level time series: *Step changes* in the signals, which are captured as changepoints, indicate sudden unexpected changes in the signals and are often associated with anomalous behavior. We compute the changepoints in the water level signal, as step changes in the water level are associated either with an unexpcted change in the pump's ability to clear the water, or external sources of water, like rainfall.
3. *The correlation between the flowrate and the pressure*: Under nominal operation, the flowrate and the pressure signals must be highly correlated. When a blockage occurs, the synchrony between the flowrate and pressure get broken. Therefore, we use the correlation of the two signals as a feature for anomaly detection.

These features were selected from a superset of features using feedback from subject-matter experts. Note that all our features are in the time-domain.

5.2. Modeling

We built a Local Outlier Factor (LOF) algorithm to estimate anomalies in the wastewater treatment pumps. Next, we present a brief overview of the algorithm and the implementation. The LOF algorithm, which is illustrated in Fig. 6, works as follows. A given data point is tested for novelty by measuring its density. Density is the inverse of reachability, which measures the distance that n. If the data point has a high density, it means that the k'th nearest neighbor is far away, i.e we need to travel far to meet the k'th nearest neighbor. On the other hand, if the point is less dense, then its k-th nearest neighbor is close-by. In Fig. 6(a), the green circle denotes the distance between the test point in the middle and its 5th (k) nearest neighbor. In subfigure (b), similar circles are shown for the 5 nearest neighbors. The density for the test point is compared to the average density of its 5 nearest neighbors. This is equivalent to comparing the circle on the left with a circle whose radius is the average of the density circles on the right in Fig. 6.

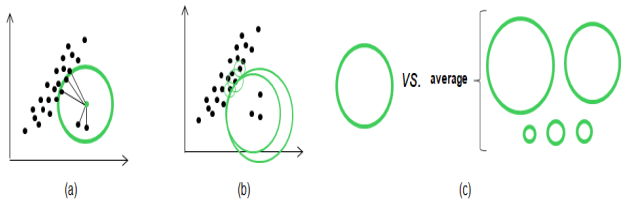


Figure 6. Working of the LoF Algorithm.

When the density of the test point is larger than the average density of of its k-nearest neighbors, the test point is deemed to be a local outlier. On the other hand, if the density of the test point is lower than the average of the k-nearest neighbors, it is consider an inlier.

The open-source machine learning library Scikit-Learn (Scikit-Learn, n.d.) was used to implement the novelty detection algorithm for wastewater pumps. The novelty detection model entails learning an LOF model from the training data, assuming that there are no outliers. The number of neighbors, k, was tuned by hand and set to 20 for the intra-cycle (peak) model and 5 for the inter-cycle model (off-peak). Automated grid-based tuning of other hyperparameters of the model is planned as future work.

Additionally, we also tested Isolation Forests, for the novelty detection model. The overall performance of the LoF algorithm was beter and thus, it was chosen for the pipeline.

5.3. Alerting Logic based on the Health Index

The LoF algorithm described in the previous subsection detects anomalies at a cycle or a sub-cycle level depending on the time of the day. The predictions themselves are not very actionable. We synthesize a Health Index (HI) from the individual predictions to quantify the *recency* and *frequency* of the anomalies.

The HI is computed by scaling a *raw score* (rs), which is updated after each invocation of either of the two novelty detection models.

$$rs_i = rs_{i-1} + \begin{cases} -20 \text{ i-th prediction is not-novel (inlier)} \\ +10 \text{ i-th prediction is novel (outlier).} \end{cases} \tag{1}$$

Eq. 1 describes the update of the raw score after every prediction. If the current prediction is nominal, then the raw score is decremented by 20. If the prediction is novel (outlier), then the raw score is incremented by 10.

The HI is calculated as follows.

$$HI = 100.\frac{1}{3}.\log_{10}\left[max\{rs, 1\} + 0.55\right] \tag{2}$$

The function in Eq. 2 plotted in Fig, 7. The HI grows rapidly for small values in the rs: any anomalies after a prolonged nominal period lead to a rapid increase in the HI. As the HI increases to larger values, the growth slows down. The slow

growth reflects inertia against too many anomalies as they do not present any new insights about the system.

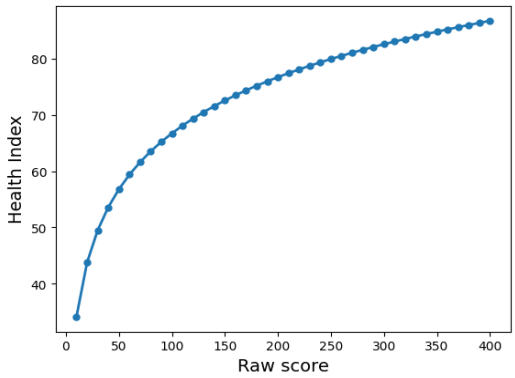


Figure 7. Scaling raw scores to obtain the HI.

When the system recovers from anomalous behavior, the raw score is decremented. The impact on the HI depends on the current value of the index. If the HI is already elevated, the HI reduces slowly with every decrement of the raw score. This ensures that the HI does not reduce quickly if there were many anomalies in the recent past. As time passes, the health index decelerates faster as it reduces in value. Thus, if the HI is already low, an intermittent anomaly will cause a small blip in the HI that will recover back quickly.

The HI can be used to create an alert when a pre-defined threshold is crossed. The threshold for alerting is meant to be context-specific but is generally dependent on a couple of factors: asset criticality and the desired sensitivity. If the asset is highly critical, then a relatively low threshold will ensure that alerts would be highly responsive to any observed anomalies. On the other hand, a lower criticality asset can afford the alerts to be generated at a much higher threshold. The overall sensitivity of the system may be used to set the specific thresholds for each asset after a commissioning phase. Subject-matter experts could weigh-in on the process of setting the thresholds. Finally, the thresholds may be reviewed periodically and tuned to maintain the performance of the overall system.

6. RESULTS

In this section, we describe the experiment conducted at the Gladstone Regional Council's wastewater treatment station in Australia and the results obtained using our cloud to edge solution. The A01 site, which consists of three pumps, was chosen for the controlled experiment.Nominally, two of the bigger pumps are turned OFF and a relatively smaller pump is used to circulate the water from an upstream site, S01, to downstream sites of the wastewater treatment facility.

A blockage was simulated on the pump at A01 on August 29, 2023. The simulation entailed closing a valve at A01 to block the flow, thereby mimicking a pump blockage. The valve was closed gradually in a stepwise fashion at 8:07 am, 8:16 am, 8:28 am, 8:41 am (most severe blockage was attained). The valve was brought back to Normal at 8:56 am.

The goal was to detect the blockage event using the proposed pump monitoring system.

The pump monitoring system was commissioned and started collecting data on August 19, 2023. Data was collected until November 23, 2023. The data was divided into training and inference as shown in Fig. 8. The data for the period spanned in yellow was used to train the anomaly detection models. The data spanned in blue, between August 27th and August 29th was used to test the pump monitoring system, including the blockage event.

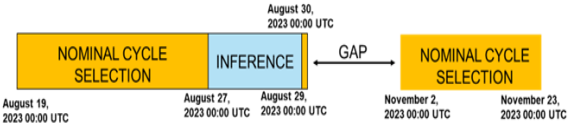


Figure 8. Time windows used for training and testing the pump monitoring system.

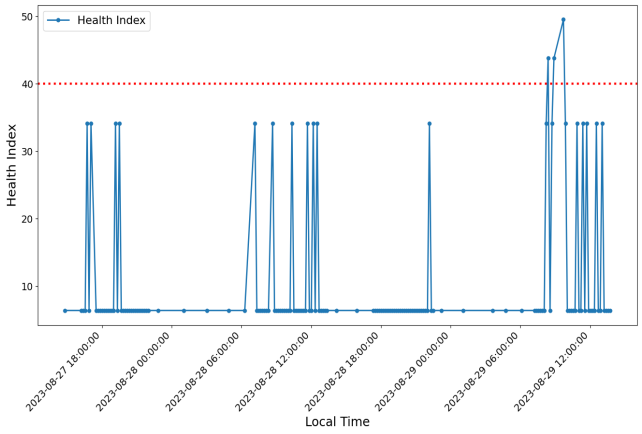
Fig. 9a illustrates the HI estimated during the test period, which lasted 3 days. The HI for the blockage event, which was done on the second day, is shown separately in Fig. 9b. The threshold for alarming was set at 40 and is shown in red. *Despite intermittent blips in the HI, the threshold was crossed only during the blockage event. Thus, the proposed system correctly raised the alarm during the blockage event and no false alarms were raised.*

We describe additional aspects of the HI estimation from Figs. 9a and 9b. Each prediction from the anomaly detection models triggers an update to the HI (see the flowchart in 5. During peak hours, the corresponding model is invoked every 10 mins after the start of a cycle. This leads to the HI being updated more frequently, as shown by the closely spaced data points. During off-peak hours, the updates are made at the end of the cycle, which tend to be longer than 10 mins. The HI updates are less frequent during these times. The HI was initially low at the start of the test period. Then the models detected intermittent anomalies several times during the test period. These resulted in the blips in the HI shown in Fig. 9a.

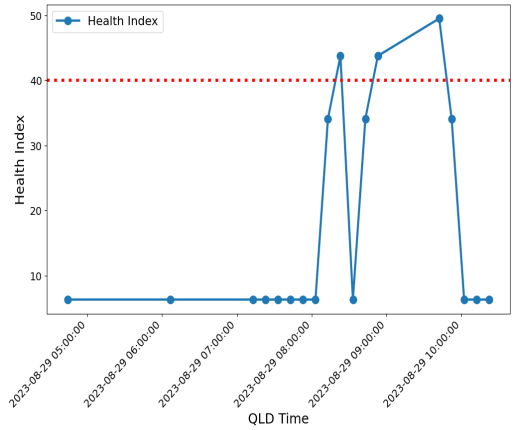
We focus on the day of the blockage event in Fig. 9b. The early hours of day, which are off-peak, had a couple of nominal cycles, as shown in Fig. 9b. The cycle that experienced the blockage started just after 7 pm. The first few 10-minute windows were nominal and then peak-model detected successive anomalies leading to the rise in the HI. The HI crossed the alarming threshold of 40 during the cycle.

A detailed look at some of the cycles scored by the peak model, including the blockage cycle is shown in Fig 10. We plot the motor amps, flow rate, water level and the pressure signals throughout the cycle. The initial ramp-up and ramp-down are also shown, but are not used for feature generation and scoring. The ten-minute windows are shown by vertical dotted lines. If a window was scored as nominal by the peak model, the vertical line at the beginning of the window is colored green. An anomalous window has a red dotted vertical line at the beginning.

Fig. 10a shows the blockage cycle and the predictions made by the model. The mock blockage involved closing the valve slowly starting at 8:00 am. We can see that the window around 8:15 am, and the subsequent window were called as anomalous by the model. As the valve further closes, the correlation between the flow rate and the pressure breaks down. This is picked up by the model, resulting in anomalous windows around 8:45 am. As the controlled experiment ended and the valve was restored, the system recovers and the corresponding windows are labeled as nominal.



(a) HI during the entire test period between August 27th and 29th.



(b) HI during the controlled blockage event.

Figure 9. HI estimated by the pump monitoring system.

Fig. 10b shows the performance of the peak model on a subsequent cycle. The cycle starts off with an anomalous window, which could be explained by the oscillatory behavior of the flow rate, resembling the *hammer effect*. Despite not being labeled anomalous, we believe that out model correctly scores and further demonstrates the generalizability of the model. We also see the hammer effect in a subsequent 10-min window. At around 13:00 hours, the model predicts an anomaly, which can be explained by the relatively higher amp values.

The model also predicts potential false positives, as shown in Figs. 10c-d. All the 10-minute windows of these cycles seem to be nominal based on visual inspection of the data. As shown in the figures, our model predicts some of the windows as anomalous. The false positive rate of the models can be tuned to control the behavior. Specifically, in the Scikit-Learn implementation of the Local Outlier Factor algorithm, the `score_samples()` function can be used to measure the degree to which an input is an inlier. The inlier score is the opposite of local outlier factor and it can be thresholded to accept a predefined level of false positivies. This threshold can even be tuned using an RoC curve. Finally, intermittent anomalies do not spike the HI beyond the alarming threshold and therefore the system is robust to infrequent false positives.

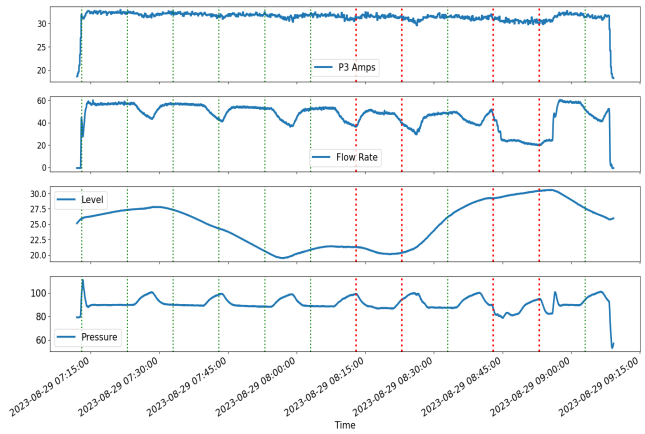
7. CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

In this paper, we presented a hybrid cloud-to-edge monitoring system that goes beyond traditional SCADA systems in enabling machine-learning-based predictive maintenance applications for industrial assets. Our system does not need additional sensors as it acquires the raw sensor data that is conventionally collected and transmitted over dedicated protocols by the SCADA system. We applied our system to the monitoring or wastewater treatment pumps. Blockage is a common problem for such pumps and we demonstrated that our proposed system is able to raise relevant alarms in a controlled experiment. The time series anomaly detection pipeline scores off-peak and peak-time cycles using different models and the predictions are then synthesized into a health index.

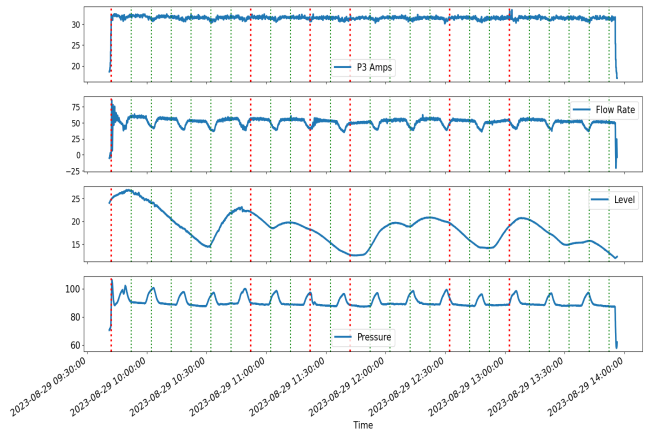
In the future, we will go beyond anomaly detection and train models for more granular insights. These could entail remaining useful life or degradation patterns based on comparing an asset with others in a fleet. Federated learning could be used to train the models on the edge. The feature generation pipeline will also be enhanced by adding time and frequency-domain features during the initial ramp-up and ramp-down periods of the motors.

REFERENCES

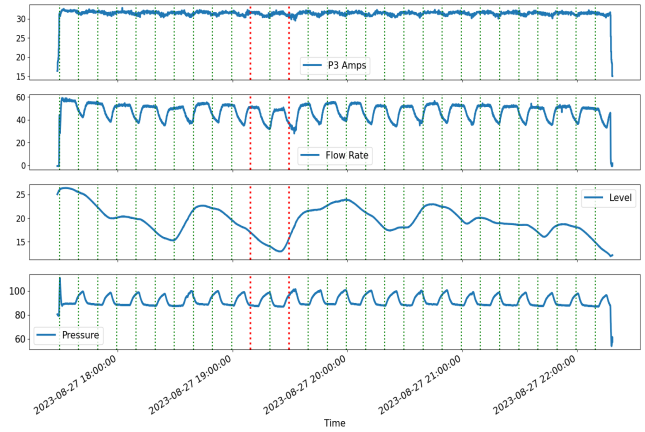
Concetti, L., Mazzuto, G., Ciarapica, F. E., & Bevilacqua, M.



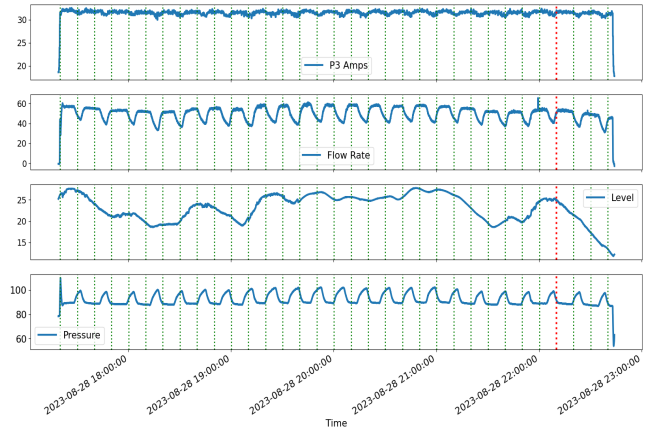
(a) Blockage cycle scored by the peak model.



(b) Example of a peak cycle with anomalies.



(c) Example of a peak cycle with potential false positives.



(d) Example of a peak cycle with potential false positives.

Figure 10. A detailed look at some of the cycles scored by the peak model.

(2023, 03). An unsupervised anomaly detection based on self-organizing map for the oil and gas sector. *Applied Sciences*, *13*, 3725. doi: 10.3390/app13063725

Giro, R. A., Bernasconi, G., Giunta, G., & Cesari, S. (2021). A data-driven pipeline pressure procedure for remote monitoring of centrifugal pumps. *Journal of Petroleum Science and Engineering*, *205*, 108845.

Moreno-Rodenas, A. M., Duinmeijer, A., & Clemens, F. H. (2021). Deep-learning based monitoring of fog layer dynamics in wastewater pumping stations. *Water Research*, *202*, 117482.

Mosallam, A., Medjaher, K., & Zerhouni, N. (2013). Nonparametric time series modelling for industrial prognostics and health management. *International Journal of Advanced Manufacturing Technology*, 1-25.

Scikit-Learn. (n.d.). *LocalOutlierFactor.* `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html`. (Accessed: 2024-06-16)

Šenk, I., Tegeltija, S., & Tarjan, L. (2024). Machine learning in modern scada systems: Opportunities and challenges. In *2024 23rd international symposium infoteh-jahorina (infoteh)* (p. 1-5).

Trstenjak, B., Palasek, B., & Trstenjak, J. (2019, Oct.). A decision support system for the prediction of wastewater pumping station failures based on cbr continuous learning model. *Engineering, Technology amp; Applied Science Research*, *9*(5), 47454749.