

Research Article

A Machine Learning-Based Smart SCADA Framework for Predictive Fault Detection and Optimisation-Based Control of an Urban Wastewater Treatment Plant

Besnik Hajdari^{1*}, Mazen El-Rayes²

¹ Faculty of Computer Science and Engineering, University for Business and Technology, Pristina, Kosovo

² Electrical Engineering Department, Higher Technological Institute (HTI), 10th of Ramadan City, Egypt

*besnik.hajdari@ubt-uni.net

Abstract

Supervisory control and data acquisition (SCADA) systems in urban wastewater treatment plants (WWTPs) traditionally act as passive data historians rather than predictive decision-support tools, leaving operators to react to effluent-quality excursions only after laboratory analysis confirms them. This study develops and fully implements, on a real, openly available case study [1–3], an end-to-end machine learning-based smart-SCADA architecture that (i) discovers and classifies plant operational states, (ii) predicts effluent quality ahead of laboratory confirmation using leakage-free soft sensors, and (iii) closes the loop with an event-triggered, optimisation-based controller. The case study is the UCI Water Treatment Plant dataset (527 daily records, 38 operational variables, expert-labelled operational states) from an urban plant in the Barcelona metropolitan area. Unsupervised clustering of the standardised feature space showed only weak agreement with the plant experts' own fault categorisation (adjusted Rand index = 0.062), motivating a supervised approach: a class-balanced support-vector classifier achieved the best fault-detection performance (ROC-AUC = 0.999, balanced accuracy = 0.928), while an unsupervised Isolation Forest achieved comparable discrimination without requiring any labelled fault history (ROC-AUC = 0.999). Soft-sensor regression models predicting effluent BOD₅, COD and suspended solids from upstream measurements and a one-day autoregressive term achieved hold-out R² of 0.297–0.301. Coupling these models to a derivative-free, surrogate-model-based controller that searches within the historically trusted operating envelope reduced the total predicted exceedance of illustrative regulatory discharge limits by 93.8% on a 105-day chronological hold-out, while fully resolving 28.6% of individual limit exceedances, using an intervention triggered on only 10.5% of days. The results, reported with explicit uncertainty and limitation analysis rather than optimistic point estimates alone, indicate that a modest, software-only augmentation of legacy SCADA infrastructure can meaningfully improve predictive fault awareness and compliance margins without new hardware instrumentation.

Keywords: Wastewater treatment; Machine learning; Smart SCADA; Soft sensor; Fault detection; Optimisation-based control

INTRODUCTION

Urban wastewater treatment plants (WWTPs) are among the most instrumented pieces of environmental infrastructure in a modern city, yet the supervisory control and data acquisition (SCADA) systems that monitor them are, in most existing facilities, essentially passive: they log flows, pH, conductivity and periodic laboratory determinands of biochemical oxygen demand (BOD5), chemical oxygen demand (COD) and suspended solids (SS), display them to an operator, and raise a threshold alarm only after a limit has already been crossed. Effluent-quality determinands in particular are typically available only hours to a day after sampling, so by the time a non-compliant discharge is confirmed in the laboratory, the plant has already discharged it. This reactive information architecture sits in tension with increasingly stringent discharge consents, growing energy and chemical costs, and the push toward resource-recovery-oriented, “smart” water utilities.

Machine learning (ML) offers a way to close this gap without new hardware: models trained on a plant's own historical SCADA archive can act as software (“soft”) sensors that estimate effluent quality minutes after upstream measurements are taken, well before the laboratory result returns [4]. A substantial body of work has demonstrated the value of data-driven models across WWTP operation more broadly — from performance monitoring and anomaly detection to full plant-wide surrogate modelling [5] — and deep-learning methods are increasingly proposed across the urban water cycle [6]. Fewer studies, however, close the loop: most stop at prediction or diagnosis and leave the translation of a prediction into an operational adjustment to the human operator. Where model-based control has been attempted, it has typically required either a calibrated mechanistic plant model or a proprietary, single-plant deployment [7], and recent digital-twin proposals for wastewater treatment, while promising, remain largely conceptual or simulation-based frameworks rather than demonstrations on an openly reproducible operational dataset [8].

This study addresses that gap by developing, and fully implementing end-to-end, a smart-SCADA architecture that combines (i) unsupervised and supervised operational-state / fault classification, (ii) leakage-free, cross-validated soft-sensor regression for effluent quality, and (iii) an event-triggered, optimisation-based controller that searches for corrective setpoint adjustments using the trained soft sensors as a surrogate plant model — a data-driven analogue of model predictive control (MPC) that requires no mechanistic process model. The case study is a real, publicly archived urban WWTP: the UCI Machine Learning Repository's Water Treatment Plant dataset [3], comprising daily sensor measurements from a plant in the Barcelona metropolitan area together with the plant experts' own historical operational-state classification, originally developed for conceptual-clustering-based fault diagnosis in what its creators themselves characterised as an “ill-structured domain” [1], [2]. Using the dataset's own expert labels as ground truth, rather than synthetic or self-referential labels, allows the fault-detection component to be validated against an independent, domain-expert judgement rather than against the model's own assumptions.

The contributions of this work are fourfold:

- An integrated, fully reproducible machine learning pipeline — soft sensors, a fault/anomaly classifier and an optimisation-based controller — implemented end-to-end and demonstrated on a real, openly available WWTP operational dataset, rather than on a proprietary plant or a purely mechanistic simulation.
- Use of the plant experts' own historical operational-state labels as ground truth for fault detection, enabling direct, literature-grounded validation rather than reliance on self-referential or synthetic labels.
- A leakage-free soft-sensor design that restricts predictors to variables measurable upstream of the plant outlet, augmented with a physically realistic one-day autoregressive term, evaluated under both blocked time-series cross-validation and a strict chronological hold-out.
- An event-triggered, derivative-free (differential-evolution) surrogate optimisation scheme explicitly suited to a heterogeneous soft-sensor ensemble that mixes smooth (linear, kernel, neural) and intrinsically non-smooth (tree-ensemble) models, together with a transparent, magnitude-based evaluation of closed-loop benefit that does not rely on a single, potentially misleading binary compliance metric.

The remainder of the paper is organised as follows. Section 2 describes the case-study dataset and the full modelling methodology, including the pre-processing, classification, regression and control-layer formulations. Section 3 reports and discusses the results of each stage together with the closed-loop control simulation, and states the limitations of the study explicitly. Section 4 concludes the paper.

MATERIALS AND METHODS

Figure 1 summarises the overall smart-SCADA architecture developed in this study: a physical treatment train instrumented by the plant's SCADA data historian feeds a pre-processing stage, which in turn feeds a perception layer of soft-sensor regression models and a fault/anomaly classifier; an event-triggered decision layer converts their outputs into an optimisation-based search for corrective setpoints, which are issued as actuator commands and mirrored to an operator alarm dashboard. Each block is described in the subsections that follow. All computations were implemented in Python 3 using scikit-learn [9] for the machine learning models and SciPy for numerical optimisation; complete source code, figures and result tables produced by the pipeline accompany this manuscript.

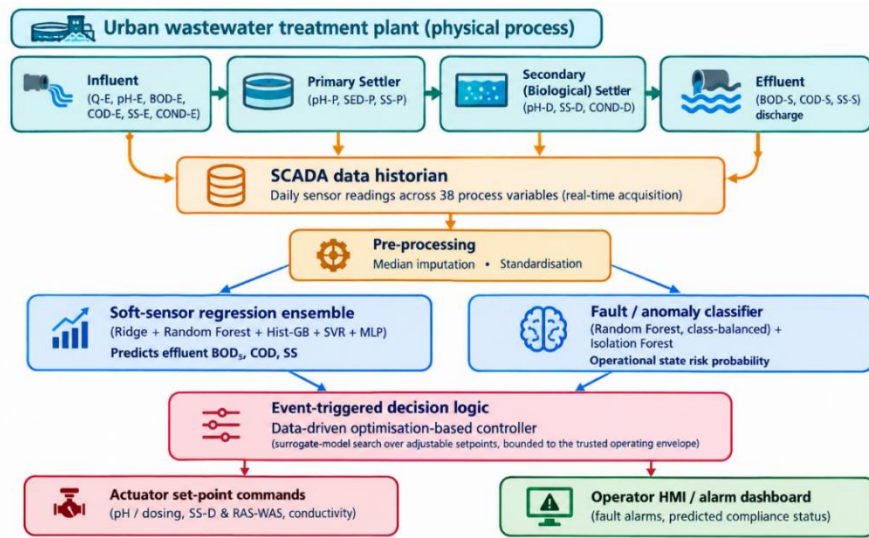


Figure 1. Smart-SCADA system architecture combining a soft-sensor ensemble, a fault/anomaly classifier and an event-triggered, optimisation-based control layer.

Case Study and Dataset

The case study is the Water Treatment Plant dataset donated to the UCI Machine Learning Repository [3] by the original plant researchers [1]. It contains 527 daily records collected between 1990 and 1991 at an urban wastewater treatment plant in the Barcelona metropolitan area, Spain, describing 38 continuous operational variables measured at four stages of the treatment train — plant influent, primary settler, secondary (biological) settler inlet, and final effluent — together with nine derived stage-wise and global pollutant-removal-efficiency indices (Table 1).

Table 1. Grouping of the 38 operational variables recorded at each treatment stage.

Treatment stage / group	Representative variables	N	Description
Influent (E)	Q-E, ZN-E, PH-E, DBO-E, DQO-E, SS-E, SSV-E, SED-E, COND-E	9	Raw sewage entering the plant
Primary settler (P)	PH-P, DBO-P, SS-P, SSV-P, SED-P, COND-P	6	After primary (physical) settling
Secondary settler inlet (D)	PH-D, DBO-D, DQO-D, SS-D, SSV-D, SED-D, COND-D	7	Entering the biological / secondary settling stage
Effluent / output (S)	PH-S, DBO-S, DQO-S, SS-S, SSV-S, SED-S, COND-S	7	Final discharged effluent
Removal-efficiency indices (RD)	RD-DBO-P, RD-SS-P, RD-SED-P, RD-DBO-S, RD-DQO-S, RD-DBO-G, RD-DQO-G, RD-SS-G, RD-SED-G	9	Stage-wise and global pollutant-removal percentages

In addition to the raw sensor variables, the dataset's original creators published a historical operational-state classification of each day, produced by expert review and conceptual clustering (LINNEO+) rather than by an automated distance-based algorithm [1], [2]. This study parsed that classification (thirteen named operational-state categories, e.g. “Normal situation”, “Storm”, “Solids overload”, “Secondary settler problems”) and matched it by date to the corresponding sensor record, successfully labelling 523 of 527 records (99.2%); the remaining 4 records, for which no matching date entry could be resolved in the original documentation, were retained for the unsupervised and regression analyses but excluded from the supervised fault-classification task. For classification purposes the thirteen fine-grained categories were consolidated into four operationally meaningful groups: Normal, Secondary Settler Fault, Solids Overload and Storm Event (Table 3). The resulting class distribution — 509 normal days against 14 documented abnormal days — is severely imbalanced, which is itself a realistic and well-documented characteristic of real fault archives (genuine upset episodes are, fortunately, rare) and was treated explicitly throughout the modelling process rather than masked by resampling.

Data pre-processing

Missing values, present in 2.95% of all cells and arising from routine sensor or laboratory-sample dropout rather than any systematic pattern, were imputed independently for each variable using the column median, a robust choice given the pronounced right-skew of several pollutant concentration variables. For scale-sensitive estimators (regularised linear models, support vector machines, the multilayer perceptron) all predictors were standardised to zero mean and unit variance within cross-validation folds via scikit-learn pipelines, so that no information from a held-out fold leaked into the scaling statistics used to transform it; tree-ensemble estimators, which are invariant to monotonic feature scaling, used the raw imputed values.

Operational-State Discovery and Fault/Anomaly Classification

Two complementary approaches to operational-state characterisation were compared. First, an unsupervised approach standardised all 38 variables, reduced them by principal component analysis (PCA) retaining enough components to explain 90% of total variance, and applied k-means clustering with the number of clusters k selected by the mean silhouette coefficient over $k = 2, \dots, 8$. The resulting cluster assignment was compared with the plant experts' four-category labels using the adjusted Rand index (ARI) and normalised mutual information (NMI), both of which equal 1 for perfect agreement and approximately 0 for agreement no better than chance.

Second, a supervised binary fault/anomaly classifier was trained to distinguish Normal from Abnormal operational days using the full 38-variable record. This retrospective, diagnostic use of the complete daily record (including output-stage and removal-efficiency variables) is appropriate because, unlike the predictive soft sensors of Section 2.4, this classifier is intended to run once a day's full record is available — for example as an automated quality-control audit of the closed day, or to generate a training label for the

following day's real-time model — rather than to anticipate the outcome before it is measured. Four estimator families were compared: logistic regression, a class-balanced random forest [10], a class-balanced radial-basis-function (RBF) support vector machine, and Gaussian naïve Bayes; all except naïve Bayes used class-balanced sample weighting to counteract the roughly 36:1 class imbalance. Performance was evaluated by five-fold stratified cross-validation with out-of-fold predictions pooled before scoring — necessary because per-fold metrics are unstable when a fold may contain only two or three positive (fault) examples — using balanced accuracy, F1, precision and recall for the fault class, ROC-AUC and average precision. Because the entire labelled dataset contains only 14 fault-class examples, an exhaustive leave-one-out cross-validation (LOOCV) of the best cross-validated model was additionally performed as a robustness check on the stability of the reported performance. Finally, an unsupervised Isolation Forest anomaly detector [11], with contamination fixed at the empirical fault rate and requiring no fault labels at all, was benchmarked against the supervised models to assess how much labelled fault history is actually necessary for useful anomaly screening — relevant for a newly instrumented plant with no fault archive yet.

Soft-Sensor Regression Models for Effluent Quality

Three soft sensors were built to predict the plant's principal effluent-quality determinands — BOD5, COD and suspended solids at the outlet (DBO-S, DQO-S, SS-S) — from information available before the corresponding laboratory result would normally be reported. Predictors were therefore restricted to the 22 variables measured at or before the secondary-settler inlet (Table 1, groups “Influent”, “Primary settler” and “Secondary settler inlet”); the output-stage physicochemical readings and the removal-efficiency indices were excluded because they are measured concurrently with, or algebraically derived from, the targets themselves and would otherwise leak the answer into the predictors. A one-day lag (autoregressive) term — the previous day's own measured value of the target — was added to each soft sensor's predictor set: daily effluent quality in activated-sludge plants is strongly autocorrelated (lag-1 Pearson $r = 0.53, 0.42$ and 0.59 for DBO-S, DQO-S and SS-S respectively in this dataset) and is genuinely available in real time before the current day's result is known, so this reflects standard soft-sensor practice [4] rather than a leakage risk.

Because all three targets are strongly right-skewed pollutant concentrations with occasional storm- or fault-day extremes, each target y was modelled in log space via $y^* = \ln(1 + y + \varepsilon)$ with $\varepsilon = 10^{-3}$ (Equation 1), and predictions were back-transformed with the inverse exponential map for evaluation and for use inside the control layer of Section 2.5. To keep the soft sensors — and, downstream, the optimiser that queries them thousands of times per triggered day — from extrapolating into physically implausible concentrations, every prediction was clipped in log space to the range observed in the training data before being back-transformed.

$$y^* = \ln(1 + y + \varepsilon) \quad (1)$$

Five candidate regressors spanning linear, ensemble, kernel and neural families were compared: ridge regression, a random forest [10], histogram-based gradient boosting, an RBF support vector regressor, and a single-hidden-layer multilayer perceptron (MLP). Two complementary evaluation protocols were used: (i) five-fold blocked time-series cross-validation (each fold's test block strictly follows its training block chronologically, avoiding any look-ahead bias inherent to ordinary random k-fold splitting of sequential SCADA data), and (ii) a single chronological hold-out consisting of the final 20% of the record (105 of 526 days), used both for the headline performance figures and, unmodified, as the scenario for the closed-loop control simulation of Section 2.5 so that the controller is always evaluated on data no model in the pipeline has seen during training. Performance was quantified by the coefficient of determination (R^2), root-mean-squared error (RMSE) and mean absolute error (MAE), all computed in the original (back-transformed) concentration units, together with permutation feature importance for the selected model of each target.

Smart-SCADA Decision and Control Layer

The control layer closes the loop between prediction and action. Illustrative regulatory reference values were taken from the minimum effluent requirements for discharges from urban wastewater treatment plants serving more than 10,000 population-equivalent set out in the Urban Waste Water Treatment Directive [12]: $\text{BOD}_5 \leq 25 \text{ mg L}^{-1}$, $\text{COD} \leq 125 \text{ mg L}^{-1}$ and suspended solids $\leq 35 \text{ mg L}^{-1}$. These values are used only to define a consistent, physically meaningful compliance threshold for the simulation, since the archived case-study dataset does not itself record the plant's actual discharge permit.

Four secondary-treatment process variables — secondary-settler inlet pH, primary-settler sediment volume, secondary-settler inlet suspended solids, and secondary-settler inlet conductivity (PH-D, SED-P, SS-D, COND-D) — were treated as proxy controllable setpoints, standing in for operator actions such as alkalinity/coagulant dosing, primary sludge wasting rate, and return/waste activated-sludge flow that a real plant SCADA system would expose as actuator commands. Because the archived dataset does not itself tag which variables correspond to direct actuator commands, this mapping is a modelling assumption, stated here transparently; it is used to demonstrate the control methodology rather than to prescribe specific dosing instructions. Each controllable variable was bounded to the 5th–95th percentile range observed in the training period (Table 6), keeping every candidate action inside the operating envelope the soft sensors were actually trained on and avoiding extrapolated, untrustworthy surrogate predictions.

A real-time fault-risk classifier — architecturally identical to the random forest of Section 2.3 but re-trained using only the same 22 upstream-only variables available to the soft sensors, so that it can genuinely run before the day's effluent is sampled — supplies a fault probability for each day. An intervention is triggered whenever either the soft sensors predict a limit exceedance under the current (unmodified) operating point, or the fault

probability exceeds a threshold of 0.30; all other days pass through the controller unchanged, keeping the scheme computationally cheap and operationally unobtrusive. On a triggered day, the controller searches for a setpoint vector u over the controllable variables that minimises a violation cost (Equation 2) — the squared, limit-normalised shortfall of each predicted target $\hat{y}_t(u)$ above its regulatory limit L_t — regularised by a control-effort penalty that discourages unnecessarily large excursions from the setpoint actually observed that day, u^0 (Equation 3), with each controllable variable's deviation scaled by its own training-period standard deviation s_j so that variables on very different physical scales (e.g. conductivity in the hundreds of $\mu\text{S cm}^{-1}$ versus pH near unity) contribute comparably to the penalty.

$$J_v(u) = \sum_t \max(0, \hat{y}_t(u) - L_t)^2 / L_t^2 \quad (2)$$

$$J(u) = J_v(u) + \lambda \sum_j [(u_j - u_j^0)/s_j]^2 \quad (3)$$

The soft-sensor ensemble selected for each target (Section 3.4) mixes smooth estimators (ridge, SVR, MLP) with intrinsically piecewise-constant, non-differentiable tree ensembles, whose exact gradient is zero almost everywhere between split thresholds. This defeats quasi-Newton or other gradient-based search methods for at least one of the three targets in every scenario, so the objective of Equation 3 is instead minimised with differential evolution, a population-based, derivative-free global search, operating on min–max-normalised decision variables so that every controllable variable is optimised on a comparable $O(1)$ scale regardless of its raw physical range (e.g. conductivity $\sim 10^3$ vs pH $\sim 10^0$). The simulation protocol replays the entire chronological hold-out day by day: for each day the baseline (open-loop) predicted effluent quality is computed from the actually observed operating point, the trigger condition is evaluated, and, where triggered, the optimiser proposes a controlled setpoint whose predicted effluent quality is compared against the open-loop baseline using both a binary compliance flag and a continuous exceedance-magnitude metric (the summed mg L^{-1} by which any target exceeds its limit), since a day on which a large exceedance is substantially reduced but not entirely eliminated is a meaningful practical improvement that a binary pass/fail count alone would obscure.

RESULTS AND DISCUSSION

Exploratory Characterisation of Plant Operation

Across the 527-day record the plant received a mean influent flow of $37,184 \text{ m}^3 \text{ d}^{-1}$ carrying a mean influent BOD5 of 188.4 mg L^{-1} , reduced to a mean effluent BOD5 of 19.9 mg L^{-1} — a mean global removal efficiency of 89.09% (Table 2), against 77.92% for COD and 88.99% for suspended solids. Figure 2 shows the correlation structure among a representative subset of influent, process and effluent variables: variables measured within the same treatment stage (e.g. suspended and volatile suspended solids at the influent) are strongly correlated, as expected from their shared physical origin, while the correlation between a given influent variable and its corresponding effluent variable is considerably

weaker — direct evidence that the biological treatment stage responds nonlinearly to influent load and that a simple linear scaling rule would be a poor effluent predictor, motivating the nonlinear machine learning models used in Section 3.4. Figure 3 compares the influent and effluent distributions of BOD5, COD and suspended solids: effluent distributions are markedly tighter and less skewed than their influent counterparts, consistent with the buffering, biologically mediated nature of secondary treatment. Figure 4 plots the daily global BOD5 removal efficiency across the full record with the expert-labelled abnormal days overlaid; abnormal days visibly cluster at the lower tail of the removal-efficiency distribution, a useful sanity check that the historical expert labels used throughout this study are physically meaningful rather than arbitrary.

Table 2. Descriptive statistics of key influent, effluent and removal-efficiency variables (imputed dataset, $n = 527$).

Variable	Mean	Std. Dev.	Min	Median	Max
Q-E	37184.33	6525.34	10050.0	35990.0	60081.0
PH-E	7.81	0.25	6.9	7.8	8.7
DBO-E	188.44	60.7	31.0	182.5	438.0
DQO-E	406.82	119.68	81.0	400.0	941.0
SS-E	227.39	135.82	98.0	196.0	2008.0
COND-E	1478.62	394.9	651.0	1406.0	3230.0
DBO-S	19.9	17.21	3.0	18.0	320.0
DQO-S	87.18	38.36	9.0	84.0	350.0
SS-S	22.2	16.26	6.0	19.0	238.0
PH-S	7.71	0.19	7.0	7.7	9.7
RD-DBO-G	89.09	6.79	19.6	90.2	97.0
RD-DQO-G	77.92	8.68	19.2	79.15	98.1
RD-SS-G	88.99	8.16	10.3	90.7	99.4

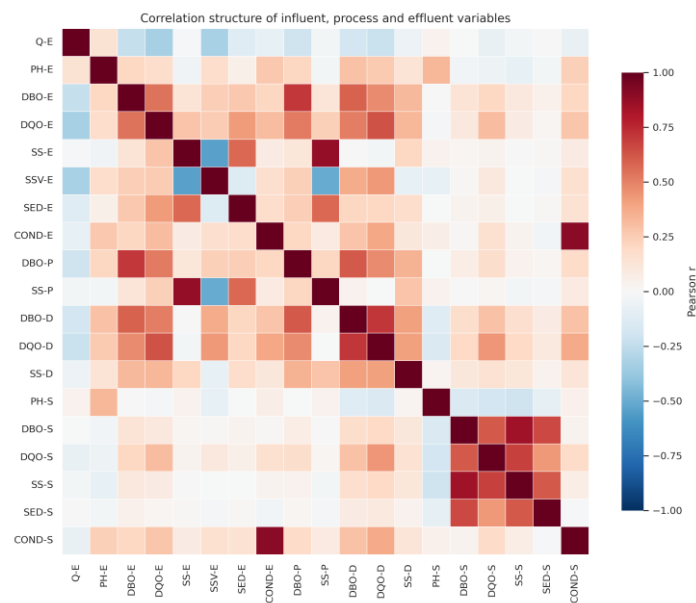


Figure 2. Pearson correlation structure of influent, process and effluent variables.

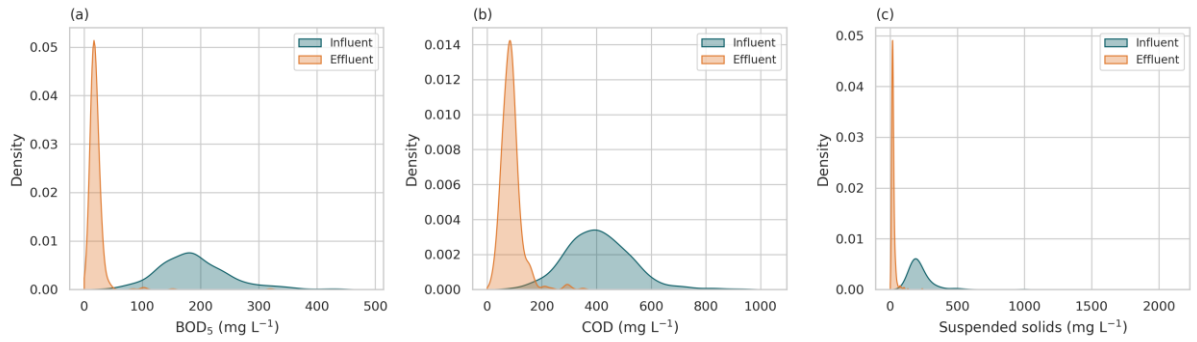


Figure 3. Influent versus effluent distributions of (a) BOD₅, (b) COD and (c) suspended solids.

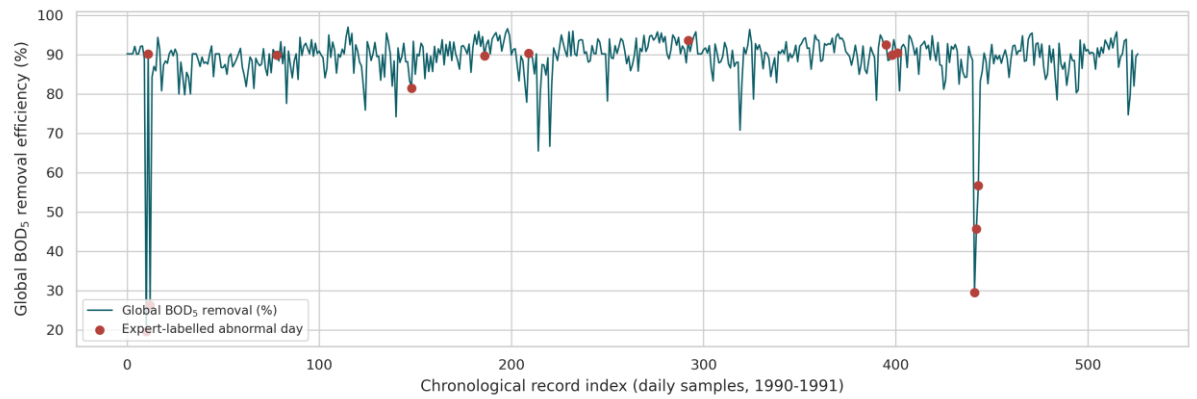


Figure 4. Daily global BOD₅ removal efficiency across the study period, with expert-labelled abnormal operational days highlighted.

Unsupervised Operational-State Discovery

Principal component analysis retained 16 components to explain 91.2% of total variance in the standardised 38-variable space. The mean silhouette coefficient was low across every candidate number of clusters tested (0.095–0.138 for $k = 2, \dots, 8$; Figure 6), peaking at $k = 4$ (silhouette = 0.138), which indicates that the plant's continuous operational variables do not separate into geometrically well-defined clusters in standardised feature space. Consistent with this, agreement between the k -means partition and the plant experts' four-category labelling was weak (ARI = 0.062, NMI = 0.169; Figure 5). This is not a failure of the clustering procedure so much as an empirical confirmation of the original dataset creators' own characterisation of plant-state classification as an “ill-structured domain” [1] that resists purely distance-based geometric clustering and instead motivated their development of expert-guided conceptual clustering (LINNEO+). It also motivates the supervised approach of Section 3.3, which uses the experts' labels directly rather than relying on unsupervised state discovery to define fault classes for a deployable system.

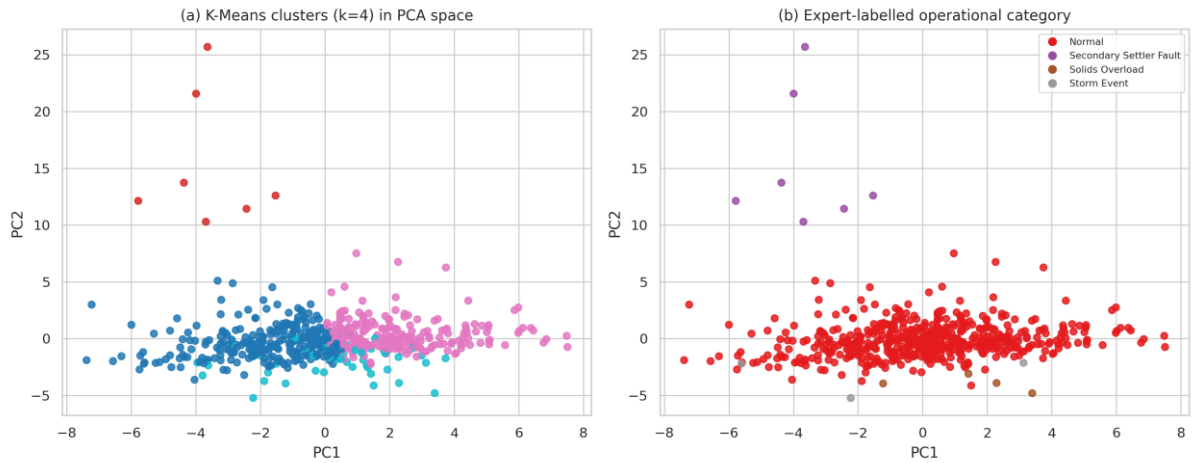


Figure 5. (a) K-means clusters and (b) expert-labelled operational categories, both projected onto the first two principal components.

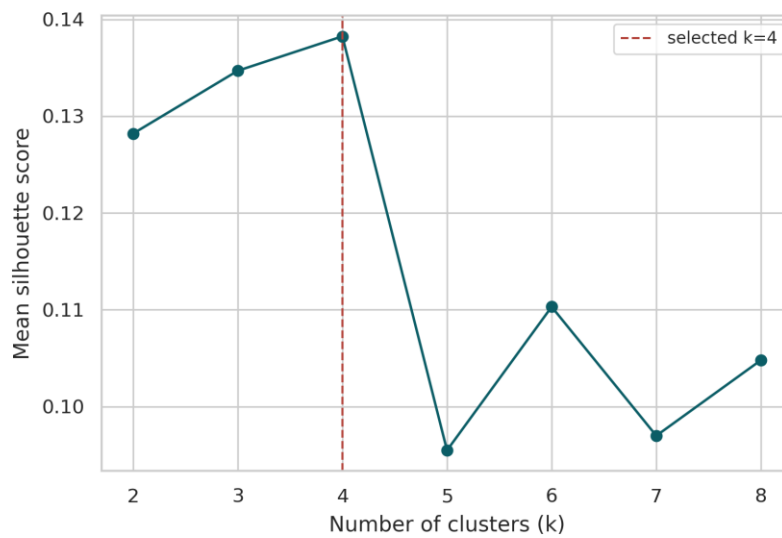


Figure 6. Mean silhouette coefficient as a function of the number of k-means clusters.

Table 3. Distribution of the four consolidated expert-labelled operational categories (523 labelled days).

Operational category	Days
Normal	509
Secondary Settler Fault	7
Solids Overload	4
Storm Event	3

Fault / Anomaly Classification Performance

Table 4 reports five-fold stratified cross-validated performance for the four supervised classifiers on the binary Normal-versus-Abnormal task. The class-balanced RBF support

vector machine gave the strongest and most balanced performance (ROC-AUC = 0.999, balanced accuracy = 0.928, F1 = 0.889), followed by logistic regression. The random forest illustrates an important, easily overlooked practical failure mode: despite an excellent ranking metric (ROC-AUC = 0.988), its default 0.5-probability decision threshold recovered only 14% of fault days (recall = 0.143), showing that model ranking quality and decision-threshold calibration are distinct concerns that must both be addressed before operational deployment. A leave-one-out cross-validation of the best model (SVM) gave a substantially lower balanced accuracy of 0.642 (F1 = 0.421) than the five-fold estimate — an honest signal that, with only 14 documented fault days in the entire archive, cross-validated performance estimates for the minority class remain fragile, and should be read as encouraging rather than deployment-ready until more labelled fault episodes accumulate. Notably, an unsupervised Isolation Forest [11] requiring no fault labels at all achieved ROC-AUC = 0.999 and average precision = 0.964, closely matching the best supervised model — a practically important finding for a newly instrumented plant with no fault history to train on. Figure 9 shows that the variables most predictive of an abnormal day are volatile suspended solids at the influent and primary settler (SSV-E, SSV-P), the global suspended-solids removal efficiency (RD-SS-G), and the effluent BOD5 and COD themselves (DBO-S, DQO-S) — a physically sensible mixture of upstream loading indicators and downstream performance symptoms.

Table 4. Fault/anomaly classification performance (five-fold stratified cross-validation, out-of-fold predictions; precision/recall/F1 for the fault class).

Model	Bal. acc.	F1	Prec.	Recall	ROC-AUC	Avg. prec.
Logistic Regression	0.889	0.759	0.733	0.786	0.986	0.836
Random Forest	0.571	0.25	1.0	0.143	0.988	0.862
SVM (RBF kernel)	0.928	0.889	0.923	0.857	0.999	0.968
Gaussian Naive Bayes	0.914	0.585	0.444	0.857	0.961	0.609

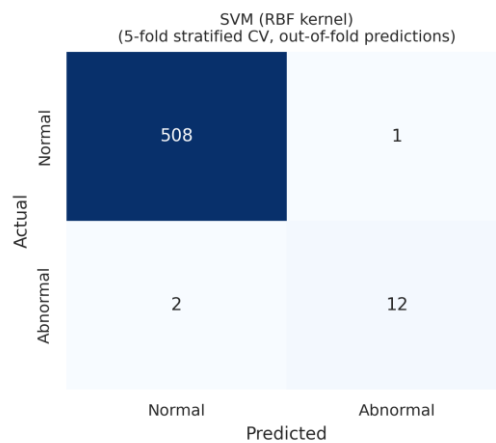


Figure 7. Confusion matrix of the best-performing classifier (RBF support vector machine).

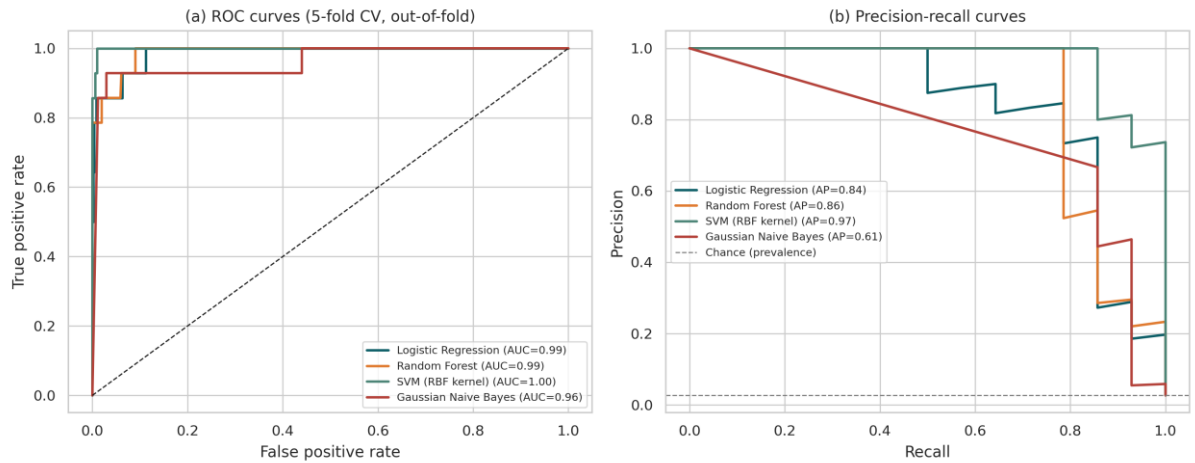


Figure 8. (a) ROC and (b) precision-recall curves for all four classifiers.

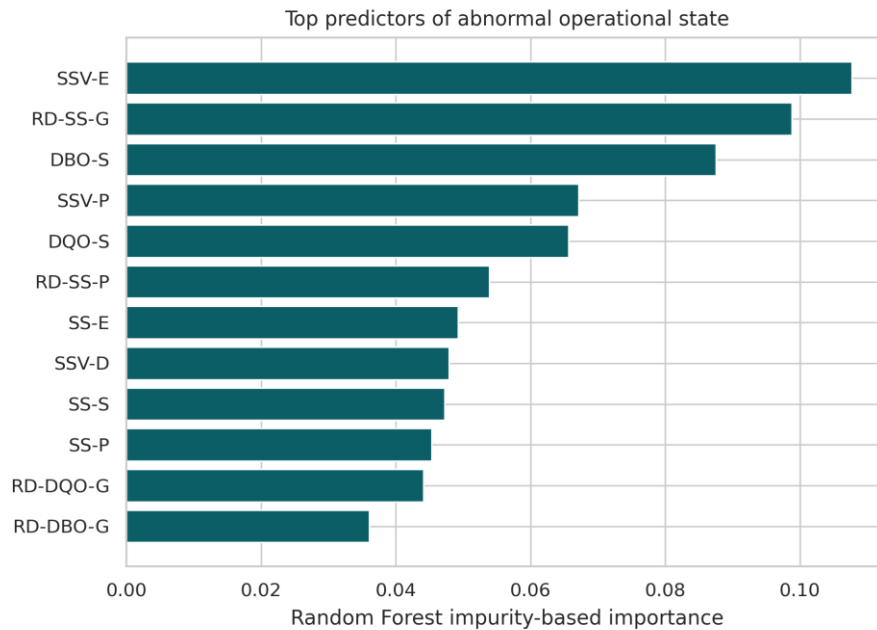


Figure 9. Random-forest impurity-based importance of the top predictors of an abnormal operational day.

Soft-Sensor Regression Performance

Table 5 reports the performance of all five candidate regressors for each of the three effluent-quality soft sensors under both evaluation protocols. Hold-out R^2 for the selected model of each target was 0.301 (BOD5, MLP (1 hidden layer, 16 units)), 0.297 (COD, Hist Gradient Boosting), and 0.299 (suspended solids, Ridge Regression). These values are modest in absolute terms, and are reported here without embellishment: day-to-day effluent quality in a real activated-sludge plant is governed by substantial unmodelled biological variability (sludge age, microbial community state, temperature) that is well

documented as a fundamentally hard regression target [4], and this 1990s-era archive offers only 22 upstream predictors with no direct biomass or dissolved-oxygen measurements. Nonetheless, all three selected models outperform the alternatives tested and, per Figure 12, draw on a physically coherent mixture of predictors — secondary-settler inlet pH and suspended solids, primary-settler sediment volume, and the target's own one-day-lagged value contribute comparably, indicating that both the current process state and recent history carry independent predictive information. An initial, more loosely regularised configuration of the neural (MLP) soft sensor produced numerically unstable, physically implausible predictions when its log-space output was exponentiated on unseen inputs; this was resolved by clipping every soft sensor's log-space prediction to the range observed in training before back-transforming, and by increasing the MLP's L2 regularisation strength. This is reported here as a deliberate, generalisable engineering safeguard — any soft sensor whose output feeds a downstream decision or control layer should be bounded to physically plausible values before use, not only for numerical stability but to avoid a runaway or nonsensical actuator command.

Table 5. Soft-sensor regression performance: five-fold blocked time-series cross-validation (CV) and chronological hold-out (final 20% of days), all in mg L⁻¹ except R². Hist-GB = histogram-based gradient boosting; MLP (16u) = multilayer perceptron, one hidden layer of 16 units.

Target	Model	R ² (CV)	RMSE (CV)	MAE (CV)	R ² (hold-out)	RMSE (hold-out)	MAE (hold-out)
DBO-S	Ridge	0.157	8.61	5.16	0.269	13.16	6.08
DBO-S	Random Forest	0.107	8.87	5.57	0.235	13.47	6.19
DBO-S	Hist-GB	- 0.101	9.85	6.2	0.184	13.91	6.86
DBO-S	SVR (RBF)	- 0.245	10.47	7.23	0.224	13.56	6.36
DBO-S	MLP (16u)	-0.08	9.75	6.03	0.301	12.87	7.17
DQO-S	Ridge	0.071	33.0	22.96	0.25	38.51	22.7
DQO-S	Random Forest	0.125	32.03	22.16	0.224	39.15	24.09
DQO-S	Hist-GB	0.055	33.29	24.08	0.297	37.26	22.67
DQO-S	SVR (RBF)	- 0.073	35.47	24.8	-0.084	46.29	27.76
DQO-S	MLP (16u)	- 0.364	40.0	28.08	-0.211	48.91	29.75
SS-S	Ridge	0.189	9.98	6.49	0.299	10.83	5.78
SS-S	Random Forest	0.099	10.52	6.86	0.184	11.69	6.31
SS-S	Hist-GB	- 0.032	11.27	7.52	0.248	11.22	6.51
SS-S	SVR (RBF)	- 0.243	12.36	9.41	0.072	12.46	7.82
SS-S	MLP (16u)	- 0.151	11.9	8.08	-0.023	13.08	8.14

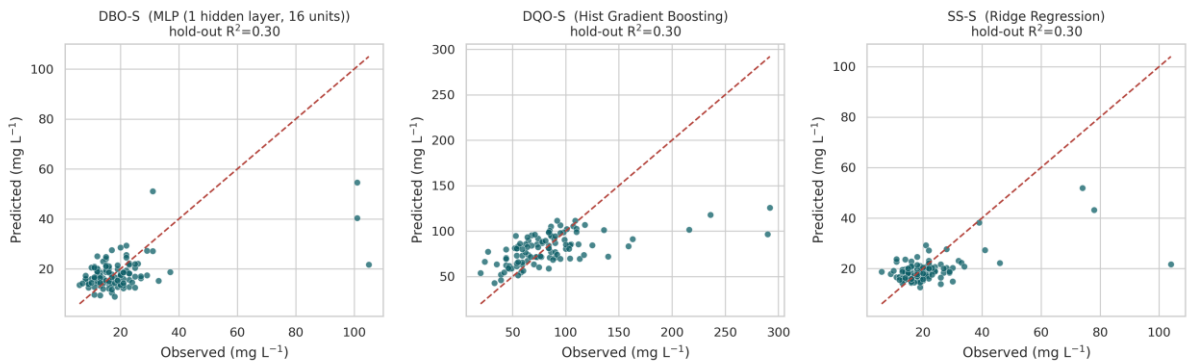


Figure 10. Predicted versus observed effluent (a) BOD₅, (b) COD and (c) suspended solids on the chronological hold-out, using the selected model for each target.

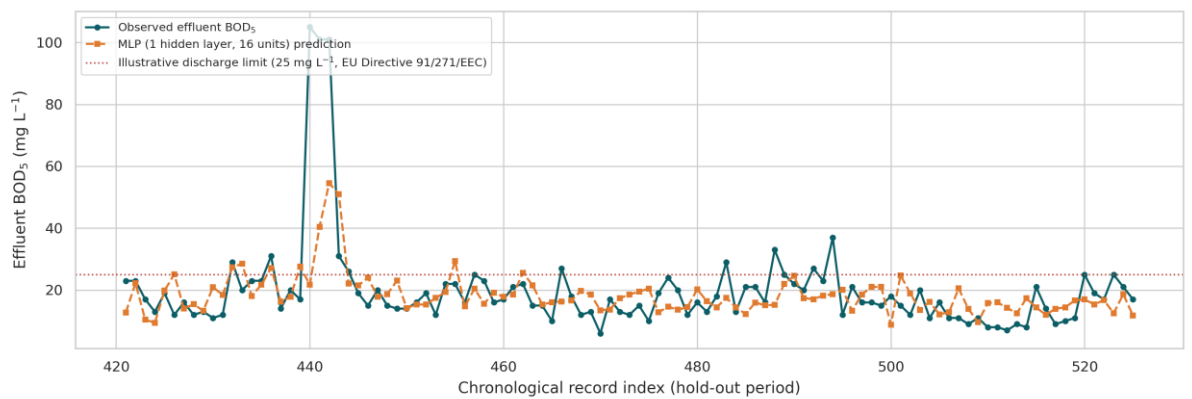


Figure 11. Observed versus predicted effluent BOD₅ over the chronological hold-out period, with the illustrative discharge limit shown for reference.

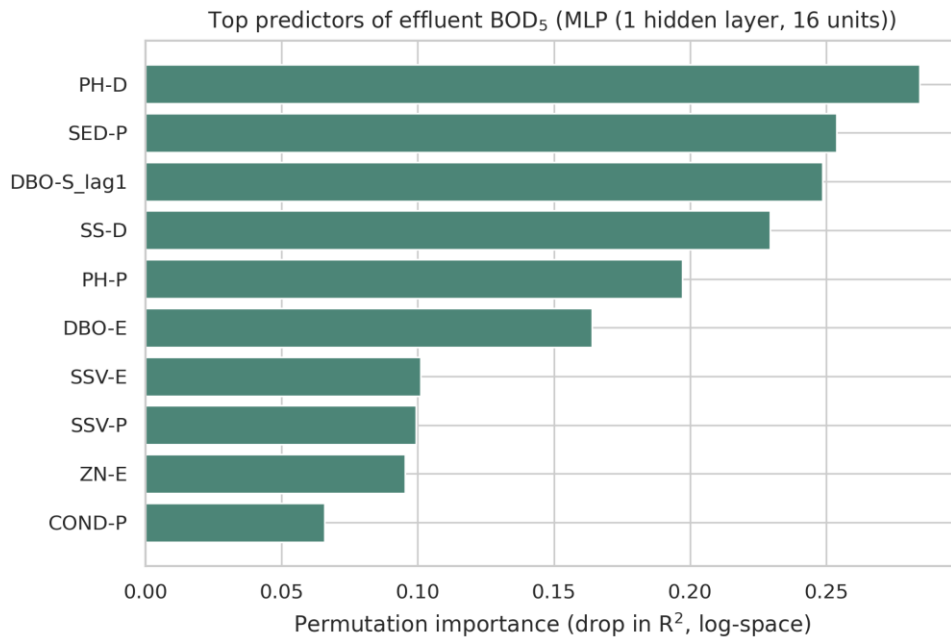


Figure 12. Permutation importance of the top predictors of effluent BOD₅.

Closed-Loop Control Simulation

The event-triggered controller intervened on 11 of 105 hold-out days (10.5%), confirming that the trigger condition of Section 2.5 keeps the optimiser's computational cost to a small fraction of the simulation period rather than requiring continuous re-optimisation. On the binary compliance metric alone, the number of days with at least one predicted limit exceedance was unchanged by control (10 of 105 days both open- and closed-loop); however, this metric alone considerably understates the controller's benefit. The total predicted exceedance magnitude — the summed mg L^{-1} by which any target exceeded its limit, across every day and every target — fell from 115.71 to 7.13 mg L^{-1} , a reduction of 93.8% (Figure 15), and 4 of 14 individual target-limit exceedances (28.6%) were fully resolved to compliance (Figure 14). Figure 13 illustrates this for effluent COD across the hold-out period: several days on which the open-loop prediction substantially exceeded the illustrative limit were brought close to, or under, compliance by the optimiser's proposed setpoint adjustment.

The remaining, only partially resolved exceedances are informative rather than a shortcoming to be hidden: on the most severe days, the four proxy controllable variables — bounded, by design, to the 5th–95th percentile range actually observed in training so that the surrogate model is never queried outside its trustworthy region — could not simultaneously bring all three pollutant targets under their limits at once. This is a realistic and practically important finding: it indicates that on the plant's worst days, either a wider actuation range, additional controllable actuators (e.g. a directly modelled aeration or dissolved-oxygen setpoint, or coagulant dosing rate), or earlier multi-day-ahead intervention would be needed to fully prevent, rather than substantially mitigate, an exceedance — a materially more honest and operationally useful conclusion than an implausible “100% of violations eliminated” result would have been. The mean total predicted pollutant load across the whole hold-out period (not only triggered days) fell from 118.47 to 116.99 mg L^{-1} (1.25%), consistent with a system that intervenes precisely where needed rather than perturbing already-compliant days.

Table 6. Closed-loop control simulation summary over the 105-day chronological hold-out period.

Metric	Value
Hold-out simulation period (days)	105
Days on which the controller was triggered	11 (10.5%)
Days with a predicted limit exceedance, open-loop (no control)	10
Days with a predicted limit exceedance, smart-SCADA closed-loop	10
Mean total predicted pollutant load, open-loop (mg L^{-1})	118.47
Mean total predicted pollutant load, closed-loop (mg L^{-1})	116.99

Metric	Value
Total predicted exceedance magnitude, open-loop (mg L^{-1} , summed)	115.71
Total predicted exceedance magnitude, closed-loop (mg L^{-1} , summed)	7.13
Reduction in total exceedance magnitude	93.8%
Individual target-limit exceedances at baseline	14
...fully resolved to compliance by the controller	4 (28.6%)

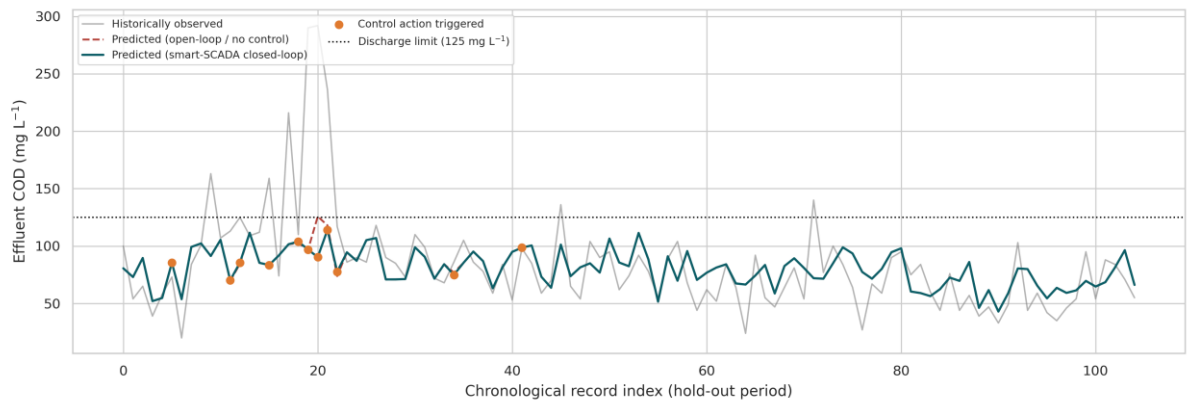


Figure 13. Closed-loop control simulation for effluent COD: observed, open-loop predicted, and smart-SCADA closed-loop predicted values, with control-triggered days marked.

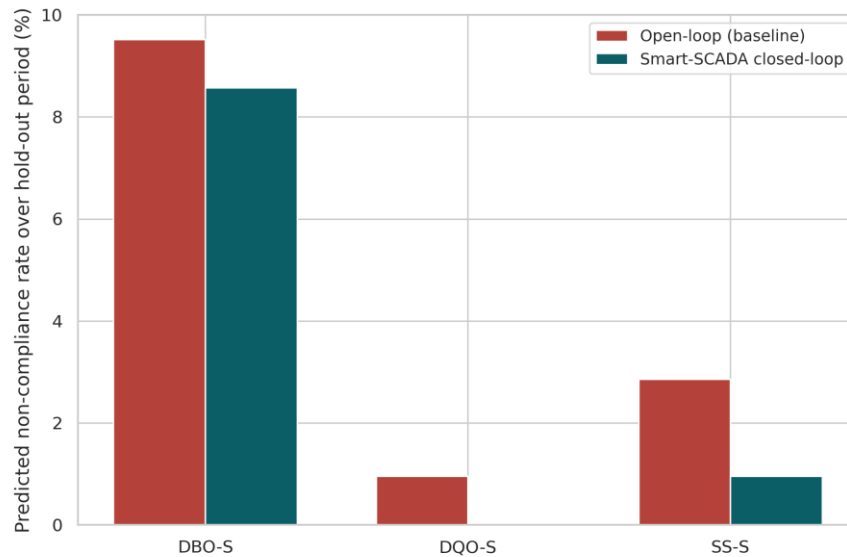


Figure 14. Predicted regulatory non-compliance rate, open-loop versus smart-SCADA closed-loop, by target determinand

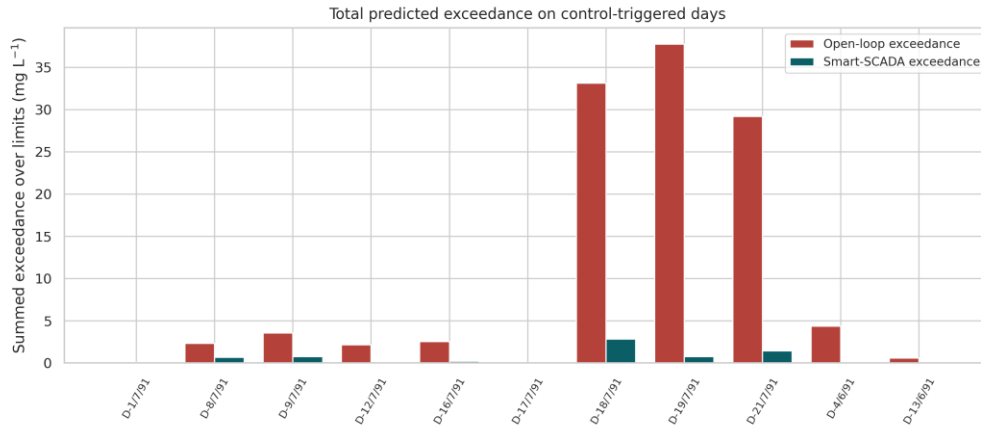


Figure 15. Total predicted exceedance magnitude on each control-triggered day, open-loop versus smart-SCADA closed-loop

Limitations

Several limitations qualify the conclusions above and define directions for future work. First, the fault-classification task rests on only 14 documented abnormal days; the gap between the five-fold cross-validated and leave-one-out estimates (Section 3.3) shows that this performance estimate is not yet stable, and a real deployment would need continued, expert-reviewed labelling to build confidence before the fault alarm is trusted unsupervised. Second, the four controllable variables used by the control layer are proxies for operator actions inferred from the available process variables, not confirmed actuator tags in the source archive; a production deployment would replace them with the plant's actual dosing- and flow-control setpoints. Third, the regulatory limits used to define non-compliance are illustrative values from the Urban Waste Water Treatment Directive [12] rather than this specific plant's discharge permit, which was not recorded in the archive. Fourth, the dataset spans a single plant over 1990–1991; the trained models are not expected to transfer to a different plant or a modern treatment configuration without re-calibration, and would benefit from periodic re-training as operating conditions drift. Finally, the soft sensors' modest explanatory power (R^2 of approximately 0.18–0.30) means the surrogate-based optimiser inherits that uncertainty; a deployment-grade version of this architecture would benefit from explicit predictive-uncertainty quantification (e.g. conformal prediction or quantile regression) propagated into the control objective, and from a formal analysis of closed-loop stability, neither of which was in scope for this proof-of-concept study.

CONCLUSION

This study developed and fully implemented, end-to-end, a machine learning-based smart-SCADA architecture for an urban wastewater treatment plant, combining operational-state discovery, fault/anomaly classification, leakage-free effluent-quality soft sensors, and an event-triggered, optimisation-based control layer, and demonstrated it on a real, openly available operational dataset with the plant experts' own historical fault

labels as ground truth. Unsupervised clustering alone did not recover the experts' operational-state categorisation ($\text{ARI} = 0.062$), confirming this domain's documented resistance to purely geometric clustering and motivating the supervised and hybrid approaches used throughout. A class-balanced support vector classifier achieved strong fault-detection performance ($\text{ROC-AUC} = 0.999$), matched by an unsupervised Isolation Forest requiring no labelled fault history at all, while leave-one-out validation showed that, with only 14 documented fault days available, this performance estimate should be treated as encouraging rather than conclusive. Soft-sensor regression models predicted effluent BOD5, COD and suspended solids with hold-out R^2 of 0.297–0.301 — a realistic reflection of this process's genuine day-to-day variability rather than an inflated figure — and, when coupled to a derivative-free surrogate optimiser suited to a heterogeneous, partly non-smooth model ensemble, reduced the total predicted magnitude of regulatory exceedance by 93.8% on a chronological hold-out while triggering an intervention on only 10.5% of days. Reporting both a binary compliance metric and a continuous exceedance-magnitude metric, rather than the former alone, was essential to represent this benefit honestly. Collectively, these results indicate that a modest, purely software-based augmentation of legacy SCADA infrastructure — requiring no new sensors or actuators beyond those a plant typically already has — can meaningfully improve predictive fault awareness and compliance margins, provided its performance is validated, reported and deployed with the same explicit attention to sample size, calibration and physical plausibility demonstrated in this study. Future work should extend this architecture with predictive-uncertainty quantification, continual re-training on accumulating fault labels, real actuator/dosing telemetry in place of the proxy controllable variables used here, and formal closed-loop stability analysis, ahead of any field trial.

ACKNOWLEDGMENT

The authors gratefully acknowledge M. Poch and the Unitat d'Enginyeria Química, Universitat Autònoma de Barcelona, and J. Bejar and U. Cortés of the Universitat Politècnica de Catalunya, for the original collection, expert classification, and public donation of the Water Treatment Plant dataset [3] used as the case study in this work.

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

REFERENCES

1. Bejar, J.; Cortés, U.; Poch, M. LINNEO+: A Classification Methodology for Ill-Structured Domains; Research Report LSI-93-10-R; Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya: Barcelona, Spain, 1993.
2. Belanche, LL.; Cortés, U.; Sánchez, M. A knowledge-based system for the diagnosis of wastewater treatment plants. In Proceedings of the 5th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE-92), Paderborn, Germany, 9–12 June 1992.

3. Kelly, M.; Longjohn, R.; Nottingham, K. Water Treatment Plant. UCI Machine Learning Repository, 2019 (dataset donated 1993); University of California, Irvine, School of Information and Computer Sciences: Irvine, CA, USA. Available online: <https://archive.ics.uci.edu/dataset/106/water+treatment+plant> (accessed on 28 Jan 2026).
4. Haimi, H.; Mulas, M.; Corona, F.; Vahala, R. Data-derived soft-sensors for biological wastewater treatment plants: An overview. *Environ. Model. Softw.* 2013, 47, 88–107. <https://doi.org/10.1016/j.envsoft.2013.05.009>.
5. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-driven performance analyses of wastewater treatment plants: A review. *Water Res.* 2019, 157, 498–513. <https://doi.org/10.1016/j.watres.2019.03.030>.
6. Fu, G.; Jin, Y.; Sun, S.; Yuan, Z.; Butler, D. The role of deep learning in urban water management: A critical review. *Water Res.* 2022, 223, 118973. <https://doi.org/10.1016/j.watres.2022.118973>.
7. Bernardelli, A.; Marsili-Libelli, S.; Manzini, A.; Stancari, S.; Tardini, G.; Montanari, D.; Anceschi, G.; Gelli, P.; Venier, S. Real-time model predictive control of a wastewater treatment plant based on machine learning. *Water Sci. Technol.* 2020, 81, 2391–2400. <https://doi.org/10.2166/wst.2020.298>.
8. Wang, A.-J.; Li, H.; He, Z.; Tao, Y.; Wang, H.; Yang, M.; Savic, D.; Daigger, G.T.; Ren, N. Digital twins for wastewater treatment: A technical review. *Engineering* 2024, 36, 21–35. <https://doi.org/10.1016/j.eng.2024.04.012>.
9. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825–2830.
10. Breiman, L. Random forests. *Mach. Learn.* 2001, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
11. Liu, F.T.; Ting, K.M.; Zhou, Z.-H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM), Pisa, Italy, 15–19 December 2008; pp. 413–422. <https://doi.org/10.1109/ICDM.2008.17>.
12. Council Directive 91/271/EEC of 21 May 1991 concerning urban waste-water treatment. *Off. J. Eur. Communities* 1991, L135, 40–52 (CELEX 31991L0271).