Proceedings Paper:
Ahmed, H. (2026) Unsupervised anomaly detection in water distribution SCADA system using Gaussian mixture models with GMM-lime explainability. In: 2026 1st International Conference on Human Centric Artificial Intelligence (ICHCAI). 2026 1st International Conference on Human Centric Artificial Intelligence (ICHCAI), 27-28 May 2026, Halden, Norway. Institute of Electrical and Electronics Engineers (IEEE). ISBN: 9798331551193.

https://doi.org/10.1109/ichcai70183.2026.11607578

Unsupervised Anomaly Detection in Water Distribution SCADA System Using Gaussian Mixture Models with GMM-LIME Explainability

Hafiz Ahmed*†
*University of Sheffield, Sheffield S10 2TN, United Kingdom (E-mail: hafiz.h.ahmed@ieee.org)
†Autonex Systems Limited, Coventry CV1 2NT, United Kingdom

***Abstract*—Reliable fault detection in existing benchmark water distribution systems (WDS) dataset is complicated by severe class imbalance, high sensor dimensionality, and significant distribution shift between historical fault signatures and faults observed during operation. This paper presents a three-stage framework that addresses each challenge in turn. Six supervised classifiers are evaluated on a 43-sensor WDS dataset and shown to perform poorly on fault-sensitive metrics, with the best F1 score of only 0.41 (Random Forest). A diagnostic analysis using Mahalanobis distance shows a 30-fold gap between training and test fault distributions, explaining why supervised training fails. Recursive feature elimination (RFE) driven by a Random Forest then reduces the sensor set from 43 to 15, retaining the sensors with the highest joint discriminative value. A Gaussian mixture model (GMM) trained exclusively on normal data subsequently achieves F1 = 0.533, AUC-ROC = 0.853, and AUPR = 0.582, outperforming all supervised baselines. For interpretability, the GMM-LIME (local interpretable model-agnostic explanations) framework replaces standard Gaussian perturbation with GMM-structured sampling, improving average Jaccard similarity from 0.544 to 0.684 for fault instances across five repeated runs. The proposed system offers a practical, interpretable route to assist human operators in WDS SCADA fault monitoring where labelled fault data are scarce and non-representative.**

***Index Terms*—Water Distribution Systems, SCADA, Anomaly Detection, XAI, Gaussian Mixture Model.**

I. Introduction

Water distribution systems (WDS) are critical infrastructure whose failures, whether from aging pipes or cyber-physical attacks, carry consequences well beyond immediate repair costs [1], [2]. The BATADAL (BATtle of the Attack Detection ALgorithms) dataset used in this study captures adversarial WDS scenarios, including deception attacks, replay attacks, and malicious actuator manipulation, logged over a long-period SCADA simulation [1]. Sensor networks now standard in modern WDS generate rich multivariate time-series data that are, in principle, well suited to data-driven detection. Two structural problems, however, make this harder in practice.

The first is class imbalance. Faults account for just 3.8% of training samples here, a skew that is typical across WDS datasets [3]. Classifiers trained on such data gravitate toward majority-class prediction. Balancing techniques such as synthetic minority oversampling (SMOTE) [4] partially address this, but oversampling before the train-test split can introduce leakage that inflates reported metrics without genuine detection improvement [5]. The second problem is distribution shift. Fault signatures at deployment frequently bear little resemblance to training faults, because network conditions, demand patterns, and attack types all evolve. We show this empirically through a Mahalanobis distance analysis that reveals a 30-fold gap between mean training and test fault distances from the normal centroid.

Interpretability makes the situation further challenging. A binary alarm from a black-box model tells field engineers nothing about which sensors triggered it or where the fault lies [6]. Standard XAI methods such as SHAP (SHapley Additive exPlanations) [7] and LIME [8] were designed for supervised differentiable models, and applying them to an unsupervised detector introduces a faithfulness problem. Authors in [9] addressed LIME instability by replacing its Gaussian perturbation step with GMM-structured sampling, substantially improving explanation consistency on human gait data. We extend their GMM-LIME framework to WDS sensor data, where physical coupling between pressure and flow measurements makes covariance-respecting perturbation especially important.

This paper makes four contributions: (i) a systematic benchmark of six supervised classifiers with Mahalanobis-based diagnostic analysis of their failure; (ii) Random Forest-driven recursive feature elimination reducing 43 sensors to a principled 15-sensor subset; (iii) a GMM anomaly detector trained on normal data only, achieving AUC-ROC = 0.853 and F1 = 0.533, outperforming all supervised baselines; and (iv) a GMM-LIME adaptation that improves average Jaccard similarity from 0.544 to 0.684 for fault instance explanations across five repeated runs.

A. Related Work

1) Anomaly Detection in WDS: Significant research effort has gone into anomaly and attack detection in WDS over the past decade. Early approaches relied on statistical methods and simple threshold rules, but the availability of the BATADAL benchmark dataset [1] created a shared evaluation platform that stimulated a wider range of machine learning approaches. The dataset simulates the C-Town network, a mid-size WDS.

979-8-3315-5118-6/26/$31.00

Supervised classifiers have been applied extensively to this benchmark [10]. Reference [11] proposed an adaptive neural tree model for attack detection, reporting an accuracy of 0.836 on BATADAL. Reference [12] took a dual-stage approach that combined classification and regression: classification models directly flag attacks while regression models predict expected tank water levels from pump readings, and large deviations are treated as anomalies. Their combined method achieved a classification accuracy of 0.89. Authors in [13] recognized the imbalanced nature of BATADAL and applied a one-class support vector data description classifier trained only on normal samples. The design is well motivated by the class skew, though the authors noted it may struggle to generalize to previously unseen fault types over time. Deep learning has also been explored. Authors in [14] developed an LSTM-based detector and found that both data balancing and sequence-to-point learning were needed to push the F1 score above 0.78. Reference [15] used a recurrent neural network to model normal system behavior in a water treatment testbed, flagging deviations with the Cumulative Sum method. Authors in [16] proposed a hybrid dynamic graph neural network for real-time anomaly detection, showing that graph-based representations of sensor relationships improve detection of both operational faults and cyber-physical attacks. Reference [17] similarly demonstrated that graph convolutional networks detect burst events from pressure and flow measurements more effectively than methods treating sensors as independent channels.

One study reporting very high performance was the hybrid deep hierarchical clustering approach in [5], which grouped sensor readings by feature similarity before training a deep network on the clustered data, reaching an F1 score of 0.915 on BATADAL. The authors themselves flagged that this figure may be inflated by data leakage from SMOTE balancing applied before the train-test split. This concern runs through much of the WDS anomaly detection literature and motivates our decision to avoid oversampling entirely.

2) Sensor Dimensionality Reduction: The BATADAL dataset contains 43 SCADA variables. Using all of them raises computational overhead and can introduce noise that degrades model performance. Authors in [18] used graph neural networks to identify the five most informative pump unit locations out of eleven, showing that a small, well-chosen sensor set can match or exceed the performance of the full sensor array. Their key insight was that sensors should be selected not only for individual relevance but for spatial coverage and complementarity across the network. Reference [13] performed feature selection based on the physical topology of the WDS, prioritizing sensors whose readings are physically coupled to likely anomaly propagation paths. These approaches align with broader findings in the sensor placement literature [19]: redundant or co-located sensors provide diminishing returns for network coverage.

Our approach differs in that it uses recursive feature elimination driven by Random Forest out-of-bag (OOB) permutation importance. This iterative, data-driven procedure identifies the sensor subset that jointly maximizes discrimination between normal and anomalous conditions in the observed training data, without requiring any prior knowledge of network topology.

3) Explainable AI for Industrial Monitoring: XAI for industrial monitoring generally divide into two families. Global methods characterize how a model behaves across a population of inputs, while local methods explain a specific prediction for a specific instance. Operator-facing applications typically require local explanations, since field engineers respond to individual alarms rather than population-level summaries [6].

LIME [8] is among the most widely used local explainability methods. It approximates the decision boundary of any black-box model near a target instance with a weighted linear surrogate trained on perturbed copies of that instance. LIME is model-agnostic and produces human-readable feature importance scores. Its main weakness is instability across repeated runs [20]. Because the perturbation step draws independent Gaussian samples, the same instance can yield quite different explanations on different occasions. In systems with physically correlated sensors, independent perturbation also generates unrealistic sensor combinations that the black-box model responds to unpredictably, further reducing explanation faithfulness. SHAP [7] provides attribution values based on Shapley game theory and is generally more stable than LIME, but its computational demands are considerably higher. For a surrogate-free method supporting real-time operator response, LIME's speed advantage remains practically important.

Authors in [9] identified LIME instability as a key clinical limitation in human gait analysis and proposed replacing the Gaussian perturbation step with GMM-structured sampling. A GMM fitted to training data generates perturbed samples that respect the covariance structure of real sensor readings, making the black-box model's responses more consistent and the resulting linear fit more reproducible across runs. Validation using Jaccard similarity and the overlap coefficient confirmed consistent improvement over vanilla LIME on two gait datasets. We adopt this GMM-LIME framework and apply it to WDS sensor data, where the physical coupling between pressure, flow, and tank level readings makes covariance-aware perturbation especially well motivated.

B. Research Gap

Several gaps motivate the present study. First, most supervised approaches applied to BATADAL treat the training-to-test distributional gap as a nuisance rather than characterizing it quantitatively, so poor test-set performance is reported but not explained. Second, unsupervised methods have been proposed but are seldom accompanied by the diagnostic analysis needed to confirm that observed performance is not an artefact of data leakage. Third, XAI for WDS anomaly detection remains largely unexplored. The majority of published work produces a detection decision with no accompanying explanation, and where XAI is applied it is typically SHAP on a supervised model, which does not address settings where labelled fault data are absent or non-representative. No existing work on BATADAL applies GMM-structured perturbation to improve

TABLE I
BATADL Dataset Summary Statistics.

Split	Total	Normal	Fault	Fault Rate	Imbalance Ratio
Training	12,938	12,446	492	3.8%	25.3:1
Test	2,089	1,682	407	19.5%	4.1:1

local explanation stability. The present paper addresses all three gaps within a single end-to-end framework. Additionally, beyond predictive performance, we frame anomaly detection in WDS SCADA as a human-centred decision support problem, where the practical value of an alarm depends on whether it can be interpreted, trusted, and acted upon by operators in time-critical settings.

II. Methodology

A. Dataset Description

This study utilizes the BATADAL dataset[1] [1], a benchmark for anomaly detection in SCADA systems generated via EPANET hydraulic simulation of the mid-sized, real-life C-Town water distribution network. The network comprises 388 junctions, 429 pipes, 7 storage tanks, 11 pumps, and 5 valves, all monitored through 9 PLCs. The dataset consists of 43 sensors sampled at hourly intervals, capturing nodal pressures, tank water levels, flow rates, and actuator states. The data is partitioned into a training set of 12,938 samples and a test set of 2,089 samples, with binary labels distinguishing normal operation (Class 0) from unspecified anomalies (Class 1). As summarized in Table I, a significant class imbalance exists in the training set (3.8% fault prevalence) compared to the test set ($\sim$ 20% fault rate); this discrepancy signals the distribution shift problem examined in Sec. II-D.

B. Preprocessing

All sensor readings are z-score normalized using only training-set statistics (with mean μ_j and standard deviation σ_j for feature j). Test samples are normalized with the same μ_j, σ_j so that no test information leaks into the scaling step. SMOTE is applied to the training set only, targeting a 20% minority class ratio with $k = 5$ nearest neighbours.

C. Supervised Baseline Classifiers

Six classifiers are trained on the SMOTE balanced training data: Logistic Regression (LR, ridge penalty $\lambda = 10^{-4}$), k-Nearest Neighbours (kNN, $k = 7$), Decision Tree (DT, max 20 splits), SVM with RBF kernel (SVM-RBF, cost $C = \mathrm{IR}$), Random Forest (RF, 200 trees), and RUSBoost (150 cycles, learning rate 0.1). For each classifier the decision threshold is optimized on training scores to maximize F1:

$$\tau^* = \arg\max_{\tau} \frac{2 \cdot P(\tau) \cdot R(\tau)}{P(\tau) + R(\tau)}, \tag{1}$$

where P is precision and R is recall. This step is critical under class imbalance, where the default 0.5 threshold systematically underestimates fault probability.

D. Diagnostic Analysis

Three complementary tests examine why supervised classifiers struggle on this dataset.

- **Feature Discriminability (Cohen's d):** For each sensor j, Cohen's d (d_j) is calculated between the training fault and normal subsets. Sensors with $d_j > 0.2$ carry meaningful discriminative signal.
- **Distribution Shift (Kolmogorov-Smirnov Test):** A two-sample KS test compares training fault versus test fault distributions for each sensor. Significant results ($p < 0.05$) indicate that the fault patterns learned during training have shifted at deployment.
- **Mahalanobis Distance:** The squared Mahalanobis distance from the normal training centroid $\boldsymbol{\mu}_N$ quantifies how far each sample sits from the normal operating region: Comparing mean distances of training and test faults quantifies the distribution shift in a single number.

E. RF-RFE Sensor Selection

Rather than ranking sensors by marginal importance scores, we use recursive feature elimination to find a jointly optimal subset. RFE is driven by a random forest (RF), which was chosen for two reasons. First, the RF OOB permutation importance [21] is a well-established, low-bias estimator of feature relevance that does not require a held-out validation set. Second, RF achieved the best F1/AUC balance among all supervised classifiers evaluated in Section II-C. The other classifiers are unsuitable as RFE rankers: LR coefficients are scale-dependent and conflated with regularization, kNN provides no feature importance mechanism, single decision trees exhibit high-variance importance, and SVM kernel methods do not yield per-feature attributions without costly approximations.

At each step, a fresh RF is trained on the current feature subset and OOB permutation importance is recorded: The sensor with minimum permutation importance is removed. This continues until 15 sensors remain. The choice of 15 sensors as the stopping point is motivated by the elimination curve in Fig. 5. Between 43 and approximately 22 sensors, the RF AUC-ROC remains relatively stable, reflecting high redundancy in the full sensor set. Below 22 sensors the curve steepens, indicating that sensors with genuine informational value are beginning to be removed. Stopping at 15 captures a 65% reduction in sensor count while remaining in a zone where the retained sensors carry complementary, non-redundant signal. Notably, the purpose of this sensor subset is not to maximize the performance of the supervised RF on 15 sensors, but to identify the most informative sensors for the unsupervised GMM. The GMM uses these sensors to model the normal operating manifold in a reduced feature space; removing noisy, redundant dimensions actually benefits the GMM by making the normal distribution more compact

[1] https://www.kaggle.com/datasets/minhbtnguyen/batadal-a-dataset-for-cyber-attack-detection

and easier to estimate. This is confirmed by the final GMM results: AUC-ROC = 0.853 on the 15-sensor subset substantially exceeds RF-43's AUC-ROC = 0.754 on all 43 sensors.

F. Proposed GMM Anomaly Detector

Figure 1 shows the complete proposed pipeline. Our main assumption is that while fault signatures vary unpredictably between data collection periods, normal operation is comparatively stable. By modeling only the normal regime, the detector generalizes to any fault type. The detector works through the following steps:

1) **Temporal Feature Engineering:** Raw 15-sensor readings are augmented with rolling mean and standard deviation over window sizes $w \in \{5, 15, 30\}$ samples, yielding a feature vector of dimension $15 \times (1 + 2 \times 3) = 105$ per time step. All normalization parameters are estimated from normal training samples only.
2) **PCA Projection:** PCA is applied to the normalized temporal features of normal training samples. The projection matrix $\mathbf{W} \in \mathbb{R}^{105 \times C}$ retains the C components needed to explain 95% of normal variance:
$$\mathbf{Z} = \mathbf{X}_{\text{feat}}\,\mathbf{W}. \tag{2}$$
3) **GMM on Normal PCA Scores:** A GMM is fitted to the projected normal training scores $\mathbf{Z}$:
$$p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k\, \mathcal{N}(\mathbf{z};\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k), \tag{3}$$
with mixing weights π_k, means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ estimated via the expectation-maximization algorithm [22]. The number of components K is selected by the Bayesian information criterion (BIC):
$$K^* = \arg\min_K \left[-2 \ln \hat{L}(K) + p_K \ln n\right], \tag{4}$$
where $\hat{L}(K)$ is the maximized likelihood, p_K is the number of free parameters, and n is the number of normal training samples.
4) **Anomaly Scoring and Normalization:** Each sample receives a score equal to its negative log-likelihood under the fitted GMM:
$$s(\mathbf{x}) = -\ln \sum_{k=1}^{K} \pi_k\, \mathcal{N}(\mathbf{z}(\mathbf{x});\, \boldsymbol{\mu}_k,\, \boldsymbol{\Sigma}_k), \tag{5}$$
where $\mathbf{z}(\mathbf{x}) = \tilde{\mathbf{x}}_{\text{feat}}\mathbf{W}$. Scores are normalized using IQR-based robust scaling related to normal training scores:
$$\hat{s}(\mathbf{x}) = \frac{s(\mathbf{x}) - \text{med}(s_N)}{\text{IQR}(s_N)}. \tag{6}$$
5) **Threshold Optimization:** The detection threshold τ^* is chosen on imbalanced training scores by maximizing F1 subject to a minimum recall constraint:
$$\tau^* = \arg\max_{\tau}\ F_1(\tau) \quad \text{s.t.} \quad \text{Recall}(\tau) \geq 0.25. \tag{7}$$
The recall floor prevents the optimizer from collapsing to a high-precision but low-recall solution.

Algorithm 1 summarizes the complete procedure.

Algorithm 1 Proposed GMM Anomaly Detector

Require: Training data $\mathcal{D}^{\text{tr}}$, test data $\mathcal{D}^{\text{te}}$, RFE feature set $\mathcal{F}$, window sizes $\mathcal{W}$
Ensure: Anomaly predictions $\hat{y}$ for each test sample
Training phase
Normalize all data using training statistics
Extract temporal features using $\mathcal{W}$ on the $|\mathcal{F}| = 15$ selected sensors
Normalize temporal features using normal-only training statistics
Fit PCA on normal training features; retain C components (95% variance) to obtain $\mathbf{W}$
Project normal training features: $\mathbf{Z}_N = \mathbf{X}_N\mathbf{W}$
for $k = 1$ **to** 8 **do**
    Fit GMM with k components on $\mathbf{Z}_N$; record BIC(k)
end for
Set $K^* = \arg\min_k \text{BIC}(k)$; retain best GMM $p^*(\mathbf{z})$
Compute training scores via (5) and normalize via (6)
Optimize τ^* via (7) on training scores
Inference phase
For each test sample $\mathbf{x}$: extract temporal features, normalize, project to $\mathbf{z}$
Compute normalized anomaly score $\hat{s}(\mathbf{x})$ via (5)–(6)
Predict: $\hat{y}(\mathbf{x}) = \mathbf{1}[\hat{s}(\mathbf{x}) \geq \tau^*]$
return $\hat{y}$

G. GMM-LIME Explainability

RUSBoost-15 is used as the black-box model for explanation. It is chosen because it is designed for class imbalance and represents the strongest supervised baseline on the 15-sensor subset.

Standard LIME: LIME [8] explains a prediction for instance $\mathbf{x}_0$ by finding the interpretable surrogate g^*:
$$g^* = \arg\min_{g \in \mathcal{G}}\ L(f, g, \pi_{\mathbf{x}_0}) + \Omega(g), \tag{8}$$
where f is the black-box model, $\pi_{\mathbf{x}_0}$ is an RBF proximity kernel, and $\Omega(g)$ penalizes complexity. Standard LIME generates perturbed samples by independent Gaussian draws, ignoring inter-sensor correlations.

GMM-LIME: Following [9], we replace Gaussian perturbation with GMM-structured sampling. A GMM is fitted to the normal training data. Component i is drawn with probability π_i, and a perturbed sample is then drawn from $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. This preserves the covariance structure of realistic sensor readings. The proximity kernel:
$$\pi_{\mathbf{x}_0}(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_0\|^2}{2\sigma^2}\right), \tag{9}$$
with σ set to the median pairwise distance among perturbed samples, weights the surrogate fit toward the local neighbourhood of $\mathbf{x}_0$.

Stability Evaluation: Both methods are run five times for each instance. Pairwise Jaccard similarity across runs:
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{10}$$

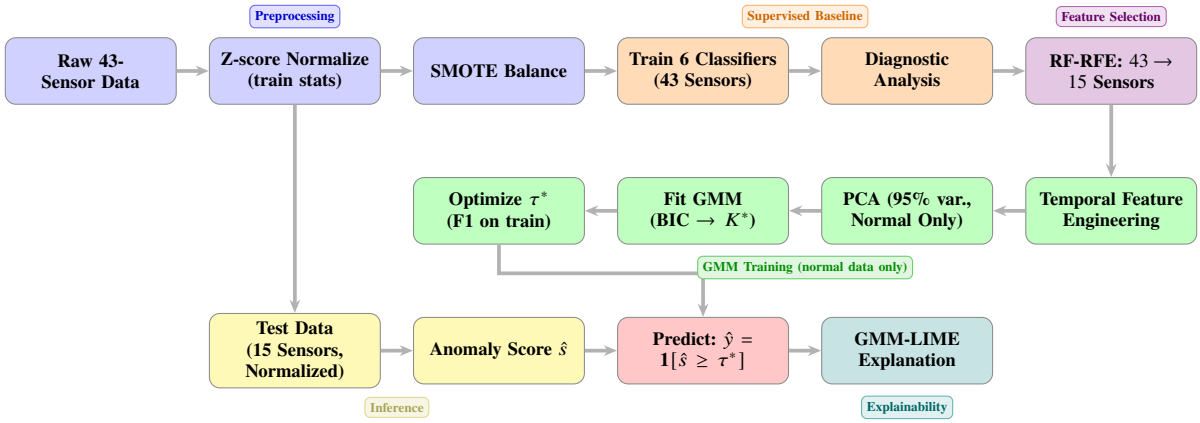

Fig. 1. Proposed GMM anomaly detection pipeline. Row 1 covers preprocessing, supervised benchmarking, and RF-RFE sensor selection. Row 2 shows GMM training on normal data only(right to left). Row 3 shows inference and GMM-LIME explanation.

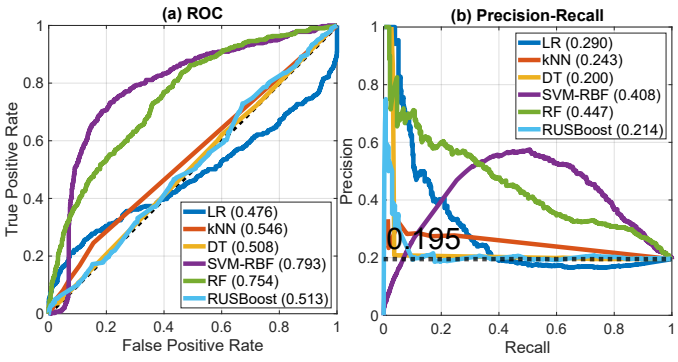

Fig. 2. ROC (a) and Precision-Recall (b) curves for all six supervised classifiers on 43 sensors. AUC-ROC and AUPR values are shown in the legend. The dashed line in (a) and (b) mark the no-skill baseline.

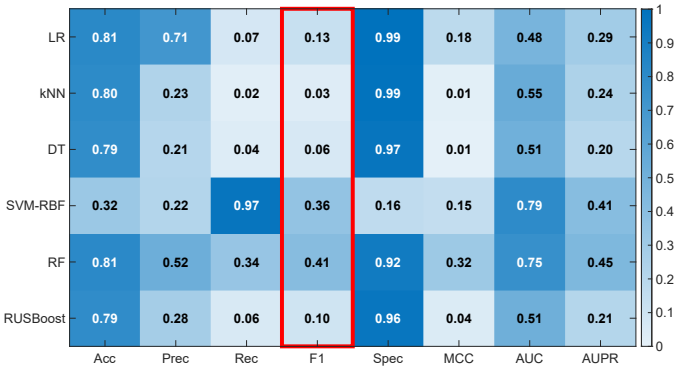

Fig. 3. Performance heatmap for all supervised classifiers across eight metrics. The red rectangle highlights the F1 column.

and overlap coefficient:

$$O(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}, \tag{11}$$

where A and B are the top-K feature sets from two runs. Higher values indicate that the explanation is consistent regardless of the random perturbation seed.

III. Results and Discussion

A. Supervised Baseline Performance

Figures 2 and 3 report test-set results for all six classifiers. Results show that most classifiers achieve high accuracy and specificity by predicting the majority class most of the time. LR reaches 99.3% specificity but detects only 7.1% of faults (F1 = 0.13). SVM-RBF on the other hand has 96.8% recall, which comes at the cost of flagging 83.2% of normal samples as faults, yielding a specificity of 0.16 that is operationally unusable. RF strikes the best balance among supervised methods, with F1 = 0.41 and AUC-ROC = 0.754, but even this falls well short of reliable detection. RUSBoost, despite being specifically designed for imbalanced data, achieves only F1 = 0.10. This outcome already hints at the distribution shift hypothesis: RUSBoost aggressively learns fault patterns from training data, but those patterns simply do not transfer.

B. Diagnostic Analysis

Figure 4 shows the results of the two diagnostic tests. Cohen's d exceeded 0.2 for only 6 of the 43 sensors, and no sensor approached a large effect size. The KS test (not shown here) found a significant distributional shift between training and test faults in 18 of the 43 sensors ($p < 0.05$). Together, these results indicate that only a small fraction of sensors carry discriminative signal in training, and for nearly half of those, the signal has shifted between the training and test conditions. The Mahalanobis distance results make the severity of this shift clear. The mean distance from the normal centroid was $\mu = 5$ for training normal samples and $\mu = 49$ for training fault samples. Test fault samples had a mean distance of $\mu = 1448$, which is approximately 30 times larger than the training fault distance. A supervised classifier calibrated on training fault signatures has no mechanism for detecting samples this far outside its training support.

C. Sensor Selection via RF-RFE

Figure 5 presents the RFE results. Panel (a) shows that AUC-ROC is relatively stable between 43 and approximately

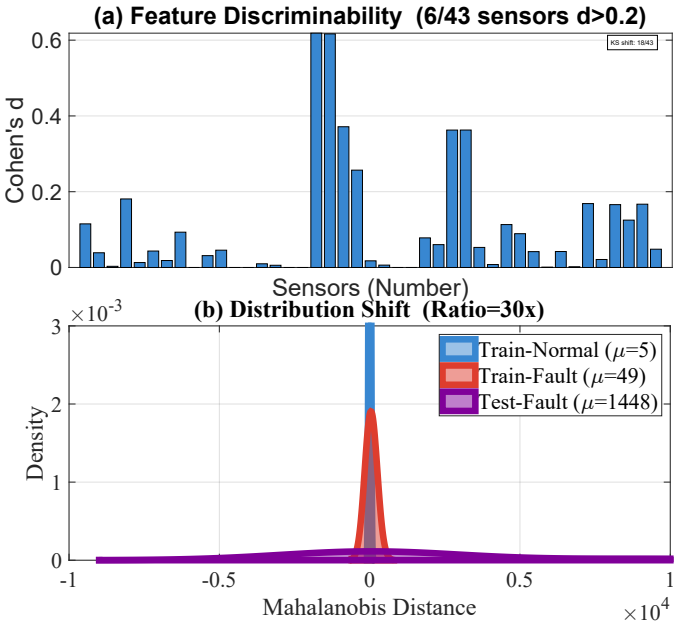

Fig. 4. Diagnostic analysis: (a) Cohen's d for each sensor and (b) Mahalanobis distance distributions.

22 sensors, then begins to fall more steeply as genuinely informative sensors are removed. At the 15-sensor cutoff, RF-15 achieves AUC = 0.556, compared with RF-43's AUC = 0.754. This reduction in supervised AUC might appear to undermine the sensor selection. There are two reasons it does not. First, the purpose of the 15-sensor subset is to provide a compact input space for the GMM, not to maximize supervised performance. Fewer redundant dimensions allow the GMM to estimate the normal distribution more accurately. Second, as the final results confirm, the GMM on 15 sensors achieves AUC = 0.853, which substantially exceeds RF-43's 0.754. The sensors identified by RFE are jointly informative for characterizing normal operating conditions, even if they are not sufficient on their own to train a reliable supervised classifier on the limited training fault data. Panel (b) shows that S20 and S32 carry the highest OOB importance scores among the retained sensors.

D. GMM Detector Performance

Table II shows the full comparative results across all nine methods. Figure 6 shows the GMM diagnostics. The BIC criterion in panel (a) selected $K = 8$ components, reflecting multiple distinct normal operating modes in the WDS data. Panel (b) shows that the score distributions of normal and fault test samples are well separated, with the threshold at 4.47 marking a sensible decision boundary. The confusion matrix in panel (c) reports TP = 169, FP = 58, FN = 238, TN = 1,624.

The proposed GMM outperforms all supervised baselines on every fault-sensitive metric. Compared with RF-43, the best supervised method, the GMM improves F1 by 0.120 (+29.1%), AUC-ROC by 0.099 (+13.1%), and AUPR by 0.135 (+30.2%). The AUPR improvement is particularly significant because it is the more informative metric under class imbalance: a model must achieve both high precision and high recall across many thresholds to attain a high AUPR. The comparative ROC and PR curves in Fig. 7 show the GMM curve consistently above all supervised baselines across the full threshold range.

TABLE II
COMPARATIVE RESULTS FOR ALL METHODS (TEST SET)

Method	Acc	Rec	F1	Spec	AUC	AUPR
LR	0.813	0.071	0.130	0.993	0.476	0.290
kNN	0.798	0.017	0.032	0.987	0.546	0.243
DT	0.785	0.037	0.063	0.966	0.508	0.200
SVM-RBF	0.321	0.968	0.358	0.168	0.793	0.408
RF-43	0.811	0.341	0.413	0.927	0.754	0.447
RUSBoost-43	0.786	0.061	0.101	0.960	0.513	0.214
RF-15 (RFE)	0.792	0.049	0.084	—	0.556	0.230
RUS-15 (RFE)	0.796	0.037	0.066	—	0.537	0.228
GMM (proposed)	**0.858**	**0.415**	**0.533**	**0.965**	**0.853**	**0.582**

TABLE III
EXPLANATION STABILITY OVER 5 RUNS

Instance	Method	Jaccard	Overlap
Fault	LIME	0.544	0.700
	GMM-LIME	0.684	0.810
Normal	LIME	0.498	0.660
	GMM-LIME	0.686	0.810

The structural reason for the GMM's advantage is worth stating explicitly. Because the GMM is trained only on normal data, it does not try to distinguish training fault signatures from normal samples. Instead, it builds a detailed probabilistic model of normal operation and flags whatever deviates from it. Test faults, including those whose signatures differ substantially from training faults, still deviate from the normal model and receive high anomaly scores.

E. GMM-LIME Explainability

Figure 8 shows feature coefficients from GMM-LIME and standard LIME across three runs for a high-anomaly fault instance. The GMM-LIME panels show noticeably more consistent feature rankings run to run. Sensor S20 and S34 appear as primary negative contributors (associated with anomalous conditions) across all three GMM-LIME runs, whereas the LIME panels show more variation in which sensors appear and with what sign. Table III and Fig. 9 quantify this difference. For the fault instance, GMM-LIME achieves a mean Jaccard index of 0.684 versus 0.544 for standard LIME, a relative improvement of 25.7%. The overlap coefficient improves from 0.700 to 0.810. For the normal instance, the Jaccard index improves from 0.498 to 0.686. Both metrics and both instances show that GMM-structured perturbation produces more reproducible explanations.

The instability in standard LIME arises from its independent Gaussian perturbation step. Because sensor readings in a WDS are physically coupled (pressure and flow measurements covary by Bernoulli's principle), independently perturbed samples often represent physically impossible operating states. The

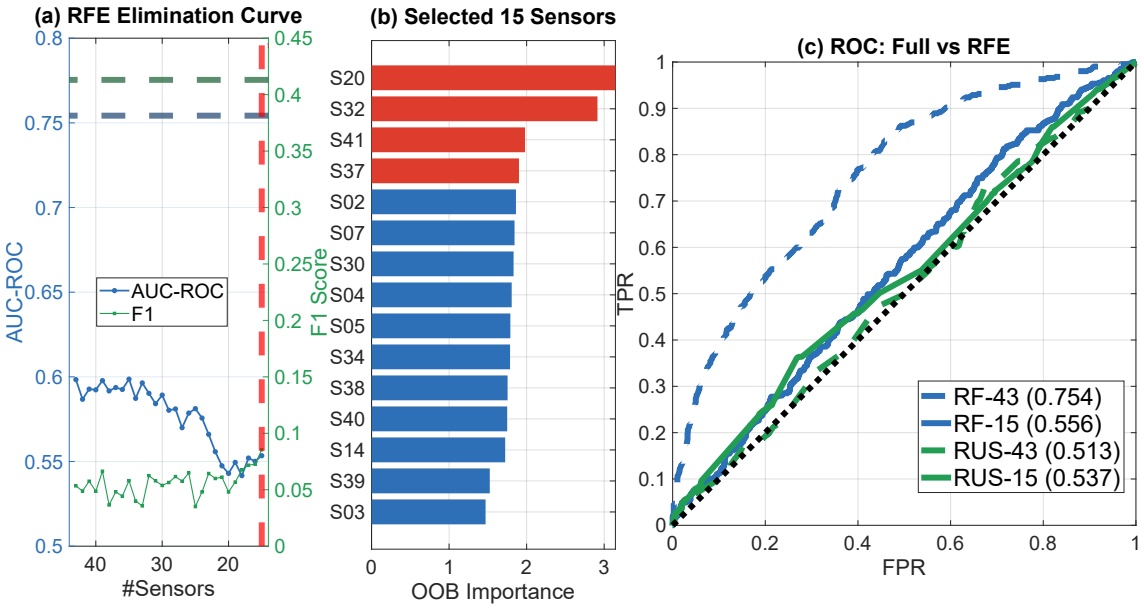

Fig. 5. RF-RFE sensor selection. (a) AUC-ROC and F1 as sensors are eliminated; the red dashed line marks the 15-sensor cutoff. (b) OOB importance of the 15 retained sensors; S20 and S32 are the top contributors. (c) ROC curves comparing full-sensor and RFE-reduced supervised models.

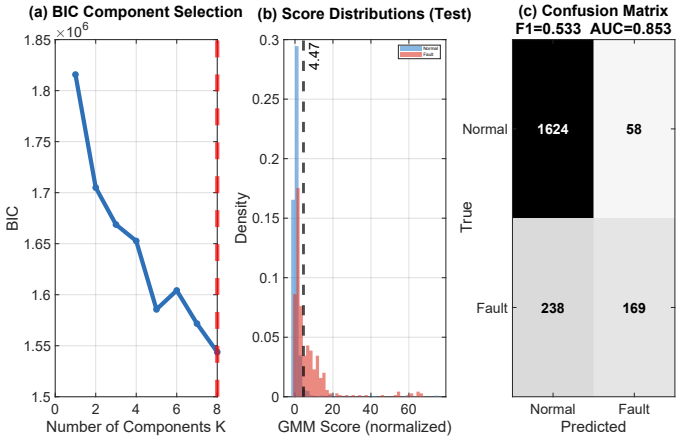

Fig. 6. GMM detector diagnostics. (a) BIC selected $K = 8$ components. (b) Normalised score distributions on the test set; the threshold of 4.47 separates the fault tail from the normal mass. (c) Confusion matrix:

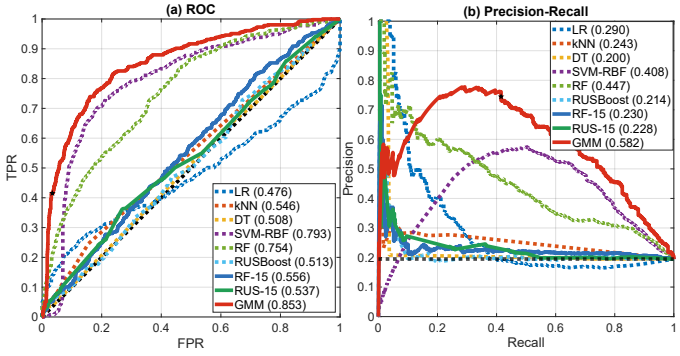

Fig. 7. Comparative ROC (a) and PR (b) curves for all nine methods. The proposed GMM lies above all supervised baselines across the full threshold range. AUC-ROC = 0.853 and AUPR = 0.582.

black-box model responds inconsistently to such samples, and the resulting linear fit changes substantially from run to run. GMM sampling avoids this: perturbed samples are drawn from a distribution fitted to real sensor data, so the black-box model responds more predictably and the surrogate is more faithful across runs.

F. Human-Centred AI Positioning and Operational Implications

The proposed framework is intended as operator-centred decision support for SCADA monitoring rather than as autonomous replacement of human judgement. In human-centred AI (HCAI), the value of a system depends not only on predictive performance, but also on whether its outputs are reliable, understandable, and usable in real operational settings [23], [24]. In dynamic control environments, operators must maintain situation awareness while acting under uncertainty and time pressure [25], [26]; accordingly, an anomaly detector should help prioritise attention, not merely raise opaque alarms.

The proposed normal-only GMM supports this role by modelling deviation from expected system behaviour rather than relying on historical fault signatures that may not transfer to deployment. This aligns with the principle of appropriate reliance in automation, where operators should use system outputs as aids without assuming that past labelled faults exhaust future abnormal conditions [27], [28]. The reduced 15-sensor subset further narrows diagnostic focus to the most informative measurements, which can help structure operator attention during alarm triage [24].

A further HCAI contribution of the framework is expla-

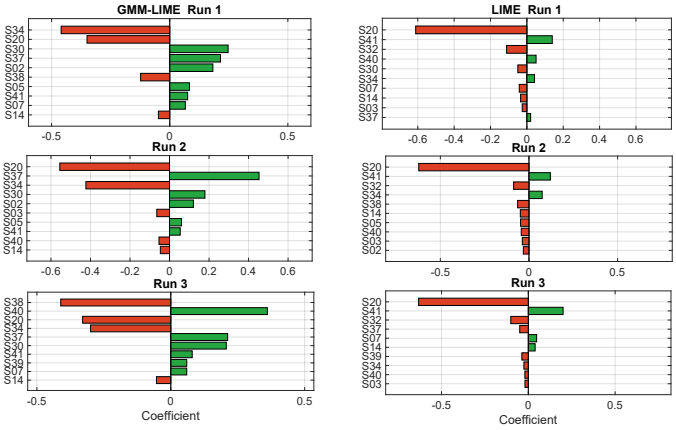

Fig. 8. Feature explanations for a high-anomaly fault instance across three runs. Left: GMM-LIME; right: standard LIME. Green bars indicate positive contributions; red bars indicate negative contributions. Sensors S20 and S34 appear consistently as primary contributors in all GMM-LIME runs, whereas LIME shows greater run-to-run variability.

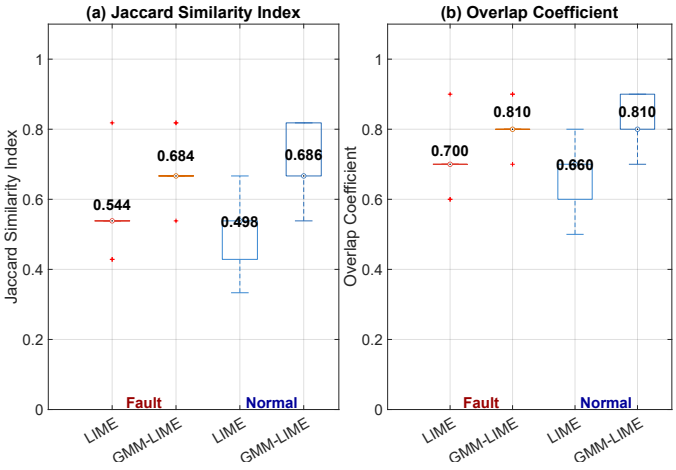

Fig. 9. Explanation stability over five runs measured by Jaccard similarity index (a) and overlap coefficient (b) for fault and normal instances. Annotated values show group means. GMM-LIME achieves higher stability on all four instance-metric combinations.

nation stability. In AI-assisted decision making, trust should be calibrated to system capability rather than increased indiscriminately [29]. Explainable AI research likewise shows that useful explanations should support understanding without imposing unnecessary cognitive burden, and should be evaluated in terms of their effect on human trust, understanding, and performance [30], [31]. By generating covariance-respecting perturbations, GMM-LIME produces more reproducible local explanations, making it more likely that operators will see the same key sensors highlighted across repeated analyses. In critical infrastructure monitoring, such consistency is important for transparent alarm review, cross-checking with process knowledge, and more dependable human-AI collaboration.

IV. Conclusion and Future Works

This paper showed that the core challenge in WDS fault detection in benchmark BATADL dataset is not model complexity but data structure. A 30-fold Mahalanobis distance gap between training and test fault distributions makes supervised generalization unreliable regardless of which classifier is used. The proposed GMM-based detector circumvents this by working entirely from the normal side of the data. It achieves AUC-ROC = 0.853 and AUPR = 0.582 on the 15-sensor RFE subset, compared with 0.754 and 0.447 for Random Forest on all 43 sensors. RF-RFE sensor selection provides a 65% reduction in monitoring requirements without sacrificing, and in fact improving, overall discrimination. The GMM-LIME adaptation produces stable, physically grounded explanations with average Jaccard similarity of 0.684 for fault instances. This stability is operationally important: an explanation that varies substantially from one run to the next cannot be trusted to guide field response decisions. Therefore, from a human centric AI perspective, the main value of the proposed framework is that it combines robust anomaly detection with stable local explanations that can support operator trust, diagnosis, and response in critical infrastructure monitoring.

Future directions include online adaptation of the normal model to account for seasonal demand shifts and extension to spatially aware anomaly attribution using WDS network topology graphs. Additionally, the present study does not include a formal user study, which is an important limitation. Future work should therefore evaluate the framework with domain practitioners such as SCADA operators, process engineers, or cybersecurity analysts. A suitable human-centred evaluation could compare standard alarming against the proposed GMM + GMM-LIME interface using measures such as time-to-diagnosis, fault localization accuracy, perceived explanation usefulness, trust calibration, cognitive workload, and willingness to act on model recommendations. Such a study would help determine whether improved explanation stability translates into better operator performance and safer decision making in real monitoring workflows.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author used ChatGPT to improve grammar, enhance readability, and refine the language of the manuscript. After using ChatGPT, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Acknowledgment

This work is supported by the CHIST-ERA funded TROCI Project (CHIST-ERA-22-SPiDDS-07) through the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/Y036344/1.

References

[1] R. Taormina, S. Galelli, N. O. Tippenhauer, E. Salomons, and A. Ostfeld, "Characterizing cyber-physical attacks on water distribution systems," *Journal of Water Resources Planning and Management*, vol. 143, no. 5, May 2017.

[2] F. Rustam, A. D. Jurcut, and M. Salauddin, "Unsupervised diffusion-guided LSTM for anomaly detection in water distribution networks," *IEEE Access*, vol. 14, pp. 58 919–58 935, 2026.

[3] M. Aghashahi, R. Sundararajan, M. Pourahmadi, and M. K. Banks, "Water distribution systems analysis symposium – battle of the attack detection algorithms (BATADAL)," in *Proceedings of the World Environmental and Water Resources Congress*, 2017, pp. 101–108.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002.

[5] J. Lachure and R. Doriya, "Intelligent sensor data analysis through hybrid deep hierarchical clustering for anomaly detection," *Transactions of the Institute of Measurement and Control*, Dec. 2024.

[6] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, p. 1135–1144.

[9] M. M. Mulwa, R. W. Mwangi, A. Mindila, M. M. Mulwa, R. W. Mwangi, and A. Mindila, "GMM-LIME explainable machine learning model for interpreting sensor-based human gait," *Engineering Reports*, vol. 6, no. 10, Feb. 2024.

[10] M. A. Habib, A. D. Jurcut, H. Ahmed, W. Wei, and M. Salauddin, "Cybersecurity in water distribution networks: A systematic review of AI-based detection algorithms," *Water*, vol. 18, no. 4, p. 519, Feb. 2026.

[11] A. C. Chen and K. Wulff, *Adaptive Neural Trees for Attack Detection in Cyber Physical Systems*, 2022, pp. 89–104.

[12] F. Rustam, M. Salauddin, U. Saeed, and A. D. Jurcut, "Dual-approach machine learning for robust cyber-attack detection in water distribution system," in *Proceedings of the 14th International Conference on the Internet of Things*, 2024, pp. 248–254.

[13] N. Kadosh, A. Frid, and M. Housh, "Detecting cyber-physical attacks in water distribution systems: One-class classifier approach," *Journal of Water Resources Planning and Management*, vol. 146, no. 8, 2020.

[14] K. Qian, J. Jiang, Y. Ding, and S. Yang, "Deep learning based anomaly detection in water distribution systems," in *Proceedings of the IEEE International Conference on Networking, Sensing and Control (ICNSC)*, 2020, pp. 1–6.

[15] J. Goh, S. Adepu, M. Tan, and Z. S. Lee, "Anomaly detection in cyber physical systems using recurrent neural networks," in *Proceedings of the IEEE 18th International Symposium on High Assurance Systems Engineering (HASE)*, 2017, pp. 140–145.

[16] G. Moraitis and C. Makropoulos, "A hybrid dynamic graph neural network framework for real-time anomaly detection," *Journal of Hydroinformatics*, vol. 26, no. 12, p. 3172–3191, Dec. 2024.

[17] A. Zanfei, A. Menapace, B. M. Brentan, M. Righetti, and M. Herrera, "Novel approach for burst detection in water distribution systems based on graph neural networks," *Sustainable Cities and Society*, vol. 86, p. 104090, Nov. 2022.

[18] F. Rustam, A. D. Jurcut, and M. Salauddin, "GSPO-LAD: A graph neural network-based sensor placement and anomaly detection in water distribution systems," *IEEE Access*, vol. 13, pp. 149 462–149 477, 2025.

[19] A. Krause, J. Leskovec, C. Guestrin, J. VanBriesen, and C. Faloutsos, "Efficient sensor placement optimization for securing large water distribution networks," *Journal of Water Resources Planning and Management*, vol. 134, no. 6, pp. 516–526, 2008.

[20] G. Visani, E. Bagli, F. Chesani, A. Poluzzi, and D. Capuzzo, "Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 91–101, 2022.

[21] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, p. 1340–1347, Apr. 2010.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., ser. Information Science and Statistics. New York, NY: Springer, Aug. 2006.

[23] M. Regona, T. Yigitcanlar, C. Hon, and M. Teo, "Building trust in artificial intelligence: A systematic review through the lens of trust theory," *ACM Computing Surveys*, vol. 58, no. 9, pp. 1–39, Feb. 2026.

[24] A. Q. Zhang, J. Suh, M. L. Gray, and H. Shen, "Effective automation to support the human infrastructure in AI red teaming," *Interactions*, vol. 32, no. 4, pp. 58–61, 2025.

[25] A. Ayodeji, M. Mohamed, L. Li, A. Di Buono, I. Pierce, and H. Ahmed, "Cyber security in the nuclear industry: A closer look at digital control systems, networks and human factors," *Progress in Nuclear Energy*, vol. 161, p. 104738, 2023.

[26] B. Aamoth, W. E. Lee, and H. Ahmed, "Net-zero through small modular reactors - cybersecurity considerations," in *IECON 2022 - 48th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, Oct. 2022, pp. 1–5.

[27] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance," *Frontiers in Computer Science*, vol. 5, Feb. 2023.

[28] M. Dubiel, S. Daronnat, and L. A. Leiva, "Conversational agents trust calibration: A user-centred perspective to design," in *Proceedings of the 4th Conference on Conversational User Interfaces*, ser. CUI 2022. ACM, 2022, pp. 1–6.

[29] L. Sanneman, M. Tucker, and J. A. Shah, "An information bottleneck characterization of the understanding-workload tradeoff in human-centered explainable AI," in *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24. ACM, 2024, pp. 2175–2198.

[30] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[31] I. Donoso-Guzmán, J. Ooge, D. Parra, and K. Verbert, *Towards a Comprehensive Human-Centred Evaluation Framework for Explainable AI*. Springer Nature Switzerland, 2023, pp. 183–204.