

What Level and Power Data Can Tell You About a Failing Submersible Pump: A Physics-Based Benchmark for Fleet-Wide Monitoring of Manhole Wastewater Pumping Stations

Mohammad Salkhordeh Haghighi¹, Seyed Mohammad Miri*²

¹ Assistant Professor, Department of Civil Engineering, University of Zabol, Zabol, Iran.

² Department of Civil Engineering, Velayat University, Iranshahr, Iran

*Corresponding Authors: Email: (✉ m.miri@velayat.ac.ir).

Received: 12 January 2026

Accepted: 10 February 2026

Available online: 11 March 2026

Abstract: Most manhole pumping stations are instrumented with a level sensor and a motor-current or power signal, and nothing else. Papers on machine-learning fault detection for pumps report accuracies above 0.95, but they are almost always validated on cycles drawn at random from the same stations used for training, and they say little about the alarm layer that an operator actually lives with. We built a physics-based digital twin of a submersible-pump wet well (mass balance, pump and system curves, efficiency and motor power) and used it to generate 400,411 pump cycles from 24 stations over 150 days, 12 of which develop rag build-up, impeller wear or a rising static head until the station can no longer empty the well. Three detectors were compared: an EWMA control chart on the pumping rate, gradient boosting on absolute cycle features, and gradient boosting on features referenced to each station's own healthy baseline. Splitting cycles at random gave a PR-AUC of 0.80 for the absolute-feature model; holding out whole stations dropped it to 0.06. Self-referencing the features kept PR-AUC at 0.73 across unseen stations. With conformal thresholds calibrated on ten days of healthy operation and a run rule of ten cycles, the self-referenced model detected 12 of 12 failures a median of 23 days ahead, at 46 false alarms per healthy station-year; the EWMA chart reached 30 days of lead time at 3.0 false alarms per station-year. Conformal exceedance rates observed on healthy stations were five to thirteen times their nominal values, because consecutive pump cycles are not exchangeable. The benchmark, the simulator and the evaluation code are released so that utilities can repeat the comparison on their own SCADA archives.

Keywords: Wastewater pumping station; submersible pump; condition monitoring; machine learning; conformal prediction; digital twin.

1. Introduction

A manhole pumping station is a small structure with a large failure cost. Two or three submersible pumps sit in a wet well a few metres across, the level sensor starts and stops them, and the only other routinely recorded signal is the motor current or power. When the impeller collects rags, when it wears, or when the rising main partly blocks, the station keeps running: the pumps simply take longer to empty the well, the energy used per cubic metre climbs, and eventually the inflow overtakes the pumps and sewage reaches the street or a watercourse. Operators would like to know about that weeks ahead, and they would like to know it without adding flow meters or vibration sensors to hundreds of small stations.

Data-driven condition monitoring for pumps is an active field. Motor-current signature analysis has been applied to wastewater pumps to detect clogage (Becker et al., 2022) and to centrifugal pumps in a desalination plant, where a rule-based harmonic method reached 91.2% diagnostic accuracy (Adaika et al., 2025). Vibration data with deep autoencoders reached an F1 of 0.995 on sewage-pump motors at a large treatment works (Krastev et al., 2024), and neural models detected inter-turn faults in pumping systems with near-perfect fit

(Dutta et al., 2022). Closer to the signals available in a manhole, Eshkofti et al. (2025) reconstructed inflow from wet-well level and pump states at one-second resolution in a Stockholm station and detected faults with physics-based residuals, and Lin et al. (2025) estimated sewer flow from manhole levels alone with a Nash-Sutcliffe efficiency above 0.95.

Two gaps remain. The first is the validation protocol. Varejão et al. (2024) showed on a public submersible-pump vibration dataset that a random k-fold split gave a mean F-measure of 0.887, while a split that respected the dependency between instances gave 0.733. In a fleet of pumping stations the dependency is stronger still, because each station has its own geometry, static head and inflow pattern; a model trained on cycles from a station will recognise that station rather than the fault. The second gap is the alarm layer. A classifier produces a score; an operator needs a threshold, and that threshold decides how many callouts the utility pays for. Conformal prediction gives distribution-free thresholds under exchangeability (Vovk et al., 2022) and has been used to monitor fleets of near-identical assets in district heating (Farouq et al., 2022), but consecutive pump cycles are autocorrelated and seasonal, so the nominal rate is not what the operator gets.

Field data that would settle these questions are scarce and rarely shared. We therefore built a physics-based twin of a manhole pumping station, seeded it with three degradation mechanisms, and used it as a benchmark. The twin is not a substitute for field data, and we make no claim about the absolute accuracy that any method would reach on a real fleet. It does let us answer three questions that a utility can act on: which features transfer to a station the model has never seen, how much lead time a realistic alarm rate buys, and how much healthy history a new station needs before it can be monitored.

2. The digital twin

2.1. Hydraulics

The wet well is a vertical cylinder of plan area A . With Q_{in} the inflow and Q_{out} the pumped flow, the level L follows the mass balance

$$A \frac{dL}{dt} = Q_{in}(t) - Q_{out}(t) \quad (1)$$

Pumps start when the level reaches L_{start} , a second pump joins at L_{high} , and all pumps stop at L_{stop} . The duty point comes from the intersection of the head-capacity curve of the pumps running in parallel and the system curve,

$$H_{sh}(1 - w) - a \left(\frac{Q}{n} \right)^2 = H_{st} + K_{sys} Q^2 \quad (2)$$

in which H_{sh} is the shut-off head, a the curve coefficient, w the fractional loss of head caused by impeller wear, n the number of pumps running, H_{st} the static lift and K_{sys} the resistance of the rising main. Motor power follows from the duty point and the efficiency,

$$P = \frac{\rho g Q H}{\eta} \quad (3)$$

with the efficiency falling away from the best efficiency point and with wear,

$$\eta = \eta_0 (1 - 0.35w) \left[1 - 2.2 \left(\frac{Q}{Q_{BEP}} - 0.85 \right)^2 \right] \quad (4)$$

which is how a real pump behaves: a worn impeller pushed away from its design point burns more energy for less flow.

2.2. Degradation and failure

Three mechanisms were implemented, each starting at a random time and developing over 8 to 25 days: rag build-up, which raises K_{sys} by up to a factor of six; impeller wear, which removes up to 45% of the head; and a rising static head, which adds up to 6.5 m of lift. A station is declared failed at the first time its pumping rate stays below 60% of its own initial rate for six hours, which is the point at which an operator would be called out. Cycles in the last 14 days before that moment are labelled as degraded, and everything earlier is labelled healthy.

2.3. Sensors and fleet

Level is recorded every 30 s with 1 cm of noise and motor power with 2% noise; there is no flow meter, which is the usual situation in a manhole station. The fleet has 24 stations with plan areas of 4 to 9 m², static heads of 6 to 13 m, dry-weather flows of 18 to 45 l/s and their own diurnal patterns and wet-weather events. Nine stations stay healthy for the whole 150 days and 12 reach the failure threshold, at a median of 95 days. Together they contain 400,411 pump cycles, of which 15,817 fall in a degraded window.

Table 1. Parameters of the simulated fleet

Quantity	Symbol	Value or range	Note
Stations	–	24	9 healthy, 15 with a developing fault
Record length (days)	–	150	30 s sampling
Wet-well area (m ²)	A	4–9	uniform across stations
Static lift (m)	H_{st}	6–13	uniform
Rising-main resistance (s ² m ^{−5})	K_{sys}	600–1400	uniform
Shut-off head (m)	H_{sh}	16–24	uniform
Pump efficiency at BEP	η_0	0.55–0.70	uniform
Dry-weather flow (l/s)	Q_{dwf}	18–45	uniform
Diurnal peaking factor	–	1.6–2.2	uniform
Wet-weather events (per year)	–	15–40	Poisson arrivals
Level noise (m)	σ_L	0.01	Gaussian
Power noise	σ_P	2%	multiplicative
Rag build-up	ΔK_{sys}	up to $6 \times K_{sys}$	ramp over 8–25 days
Impeller wear	w	up to 0.45	ramp over 8–25 days
Static-head rise	ΔH_{st}	up to 6.5 m	ramp over 8–25 days
Failure threshold	–	60% of the initial pumping rate for 6 h	triggers a callout
Label window (days)	–	14	before the failure threshold

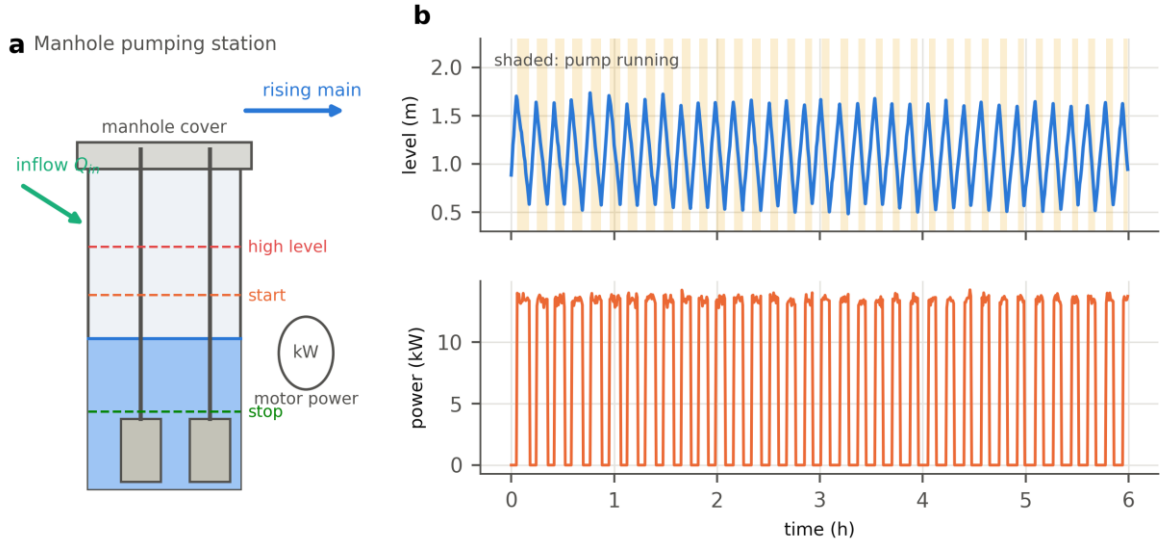


Figure 1. The simulated station: (a) layout of a two-pump manhole station with the control levels; (b) six hours of level and (c) motor power at 30 s resolution.

3. Features and detectors

3.1. Cycle features

Each pump cycle is segmented from the run signal. During the filling phase the inflow follows from the level rise, and during pumping the discharge follows from the level fall plus the inflow,

$$\hat{Q}_{in} = A \frac{\Delta L}{\Delta t} |_{fill}, \quad \hat{Q}_{out} = \hat{Q}_{in} + A \frac{|\Delta L|}{\Delta t} |_{pump} \quad (5)$$

which is the wet-well mass balance used by Eshkofti et al. (2025). From these we compute the run time, the idle time, the starts per hour, the mean motor power and the specific energy P/Q_{out} in kWh per cubic metre. Absolute features are the quantities themselves. Self-referenced features divide each quantity by the median of the same quantity over the first ten days at that station, and add 24-hour rolling medians and 72-hour slopes of the referenced signals. The difference is the whole point of the comparison: a pumping rate of 40 l/s means nothing on its own, while a pumping rate 20% below what this station did last month means a great deal.

3.2. Detectors and validation

Three detectors were compared. The EWMA control chart tracks the standardised drop in the self-referenced pumping rate with a smoothing constant of 0.05. The two gradient-boosting models (300 trees, depth 3, learning rate 0.05) use the absolute and the self-

referenced feature sets. Each model was evaluated twice: with five-fold cross-validation over cycles drawn at random, and with grouped cross-validation in which all cycles of a station are held out together. The first protocol is the one used in most published comparisons; the second is what a utility faces when it installs the model at a new station.

3.3. Alarms

Scores are turned into alarms per station. The first ten days of a station are treated as healthy and used as the calibration set of a split-conformal threshold at a nominal exceedance rate $\alpha = 0.01$ (Vovk et al., 2022). To avoid reacting to single cycles, an alarm episode is raised only after m consecutive exceedances. Lead time is measured from the first alarm episode to the failure threshold, and the false-alarm rate is counted on the healthy stations, in episodes per station-year.

4. Results

4.1. What degradation looks like in the data

Figure 2 shows one degrading station against a healthy one. The pumping rate leaves its baseline about three weeks before the station fails and falls to roughly two thirds of it; the specific energy rises by about 70%; the starts per hour fall, because each cycle takes longer. All three signals are visible at the cycle level, but they are buried in scatter of similar magnitude to the early drift, which is what makes the smoothing and the alarm rules matter.

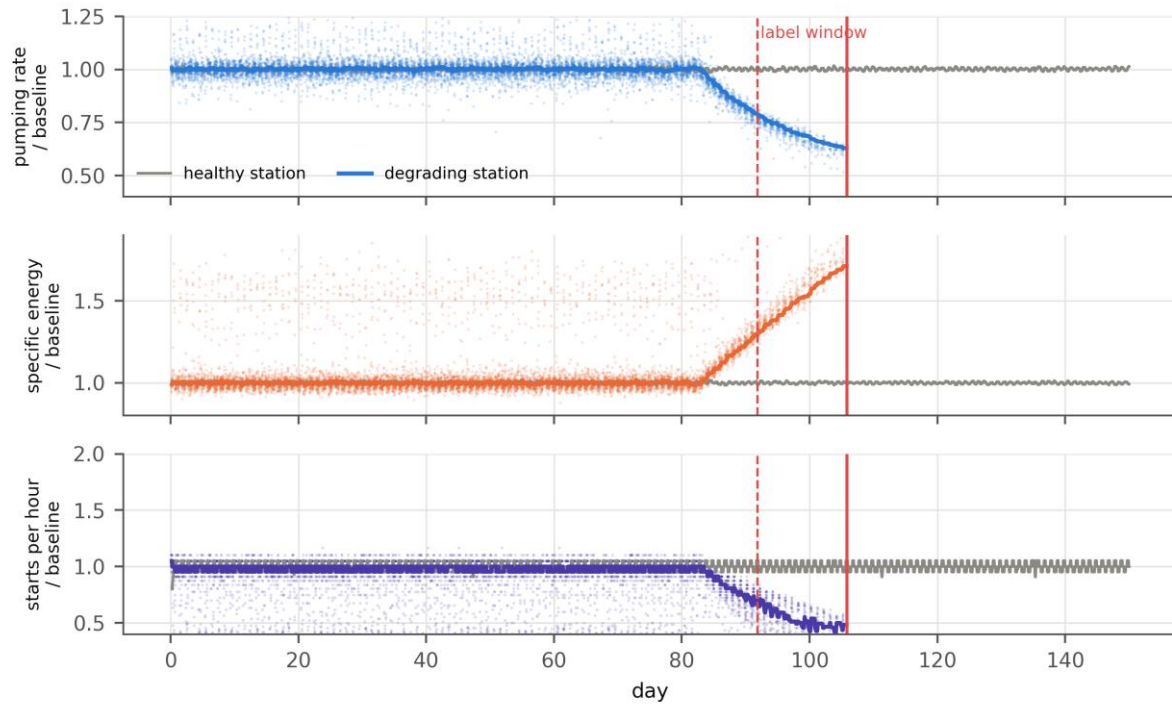


Figure 2. Cycle-level signature of degradation at one station (blue, orange, purple) against a healthy station (grey): pumping rate, specific energy and starts per hour, each divided by the station's own baseline. The solid red line is the failure threshold, the dashed line the start of the labelled window.

4.2. Validation protocol changes the answer

Table 2 and Figure 3 give the discrimination results. With cycles split at random, the model trained on absolute features reaches ROC-AUC 0.971 and PR-AUC 0.80, which would be reported as a success. Holding out whole stations, the same model collapses to ROC-AUC 0.587 and PR-AUC 0.056, barely above the prevalence of 4.0%. The model had learned which station it was looking at. Referencing the features to each station's own baseline removes that shortcut: PR-AUC falls only from 0.99 to 0.73 when the station is new. Anyone who reports a single random-split number for a fleet of pumping stations is measuring the wrong thing.

Table 2. Discrimination of degraded cycles under the two validation protocols

Detector	Validation	ROC-AUC	PR-AUC
XGB (absolute features)	random cycles	0.971	0.803
XGB (absolute features)	leave-one-station-out	0.587	0.056
XGB (self-referenced features)	random cycles	1.000	0.991
XGB (self-referenced features)	leave-one-station-out	0.984	0.733
EWMA control chart	per-station baseline	0.557	0.349

400,411 cycles from 24 stations; prevalence of degraded cycles 4.0%.

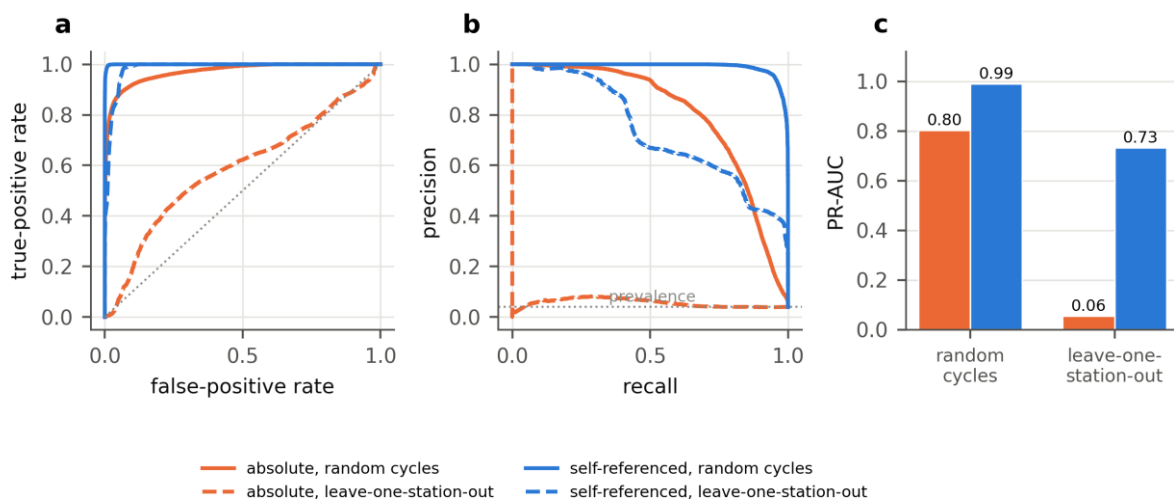


Figure 3. Effect of the validation protocol: (a) ROC curves, (b) precision-recall curves and (c) PR-AUC for the two feature sets under random-cycle and leave-one-station-out validation.

4.3. Alarms, lead time and false alarms

Discrimination is not detection. Table 3 converts the scores into alarms with conformal thresholds and a run rule of ten cycles. The self-referenced model warns at 12 of 12 stations, with a median lead time of 23 days (quartiles 18 and 34); the absolute-feature model warns

at only 12 stations and 16 days ahead. The EWMA chart, the simplest of the three, warns everywhere with a median lead of 32 days and raises 9.6 false-alarm episodes per healthy station-year against 46 for the machine-learning model.

Table 3. Alarm performance with conformal thresholds ($\alpha = 0.01$) and a run rule of ten cycles

Detector	Validation	Detected	Median lead (d)	Quartiles (d)	Stations with false alarms	False alarms per station-year
XGB (absolute features)	random cycles	12/12	16.2	15–19	2/12	48.0
XGB (absolute features)	leave-one-station-out	9/12	9.6	3–12	2/12	29.3
XGB (self-referenced features)	random cycles	12/12	23.0	18–34	7/12	45.6
XGB (self-referenced features)	leave-one-station-out	12/12	27.3	18–40	7/12	50.8
EWMA chart on pumping rate	per-station baseline	12/12	31.7	20–82	8/12	9.6

The run rule is the knob that matters (Figure 4b). Requiring a single exceedance gives the longest lead times and hundreds of false alarms per station-year. At $m = 20$ the EWMA chart still detects all 12 failing stations 30 days ahead with 3.0 false alarms per station-

year, and the self-referenced model gives 22 days at 20 false alarms. A utility that can tolerate two or three unnecessary visits per station per year can therefore expect roughly a month of warning from data it already collects.

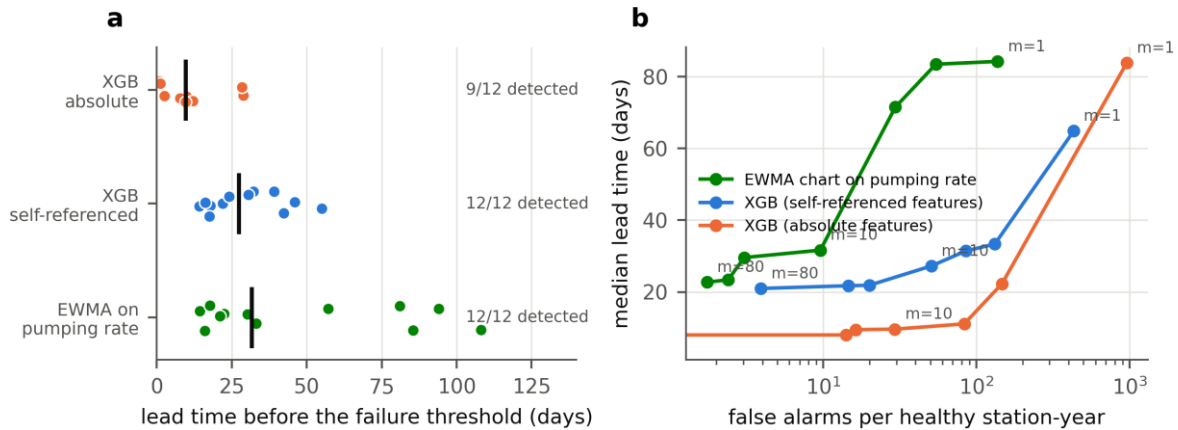


Figure 4. Alarms: (a) lead time of the first alarm episode at each failing station, with medians as vertical bars; (b) trade-off between false alarms and lead time as the run rule m varies from 1 to 80 cycles.

4.4. Conformal thresholds are optimistic, and short baselines are expensive

Figure 5a compares the nominal exceedance rate of the conformal threshold with the rate observed on healthy stations. At $\alpha = 0.01$ the observed rate is 0.125 for the EWMA chart and 0.093 for the self-referenced model, five to thirteen times the nominal value; at $\alpha = 0.001$ the gap is larger still. Pump cycles are not exchangeable: they are autocorrelated, they follow the diurnal and weekly rhythm of the catchment, and wet weather moves the whole distribution. The run rule

absorbs part of this, but a utility should calibrate the threshold on the alarm rate it can live with rather than trust the nominal level.

Figure 5b asks how much healthy history a new station needs. With three days of baseline the median lead time is 69 days but the chart raises 23 false alarms per station-year, because the baseline scatter is underestimated. Ten days brings the false alarms down to 10 per station-year, and there is little further gain beyond three weeks. Two weeks of clean data is a reasonable commissioning requirement.

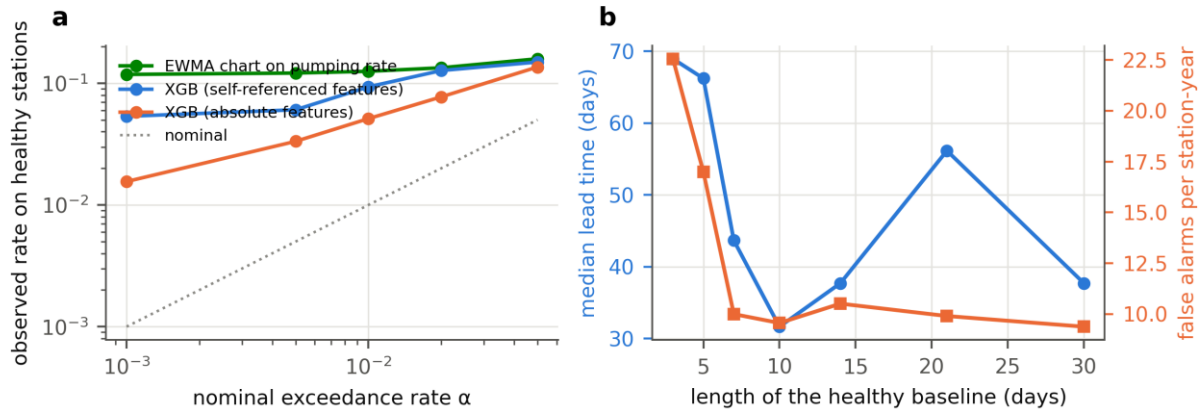


Figure 5. Threshold behaviour: (a) nominal against observed exceedance rate on healthy stations; (b) median lead time and false-alarm rate as a function of the length of the healthy baseline.

4.5. Which signals carry the information

The SHAP decomposition of the self-referenced model (Figure 6) puts the 24-hour median of the pumping rate first by a wide margin (mean $|\text{SHAP}| = 3.83$), followed by its 72-hour slope (1.81) and the 24-hour median of motor power (1.01). Single-cycle values

carry little weight. This is a useful negative result for instrumentation: the information that matters is a slow drift of the pumped volume per unit of energy, which a level sensor and a power meter resolve once they are aggregated over a day, so adding faster sensors would not help as much as keeping a clean baseline.

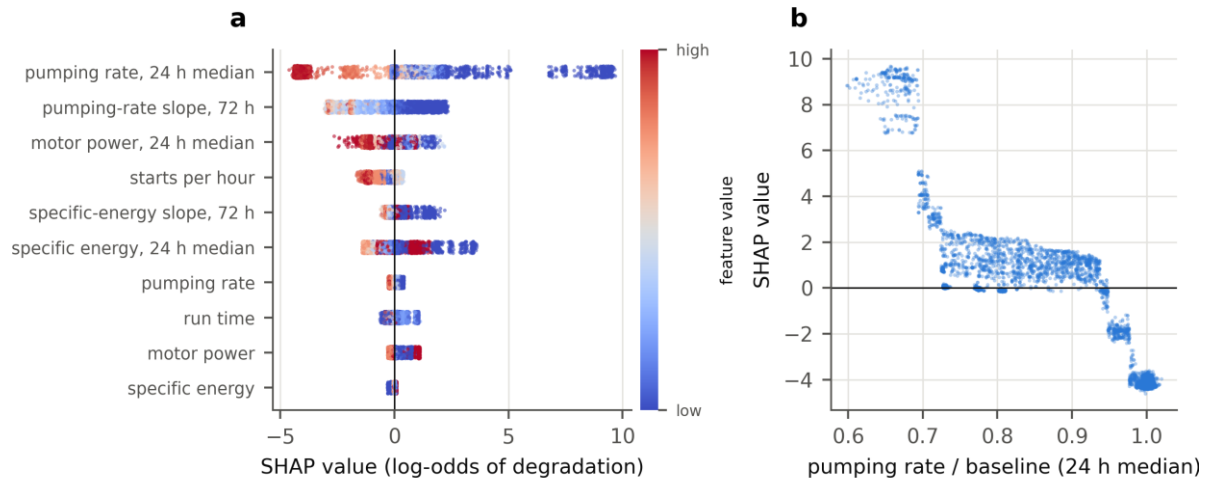


Figure 6. SHAP values of the self-referenced model: (a) contribution of each feature, coloured by the feature value; (b) dependence on the 24-hour median pumping rate.

5. Discussion

The three results point to the same practical conclusion: in a fleet of small pumping stations, the reference is the station itself. Absolute features let a model memorise stations, and they transfer badly; self-referenced features keep almost all of their skill at a station the model has never seen. The same logic explains why a plain control chart on the pumping rate competes with gradient boosting here. When the physics of the fault is a slow loss of flow at higher energy cost, and when each asset supplies its own reference, there is not much left for a flexible model to learn.

Where machine learning did help was in combining the signals. The self-referenced model detected every failing station and kept a lead time above three weeks at a run rule that a small utility could accept, and its SHAP

decomposition separates the mechanisms: rag build-up shows up first in the pumping rate and the specific energy, wear mostly in power, a rising static head in both. A model of this kind is also easy to deploy, because it needs no labels from the station where it runs. Labels come from the fleet; the threshold comes from the station.

The alarm layer deserves as much attention as the classifier. Reporting a ROC curve says nothing about how often a crew will drive to a healthy station. Our conformal thresholds, calibrated on ten days of healthy operation, produced five to thirteen times their nominal exceedance rate, and the only reason the final numbers are usable is the run rule. Fleet-level conformal schemes that pool calibration across similar assets (Farouq et al.,

2022) are a natural next step, and they can be tested on this benchmark.

The limits of the study are those of any simulation. The twin reproduces the hydraulics and the energy of a wet well, but not the electrical signatures of bearing or winding faults, not the sudden blockages that a single large rag can cause, and not sensor drift, data gaps or the operator interventions that fill real archives. Degradation is smooth and monotonic, which is kind to detectors. Inflow has a diurnal cycle and storms but no seasonal trend, and all stations share one pump type. Absolute performance figures from this benchmark should therefore not be quoted as expected field performance. What does transfer is the comparison: the ranking of feature sets, the size of the leakage effect, and the shape of the false-alarm curve. The same scripts read a SCADA export with columns for time, level, pump state and power, so a utility can repeat every figure on its own data and see how far the ranking holds.

6. Conclusions

(1) A gradient-boosting model on absolute cycle features reached PR-AUC 0.80 when cycles were split at random and 0.056 when whole stations were held out. Random splits of a fleet database overstate what a monitoring system will do at a new station.

(2) Features referenced to each station's own healthy baseline transfer: PR-AUC 0.73 across unseen stations, and every failing station was detected.

(3) With conformal thresholds and a run rule, the detectors warned a median of 23 days (machine learning) to 32 days (EWMA chart) before the failure threshold. Raising the run rule to 20 cycles cut false alarms to 3.0 per healthy station-year while keeping a lead time of 30 days.

(4) Observed exceedance rates on healthy stations were five to thirteen times the nominal conformal level, and about two weeks of clean baseline data are needed before a new station can be monitored reliably.

(5) Level and motor power, aggregated over a day, carry most of the detectable information. Extra instrumentation buys less than a clean baseline and a sensible alarm rule.

Data and code availability

The simulator, the generated cycle database, the detection and alarm scripts and the scripts that produce every figure and table are provided as supplementary material.

References

- Adaika, H., Tir, Z., Sahraoui, M., & Laadjal, K. (2025). PumpSpectra: an MCSA-based platform for fault detection in centrifugal pump systems. *Sensors*, 25(22), 6916. <https://doi.org/10.3390/s25226916>
- Becker, V., Schneider, M., Antonino-Daviu, J.A., & Urschel, S. (2022). Cloggage detection of a wastewater pump based on motor current analysis. In 2022 International Conference on Electrical Machines (ICEM) (pp. 690–695). IEEE. <https://doi.org/10.1109/ICEM51905.2022.9910776>
- Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Dutta, N., Kaliannan, P., & Shanmugam, P. (2022). Application of machine learning for inter turn fault detection in pumping system. *Scientific Reports*, 12, 12906. <https://doi.org/10.1038/s41598-022-16987-6>
- Eshkofti, K., Sandberg, H., Nilsson, M., & Barreau, M. (2025). Modeling and physics-enhanced fault detection in wastewater pump stations. *arXiv:2511.11304*.
- Farouq, S., Byttner, S., Bouguelia, M.-R., & Gadd, H. (2022). A conformal anomaly detection based industrial fleet monitoring framework: a case study in district heating. *Expert Systems with Applications*, 201, 116864. <https://doi.org/10.1016/j.eswa.2022.116864>
- Krastev, S., Ammartayakun, A., Mishra, K.J., Koduri, H., Schuman, E., Morris, D., Feng, Y., Bandi, S.S.R., Ngan, C.-K., Yeung, A., Li, J., Ko, N., Emdad, F., Rundensteiner, E., Ho, H.M.H., Wong, T.K., & Chan, J.P.C. (2024). META: deep learning pipeline for detecting anomalies on multimodal vibration sewage treatment plant data. In *Proceedings of the 16th International Joint Conference on Computational Intelligence* (pp. 461–474). <https://doi.org/10.5220/0013031600003837>
- Lin, R., Qiu, R., Hu, L., Ding, Y., & Yuan, Z. (2025). A low-cost soft sensor for sewer flow monitoring: learning from water level measurements in manholes. *Water Research*, 274, 123135. <https://doi.org/10.1016/j.watres.2025.123135>
- Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- Metcalf & Eddy, Tchobanoglous, G., Stensel, H.D., Tsuchihashi, R., & Burton, F.L. (2014). *Wastewater Engineering: Treatment and Resource Recovery* (5th ed.). McGraw-Hill Education, New York.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Varejão, F.M., Mello, L.H.S., Ribeiro, M.P., Oliveira-Santos, T., & Rodrigues, A.L. (2024). An open source experimental framework and public dataset for vibration-based fault diagnosis of electrical submersible pumps used on offshore oil exploration. *Knowledge-Based Systems*, 288, 111452. <https://doi.org/10.1016/j.knosys.2024.111452>
- Vovk, V., Gammerman, A., & Shafer, G. (2022). *Algorithmic Learning in a Random World* (2nd ed.). Springer, Cham. <https://doi.org/10.1007/978-3-031-06649-8>

