Definitional Label Leakage in Aquaculture Water Quality Anomaly Detection

Rosida Vivin Nahari[1]*, Teguh Prasetyo[2], Riza Alfita[3]
Department of Information Systems-Faculty of Engineering, Universitas Trunodjoyo Madura,
Jl. Raya Telang, Bangkalan, Indonesia[1]
Department of Mechanical Engineering-Faculty of Engineering, Universitas Trunodjoyo Madura,
Jl. Raya Telang, Bangkalan, Indonesia[2]
Department of Electrical Engineering-Faculty of Engineering, Universitas Trunodjoyo Madura,
Jl. Raya Telang, Bangkalan, Indonesia[3]

***Abstract*—Semi-supervised anomaly detection is widely proposed for Internet of Things water quality monitoring in aquaculture. However, reported high scores often result from definitional label leakage. This occurs when an observation is labelled anomalous based on a physicochemical threshold, and that same variable is used as an input feature. The benchmark then measures threshold recovery rather than true anomaly detection. To quantify this artefact, we introduce two trivial reference detectors: an all-positive predictor and a per-channel range rule. Using a rigorous leakage-aware protocol, we evaluated Isolation Forest, one-class SVM, a Liquid State Machine, a convolutional autoencoder, and an LSTM autoencoder on two corpora: the three ponds of a public aquaponics dataset that survive a sensor-integrity screen fixed before modelling, eight of eleven ponds having been rejected, and a six-pond dataset from East Java in which all positive test windows originate from two ponds. One-class SVM proved the strongest detector on both corpora, achieving a Matthews correlation coefficient up to 0.77, although with two corpora the shared detector ordering is indicative rather than established. Crucially, on the public corpus, both autoencoders and the trivial range rule performed identically poorly, with recall capped near twenty-one percent. This collapse stems from normal-class contamination caused by duration-based labelling, which inflates the percentile threshold for reconstruction-based models. Furthermore, we demonstrate that the F1 score of a zero-information all-positive predictor exceeds several trained detectors, rendering F1 unsafe as a primary metric at these prevalences. Because the labels are derived from husbandry thresholds rather than expert annotation, and because our multi-seed statistics quantify reproducibility on fixed partitions rather than generalisation across farms or seasons, these results bound the interpretation of the benchmark and not field performance. We conclude by proposing a strict reporting protocol for aquaculture anomaly detection studies, with guidance for applying it retroactively to already-published corpora.**

***Keywords*—Anomaly detection; aquaculture; water quality; label leakage; one-class SVM; liquid state machine; autoencoder**

I. Introduction

Aquaculture now supplies more aquatic food for human consumption than capture fisheries, and intensive pond culture in Southeast Asia is a substantial part of that supply [1]. In intensive systems the dominant acute production risk is a water-quality excursion: dissolved oxygen (DO) falling below the tolerance of the stock, pH moving outside the range in which gill function and ionic regulation are stable, or unionised ammonia accumulating to chronically toxic concentrations [2], [3]. Because the interval between the onset of such an excursion and mass mortality can be shorter than a manual sampling cycle, low-cost IoT sensor arrays with continuous logging have become standard, and an extensive literature now applies machine learning to the resulting multivariate time series [4].

Within that literature, semi-supervised (one-class) anomaly detection is the most common formulation. A model is fitted on data believed to represent normal operation, a score is derived — reconstruction error, isolation depth, distance to a support boundary — and observations whose score exceeds a threshold are flagged. Reported performance is often excellent, with F1 or accuracy above 0.9 being unremarkable.

This study begins from a concern about how that ground truth is obtained. Real aquaculture corpora almost never ship expert-verified anomaly annotations: nobody was standing at the pond recording which minutes were pathological. The near-universal workaround is to synthesise labels by applying published husbandry thresholds to the recorded variables. An observation is labelled anomalous if, for example, DO $<$ 2 mg L^{-1} or pH $\notin$ [6.5, 8.5]. The resulting label vector is then used to evaluate a detector whose *input features are those same variables*.

The consequence is structural rather than statistical. If the label is $y = g(x)$ for a known, deterministic, elementwise g, then a decision rule achieving perfect classification exists in closed form and requires no learning. Every score below perfection therefore quantifies not how well anomalies were detected, but how well the analyst's own thresholding function was reconstructed from unlabelled normal data. We term this condition definitional label leakage. It is not a coding defect and it cannot be removed by better cross-validation; it is a property of the labelling decision, and it silently rescales the meaning of every number in the results table.

Definitional leakage is closely related to, but distinct from, the benchmark pathologies already documented for time-series anomaly detection: trivially separable synthetic anomalies [5], the point-adjustment protocol that lets a random scorer appear near-perfect [6], and the well-known insensitivity of accuracy and F1 to class imbalance [7], [8]. Those critiques concern how scores are *aggregated*. Definitional leakage concerns whether

*Corresponding author.

the target exists independently of the input at all.

Our response is not to abandon rule-derived labels — for many corpora no alternative exists — but to make the resulting ceiling visible and quantitative. We do this by benchmarking two intentionally trivial detectors alongside the learned ones. The first is an all-positive predictor, which fixes the value that any prevalence-sensitive metric assigns to zero information. The second is a per-channel range rule that stores only the minimum and maximum of each channel over the normal training windows; because the true labelling function is exactly of this form, the range rule measures how much of the benchmark is solved by recovering channel bounds. If a deep model does not clearly exceed it, the deep model has learned nothing that the benchmark can reward.

The contributions of this study are as follows.

- We formalise definitional label leakage for rule-labelled water-quality corpora and derive its consequence: an attainable MCC of unity, hence a benchmark that measures threshold recovery (Section III-G).
- We propose two trivial reference detectors as mandatory diagnostic instruments and show empirically that their competitiveness is corpus-dependent and therefore cannot be assumed away (Sections III-G, IV-B).
- We specify a leakage-aware evaluation protocol combining per-pond chronological splitting, training-only normalisation, threshold calibration on a held-out normal subset never used for fitting, and a sensor-integrity screen whose criteria were fixed in a diagnostic dry run before any model was trained (Section III).
- We provide a ten-seed benchmark of five detectors spanning three methodological families — tree/kernel one-class methods, reservoir computing, and deep reconstruction — on two corpora, with Wilcoxon signed-rank tests, Holm correction, standardised effect sizes and bootstrap intervals (Section IV), covering the three public-corpus ponds that survive a pre-registered integrity screen and six locally monitored ponds. These statistics quantify seed reproducibility on fixed partitions rather than generalisation across farms or seasons, a scope limit we restate wherever the results are summarised.
- We report and explain two negative findings of practical importance: the collapse of both autoencoders onto the trivial range rule on the public corpus despite good ranking quality, which we trace to contamination of the normal class by the window-labelling rule; and the divergence between MCC and F1 that makes the latter unsafe as a headline metric at these prevalences (Sections IV-B, IV-F, V-A).

The remainder of the study is organised as follows. Section II reviews related work. Section III describes the corpora, labelling, protocol, detectors and statistics. Section IV reports results. Section V discusses mechanisms and threats to validity. Section VI concludes with a reporting protocol.

II. Related Work

A. Machine Learning for Aquaculture Water Quality

Reviews of machine learning in intelligent aquaculture identify water-quality prediction and abnormality detection as two of the most active application areas [4]. The physico-chemical basis is long established: Boyd's synthesis of pond water-quality requirements provides the tolerance ranges used, directly or indirectly, by most subsequent work [2], [3], and the Emerson equilibrium relation converts total ammonia nitrogen to the un-ionised fraction that is actually toxic as a function of pH and temperature [9]. Indonesian national standard SNI 7550:2009 specifies analogous ranges for tilapia grow-out in still-water ponds [10]. Publicly available labelled corpora remain scarce; the IoT aquaponics pond dataset of Udanor et al. [11] is among the few with per-pond continuous logging and is used here as our public corpus.

B. One-Class and Deep Anomaly Detection

Isolation Forest exploits the observation that anomalies are isolated by fewer random axis-parallel splits than normal points [12]. One-class SVM estimates the support of the normal distribution in a kernel-induced feature space [13], a formulation closely related to support vector data description [14]. Reconstruction-based deep methods assume that a bottlenecked network fitted on normal data will reconstruct normal inputs well and anomalies poorly; this was established for shallow autoencoders [15] and extended to sequential data through LSTM encoder–decoders [16] and to density-based hybrids [17]. Adversarially trained reconstruction schemes such as USAD extend the same principle to multivariate sensor streams and were motivated in part by the sensitivity of autoencoders to anomalies present in nominally normal training data [32] — the failure mode we quantify in Section V-A. Comprehensive surveys are available [18], [19]. Notably, these surveys repeatedly observe that deep detectors do not uniformly outperform classical one-class methods on low-dimensional tabular or short-window data — a pattern our results reproduce.

C. Reservoir Computing and Liquid State Machines

Reservoir computing fixes a random recurrent network and trains only a linear readout, avoiding backpropagation through time [20], [21], [22]. The Liquid State Machine (LSM) is the spiking instance of this idea: a sparsely connected population of leaky integrate-and-fire units projects the input into a high-dimensional transient state from which a linear readout is solved in closed form [20]. Because fitting reduces to one ridge regression, LSMs are attractive for edge deployment on the low-power hardware typical of pond-side gateways, which motivates their inclusion here. To our knowledge they have not previously been benchmarked against deep autoencoders on aquaculture water-quality data under a common thresholding protocol.

D. Evaluation Pitfalls in Time-Series Anomaly Detection

Three lines of criticism are directly relevant. First, Wu and Keogh demonstrate that widely used time-series anomaly benchmarks contain triviality, mislabelling and run-to-failure bias sufficient to make published rankings unreliable [5].

Second, Kim et al. show that the point-adjustment protocol inflates F1 so severely that a random scorer can appear state of the art, and recommend evaluation without it [6]; we accordingly report only unadjusted window-level metrics. Third, the Matthews correlation coefficient is preferable to F1 and accuracy under class imbalance because it requires all four confusion-matrix cells to be favourable [7], [8], [23]; precision–recall summaries are likewise preferable to ROC when the positive class is rare [24].

Our contribution is orthogonal and, we argue, logically prior: even a flawless aggregation protocol will report a meaningless quantity if the target variable is a deterministic function of the features. We are not aware of prior work that quantifies this condition for aquaculture corpora by means of a matched trivial baseline.

E. Protocol Hygiene

For temporally dependent data, random k-fold splitting leaks future information and inflates estimates; blocked or chronological evaluation is required [25]. Because ponds differ systematically in stocking density, substrate and shading, we additionally split within each pond rather than across ponds, so that every pond contributes to both partitions and no model can succeed by memorising pond identity.

III. Materials and Methods

Fig. 1 summarises the pipeline. Each stage is described below; all thresholds, exclusions and hyperparameters are stated explicitly so that the reported numbers are reproducible from the released code.

A. Corpora

Two corpora are used (Table I). UDANOR is the public IoT aquaponics pond corpus [11], comprising 11 pond files with timestamped DO, pH, total ammonia nitrogen (TAN) and temperature. KEDIRI comprises six ponds monitored in Kediri Regency, East Java, Indonesia, under differing management regimes (standard, biofloc, high stocking density, shaded earthen, shallow, and a sixth reference pond), with the same four variables logged.

The corpora differ deliberately. UDANOR is larger, noisier and includes ponds with evident probe faults; KEDIRI is smaller, cleaner and management-stratified. A method that behaves consistently on both is more credible than one tuned to either.

B. Physical Validity Screening

Values physically impossible for the sensor were set to missing before any other operation, using ranges that encode instrument physics rather than husbandry acceptability: pH $\in$ [0, 14], DO $\in$ [0, 20] $\mathrm{mg\,L^{-1}}$, TAN $\in$ [0, 10] $\mathrm{mg\,L^{-1}}$, temperature $\in$ [0, 45] °C. On UDANOR this affected 12.46% of pH, 17.94% of DO, 32.44% of TAN and 0.01% of temperature records; on KEDIRI no record fell outside these ranges. The contrast is itself a finding: a third of the ammonia channel of the public corpus is not physically interpretable, which is why TAN is not used as a UDANOR input.

TABLE I. Corpus Characteristics After Preparation

Property	UDANOR	KEDIRI
Pond files	11	6
Ponds retained after integrity screen	3	6
Raw records	≈1,115,000	≈42,900
Records after screening and NaN removal	305,453	42,822
Input channels	DO, pH	DO, TAN, pH
Window length L	32	16
Stride	8	4
Training windows (all)	26,716	7,473
Training windows labelled normal	22,359 (83.7%)	6,466 (86.5%)
Test windows	11,444	3,194
Test window prevalence π	0.2980	0.1334

C. Sensor-Integrity Screening and Pond Exclusion

Physical-range screening removes impossible values but not persistently faulty probes. A probe reading pH 4.8 for weeks is physically possible and statistically stable, yet it is an instrument failure, and a detector trained on it learns the failure as normality. We therefore screened at pond level using four criteria, any one of which rejected the pond:

- More than 15% of pH readings below 6.0 (stuck or uncalibrated pH probe);
- Median DO below 1.0 $\mathrm{mg\,L^{-1}}$ (persistently anoxic reading, incompatible with a stocked pond);
- More than 15% of DO readings above 15 $\mathrm{mg\,L^{-1}}$ (implausible sustained supersaturation);
- Fewer than 5,000 records (series too short for a chronological split at the chosen window length).

To prevent the screen from becoming an outcome-dependent choice, the criteria and the resulting exclusion list were fixed during a diagnostic dry run in which distributional summaries were inspected and no model was trained. On UDANOR eight of eleven ponds were rejected (IoTPond2, IoTPond6–IoTPond12), removing 772,815 records (69.3%) and leaving IoTpond1, IoTPond3 and IoTPond4. On KEDIRI no pond was rejected. Discarding 69% of a public corpus is a strong intervention and we treat it as a threat to external validity (Section V-C), but the alternative — training on probe faults and reporting the result as anomaly detection — is worse.

D. Sample-Level Labelling

No expert annotations exist for either corpus, so labels are derived from published husbandry criteria [2], [3], [10]. Crucially, and unlike common practice, we restrict the labelling rule to variables that are also model inputs, so that the leakage is total, explicit and quantifiable rather than partial and hidden.

For UDANOR, with inputs {DO, pH}, sample t is anomalous if

$$y_t = \mathbf{1}\left[\,\mathrm{DO}_t < 2.0 \;\vee\; \mathrm{pH}_t \notin [6.5,\, 8.5]\,\right] \qquad (1)$$

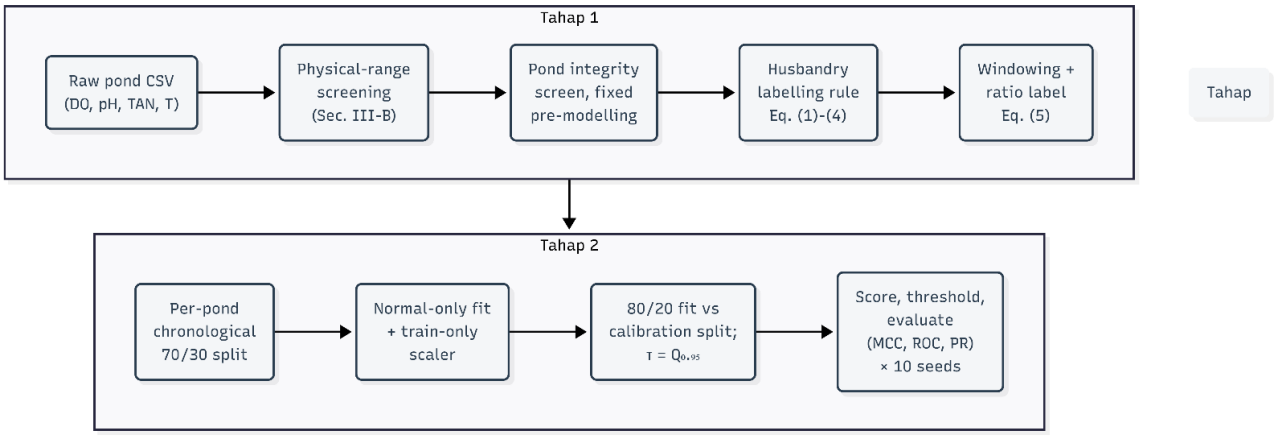

Fig. 1. Leakage-aware evaluation pipeline. Physical-range screening and pond-level integrity screening precede labelling; the chronological 70/30 split is performed within each pond; the scaler and the decision threshold are estimated exclusively from training-side normal windows, with the threshold taken from a calibration subset disjoint from the fitting subset.

For KEDIRI, with inputs {DO, TAN, pH}, the ammonia criterion is applied to the un-ionised fraction rather than to TAN, since it is NH_3 and not NH_4^+ that is toxic. Following Emerson et al. [9],

$$\mathrm{p}K_a = 0.09018 + \frac{2729.92}{T + 273.15} \tag{2}$$

$$[\mathrm{NH_3}]_t = \frac{\mathrm{TAN}_t}{1 + 10^{\,\mathrm{p}K_a - \mathrm{pH}_t}} \tag{3}$$

$$y_t = \mathbf{1}\left[\mathrm{DO}_t < 2.0 \;\vee\; \mathrm{pH}_t \notin [6.5,\, 8.5] \;\vee\; [\mathrm{NH_3}]_t > 0.05\right] \tag{4}$$

with T in °C and concentrations in $\mathrm{mg\,L^{-1}}$. Thresholds follow acute-stress and suitable-range values for warm-water culture [2], [3], [10].

Table II reports the marginal and combined firing rates. Two observations matter for interpretation. On KEDIRI the un-ionised ammonia criterion never fires, so the third input channel contributes nothing to the label; and the pH criterion is almost entirely nested inside the DO criterion (union 3,115 versus 3,108 for DO alone, i.e. pH contributes only 7 additional positive samples), so the KEDIRI label is effectively a single-variable hypoxia criterion. On UDANOR the two criteria are close to independent (36,452 + 9,015 marginal versus 45,077 union, an overlap of 390 samples). The two corpora therefore pose labelling problems of different intrinsic dimensionality, which is relevant when comparing detector rankings across them.

TABLE II. SAMPLE-LEVEL LABEL COMPOSITION

Criterion	UDANOR count (rate)	KEDIRI count (rate)
DO < 2.0 $\mathrm{mg\,L^{-1}}$	36,452 (10.65%)	3,108 (7.24%)
pH $\notin$ [6.5, 8.5]	9,015 (2.63%)	1,539 (3.59%)
Un-ionised $NH_3 > 0.05$ $\mathrm{mg\,L^{-1}}$	not applied	0 (0.00%)
Union (logical OR)	45,077 (13.17%)	3,115 (7.26%)

E. Window Construction and Window Labelling

Detectors operate on windows, so sample labels must be aggregated. The conventional choice — a window is anomalous if it contains at least one anomalous sample — is unusable here. At $L = 32$ it converts a sample prevalence of 10.65% into a window prevalence of 89.71% on UDANOR, leaving only 26% of training windows nominally normal and effectively destroying the negative class. It is also operationally wrong: a single-sample DO dip is sensor noise, whereas sustained hypoxia is the event a farmer must act on.

We therefore use a duration-aware ratio rule. For a window W of length L with sample labels y_t,

$$Y_W = \mathbf{1}\left[\frac{1}{L}\sum_{t \in W} y_t \;\geq\; \rho\right], \qquad \rho = 0.20 \tag{5}$$

A window is anomalous when at least 20% of its samples violate a criterion, i.e. when the excursion persists for at least 6.4 samples at $L = 32$ and 3.2 samples at $L = 16$. This yields the workable prevalences in Table I. Section V-A shows that this choice has an unintended consequence that explains one of our main results: with $\rho = 0.20$, a window may contain up to 19% anomalous samples and still be labelled normal, so the "normal" class is contaminated by construction.

Windows are extracted with a stride (8 for UDANOR, 4 for KEDIRI) to limit redundancy between adjacent windows while preserving temporal coverage.

F. Leakage-Aware Evaluation Protocol

Four provisions define the protocol.

1) Per-pond chronological split: Within each pond, records are sorted by timestamp and the first 70% form the training partition, the last 30% the test partition. Splitting within

rather than across ponds ensures every pond appears in both partitions; splitting chronologically prevents any use of future information [25]. Windows never straddle the cut.

2) Normal-only fitting: Only training windows labelled normal are used to fit any detector, consistent with the semi-supervised formulation.

3) Training-only normalisation: A standard scaler is fitted on the normal training windows alone and applied unchanged to the test partition.

4) Held-out threshold calibration: Every detector produces a continuous score with the convention that larger is more anomalous. Rather than allowing each library to apply its own internal offset, all detectors are thresholded identically. The normal training windows are partitioned by seed into a fitting subset (80%) and a calibration subset (20%); the threshold is the 95th percentile of the score distribution on the calibration subset, which the detector never saw during fitting:

$$\tau = Q_{0.95}\big(\{\, s(W) : W \in \mathcal{C} \,\}\big), \qquad \hat{Y}_W = \mathbf{1}\,[\, s(W) > \tau \,] \tag{6}$$

This provision matters most for the autoencoders. Calibrating on the fitting data would exploit memorisation: a network that reconstructs its training windows almost exactly would receive an artificially low threshold. Two implementation details are worth recording. The comparison is strict ($>$, not $\geq$) because the range-rule baseline scores exactly zero on in-range windows, so a non-strict comparison would flag every window and silently degenerate the baseline into the all-positive predictor; for continuous reconstruction errors the two comparisons are equivalent. And all library detectors are wrapped in a common adapter exposing `fit`, `decision_function`, `set_threshold` and `predict`, so that no detector can quietly retain its own operating point.

G. Detectors

All detectors receive the same 3-D window tensor; those requiring vector input flatten it internally to $L \cdot C$ dimensions (64 for UDANOR, 48 for KEDIRI).

1) Isolation Forest: [12]: 100 trees, default subsampling, on flattened windows.

2) OC-SVM [13]: RBF kernel, gamma = scale, $\nu = 0.05$, on flattened windows. Note that ν shapes the fitted boundary but does not set the operating point, which is governed by Eq. (6) for all detectors alike. This configuration was fixed a priori and no hyperparameter search was conducted for any detector, so every number reported below describes a default-configured model. Because Section V-A attributes OC-SVM's advantage specifically to ν, we there treat that account as a hypothesis and specify the joint sweep over ν and ρ that would test it.

3) LSM reservoir [20], [22]: a population of $N_r = 200$ leaky integrate-and-fire units with sparse random recurrent connectivity. With input $x(t) \in \mathbb{R}^C$, membrane potential v, spike vector s, and synaptic trace r:

$$v(t) = \lambda\, v(t-1) + W_{\text{in}}\, x(t) + W_{\text{res}}\, s(t-1) \tag{7}$$

$$s(t) = \mathbf{1}\,[\, v(t) > \theta \,], \qquad v(t) \leftarrow v(t) \odot \big(1 - s(t)\big) \tag{8}$$

$$r(t) = \alpha\, r(t-1) + s(t) \tag{9}$$

$W_{\text{in}} \sim U(-1, 1)$; W_{res} is Gaussian with connectivity 0.10, rescaled to spectral radius 0.9; leak $\lambda = 0.90$, firing threshold $\theta = 0.5$, synaptic decay $\alpha = 0.80$. Only the linear readout W_{out} is trained, by ridge regression reconstructing the input from the augmented state $[r(t); 1]$ with regulariser 10^{-2}, accumulated in normal-equation form so that memory does not scale with the number of windows. The anomaly score is the mean squared reconstruction error over the window.

4) CNN-AE: Encoder Conv1d($C \to 16$, $k = 3$) – ReLU – MaxPool(2) – Conv1d($16 \to 8$, $k = 3$) – ReLU – MaxPool(2); decoder mirrors it with nearest-neighbour upsampling. 15 epochs, Adam, learning rate 10^{-3}, batch size 64, MSE loss. Score is mean squared reconstruction error.

5) LSTM-AE [16]: single-layer LSTM encoder (hidden size 32), the final hidden state repeated L times, single-layer LSTM decoder and a linear projection to C channels. Same optimisation settings as CNN-AE. Score is mean squared reconstruction error.

6) All-positive baseline (trivial): Flags every window. It carries no information and therefore fixes, for each corpus, the value that a prevalence-sensitive metric assigns to zero information.

7) Range-rule baseline (trivial): stores per-channel minimum l_c and maximum h_c over normal training windows and scores a window by its largest normalised out-of-range excursion,

$$s(W) = \max_{t,c} \frac{\max(l_c - x_{t,c},\, 0) + \max(x_{t,c} - h_c,\, 0)}{h_c - l_c} \tag{10}$$

Thresholded by the same Eq. (6). Its role is diagnostic and follows from the leakage argument. Because Eq. (1) and (4) are per-sample threshold rules on the model inputs, and Eq. (5) is a monotone aggregation of them, a decision rule expressed purely in per-channel bounds attains MCC = 1 by construction — the labelling function itself is such a rule. What the range-rule baseline measures is therefore how much of the benchmark is solved simply by recovering channel bounds from unlabelled normal data. Any learned detector that does not clearly exceed it has not been shown to detect anomalies; it has been shown to be worse at threshold recovery.

H. Metrics

The primary metric is the Matthews correlation coefficient,

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{11}$$

which is high only when all four cells are favourable and is therefore robust to the imbalance present here [7], [23]. We

also report F1, precision, recall, ROC-AUC, average precision (PR-AUC) and wall-clock fitting time. No point adjustment is applied [6].

F1 is reported for comparability with the literature, not as a decision criterion. For an all-positive predictor at prevalence π, precision = π and recall = 1, so

$$\mathrm{F1}_{\text{all-positive}} = \frac{2\pi}{1+\pi} \tag{12}$$

which equals 0.4591 at $\pi = 0.2980$ and 0.2354 at $\pi = 0.1334$ — while MCC is exactly 0. Eq. (12) provides the falsification line against which every reported F1 in Section IV must be read.

I. Statistical Procedure

Each configuration is run over ten seeds (42, 123, 456, 789, 1024, 2048, 4096, 8192, 16384, 32768), controlling model initialisation, the fitting/calibration split and stochastic optimisation. For each pair of trained detectors we compute a two-sided Wilcoxon signed-rank test on paired per-seed MCC [26], adjust the ten resulting p-values by the Holm step-down procedure [27], report the standardised paired effect size $d_z = \mathrm{mean}(d)/s_{\mathrm{pooled}}$, and report a 95% percentile bootstrap interval for the mean difference from 2,000 resamples [28]. This follows the non-parametric protocol recommended by Demšar for comparing classifiers [31], with one deliberate departure: the Friedman and Nemenyi procedures he advocates require many datasets, whereas only two corpora are available here, so we test within each corpus and report the two corpora separately rather than pooling them into a single ranking. Ten seeds is the minimum for which the two-sided Wilcoxon test can reach $p < 0.05$, its attainable minimum being $2/2^{10} \approx 0.002$.

An explicit caveat is required. The test partition is fixed across seeds. Consequently these p-values, intervals and effect sizes characterise the reproducibility of a detector's performance under re-initialisation on this partition, and must not be read as generalisation intervals over ponds, seasons or farms. This is why some d_z values in Table V are numerically enormous: they reflect near-zero seed variance, not large practical superiority. We interpret them accordingly throughout and return to the point in Section V-C.

J. Implementation and Reproducibility

Python 3.12.13, NumPy 2.0.2, scikit-learn [29], PyTorch 2.11.0+cu128 [30], on an NVIDIA Tesla T4. Neural detectors run on the GPU; Isolation Forest, OC-SVM, the LSM (NumPy) and the baselines run on the CPU. Reported times are therefore not comparable across backends and are given only as an order-of-magnitude indication (Section V-C). The full pipeline, including the dry-run diagnostics that fixed the pond-exclusion list, is available at: [repository URL].

IV. Results

A. Overall Performance

Tables III and IV report means and standard deviations over ten seeds, sorted by MCC. Fig. 2 shows the per-seed MCC distributions with both trivial baselines marked.

OC-SVM is the strongest detector on both corpora and the ordering OC-SVM > Isolation Forest ≥ LSM > LSTM-AE ≥ CNN-AE is preserved across them, despite the corpora differing in size by a factor of seven, in channel count, in window length and in the intrinsic dimensionality of the labelling rule. Neither deep detector is best on either corpus, and both are last on both. Two qualifications attach to this statement wherever it recurs. First, the UDANOR rows describe the three ponds retained by the integrity screen of Section III-C and not the eleven-pond corpus as published, while on KEDIRI all 426 positive test windows originate from two of the six ponds (Section V-C); the comparison is therefore between two pond subsets. Second, the listed properties vary jointly rather than one at a time, so a preserved ordering across them is a consistency check on two corpora and identifies no single factor as responsible for it.

B. The Trivial-Baseline Check

The two corpora give opposite verdicts on the trivial range rule, which is precisely why it must be reported rather than assumed away.

On UDANOR the range rule attains MCC = 0.3957. Three of the five trained detectors fall within 0.03 of it — LSTM-AE at 0.3959, CNN-AE at 0.3918, and, at 0.4200 and 0.4225, LSM and Isolation Forest — and CNN-AE is beaten outright. The similarity is not confined to MCC: precision is ≈1.0 and recall is 0.205–0.237 for all four, against 1.0000 and 0.2091 for the rule. These detectors are not approximating the range rule loosely; they are flagging almost exactly the same windows. Only OC-SVM, at +0.335 MCC, does something qualitatively different.

On KEDIRI the range rule collapses to MCC = 0.1060 with recall 0.0223, and every trained detector exceeds it by a wide margin — from +0.289 (CNN-AE) to +0.671 (OC-SVM). Here the learned detectors add genuine value.

The mechanism behind this reversal is developed in Section V-A. Its methodological implication is immediate: the competitiveness of a trivial rule is a property of the corpus, not a constant, and cannot be inferred from published results on a different dataset. A study reporting only trained detectors on UDANOR-like data would report a deep autoencoder at MCC ≈ 0.40 as a positive result while remaining unaware that a two-number rule matches it.

C. Ranking Quality Versus Operating Point

ROC-AUC and MCC diverge sharply on UDANOR. CNN-AE reaches ROC-AUC 0.8286 and PR-AUC 0.6779 against 0.6045 and 0.4448 for the range rule, yet its MCC is lower. Its score therefore ranks anomalous windows above normal ones substantially better than the trivial rule does, while its thresholded decision is worse. Fig. 3 shows this decoupling for all detectors and both corpora.

The practical reading is that the autoencoders' failure on UDANOR is a thresholding failure rather than a representational one. The learned score contains usable information that the 95th-percentile operating point fails to exploit. This distinction is invisible to any study reporting only ROC-AUC (which would rate CNN-AE well above the trivial rule) or only

TABLE III. TEN-SEED RESULTS ON UDANOR (π = 0.2980) MEAN (SD)

Detector	MCC	F1	Prec.	Rec.	ROC-AUC	PR-AUC	Time (s)
OC-SVM	0.7310 (0.0099)	0.7676 (0.0107)	0.9941 (0.0008)	0.6253 (0.0144)	0.9838 (0.0005)	0.9663 (0.0010)	1.48 (0.31)
Isolation Forest	0.4225 (0.0110)	0.3823 (0.0150)	1.0000 (0.0000)	0.2365 (0.0114)	0.8766 (0.0145)	0.7716 (0.0243)	0.22 (0.05)
LSM Temporal	0.4200 (0.0457)	0.3794 (0.0611)	0.9984 (0.0019)	0.2359 (0.0511)	0.8047 (0.0830)	0.7230 (0.0948)	4.27 (1.61)
LSTM-AE	0.3959 (0.0006)	0.3463 (0.0007)	0.9999 (0.0004)	0.2094 (0.0005)	0.7143 (0.0060)	0.5710 (0.0027)	14.66 (0.21)
Range-rule baseline	*0.3957*	*0.3459*	*1.0000*	*0.2091*	*0.6045*	*0.4448*	*0.07 (0.01)*
CNN-AE	0.3918 (0.0121)	0.3407 (0.0163)	1.0000 (0.0000)	0.2054 (0.0116)	0.8286 (0.0360)	0.6779 (0.0463)	10.47 (0.23)
All-positive baseline	*0.0000*	*0.4591*	*0.2980*	*1.0000*	*0.5000*	*0.2980*	*0.00*

TABLE IV. TEN-SEED RESULTS ON KEDIRI (π = 0.1334) MEAN (SD)

Detector	MCC	F1	Prec.	Rec.	ROC-AUC	PR-AUC	Time (s)
OC-SVM	0.7767 (0.0110)	0.8024 (0.0094)	0.7099 (0.0143)	0.9232 (0.0184)	0.9861 (0.0015)	0.9180 (0.0078)	0.12 (0.01)
Isolation Forest	0.6799 (0.0628)	0.7204 (0.0579)	0.7339 (0.0349)	0.7108 (0.0870)	0.9637 (0.0095)	0.7359 (0.0745)	0.19 (0.01)
LSM Temporal	0.5581 (0.0618)	0.6072 (0.0587)	0.6726 (0.0408)	0.5566 (0.0805)	0.8932 (0.0291)	0.6381 (0.0786)	0.54 (0.15)
LSTM-AE	0.4548 (0.0640)	0.5201 (0.0604)	0.5605 (0.0422)	0.4866 (0.0741)	0.8903 (0.0205)	0.4615 (0.0375)	4.26 (0.54)
CNN-AE	0.3949 (0.0962)	0.4611 (0.0940)	0.5209 (0.0628)	0.4178 (0.1126)	0.8849 (0.0293)	0.4380 (0.0475)	3.07 (0.27)
Range-rule baseline	*0.1060 (0.0102)*	*0.0432 (0.0070)*	*0.6771 (0.0110)*	*0.0223 (0.0037)*	*0.5103 (0.0018)*	*0.1485 (0.0021)*	*0.01 (0.00)*
All-positive baseline	*0.0000*	*0.2354*	*0.1334*	*1.0000*	*0.5000*	*0.1334*	*0.00*

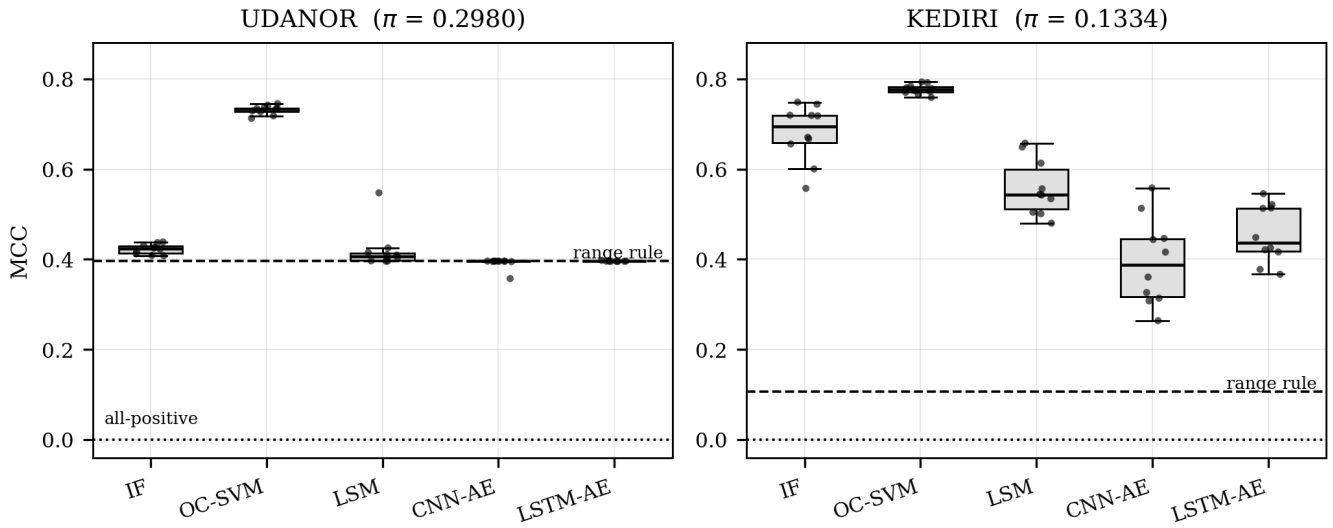

Fig. 2. Per-seed MCC distributions by detector and corpus (n = 10 seeds). Dashed line: range-rule baseline. Dotted line: all-positive baseline (MCC = 0). On UDANOR (left) three trained detectors overlap the trivial range rule; on KEDIRI (right) all trained detectors clear it.

F1 (which would rate it below), and it is the reason we report both families of metric.

D. Statistical Comparisons

Tables V and VI report all ten pairwise comparisons per corpus.

On UDANOR, eight of ten comparisons survive Holm correction, but several surviving differences are practically negligible: Isolation Forest exceeds LSTM-AE by 0.0266 MCC and LSM exceeds CNN-AE by 0.0282, differences roughly an order of magnitude smaller than the 0.335 MCC margin by which OC-SVM clears the trivial range rule. Statistical separability under fixed-test-set seed variation is not evidence of practical relevance, and the extreme d_z values (up to 47.73) are artefacts of near-zero seed variance rather than measures of practical superiority; we do not interpret them as such. The two indistinguishable pairs are Isolation Forest versus LSM and CNN-AE versus LSTM-AE. On KEDIRI, where seed variance is larger and better reflects real instability, nine of ten comparisons separate and the two autoencoders remain indistinguishable from each other.

E. Seed Stability

Seed sensitivity varies by an order of magnitude and follows the model family. The LSTM-AE on UDANOR is almost deterministic (SD 0.0006), which is itself a symptom of

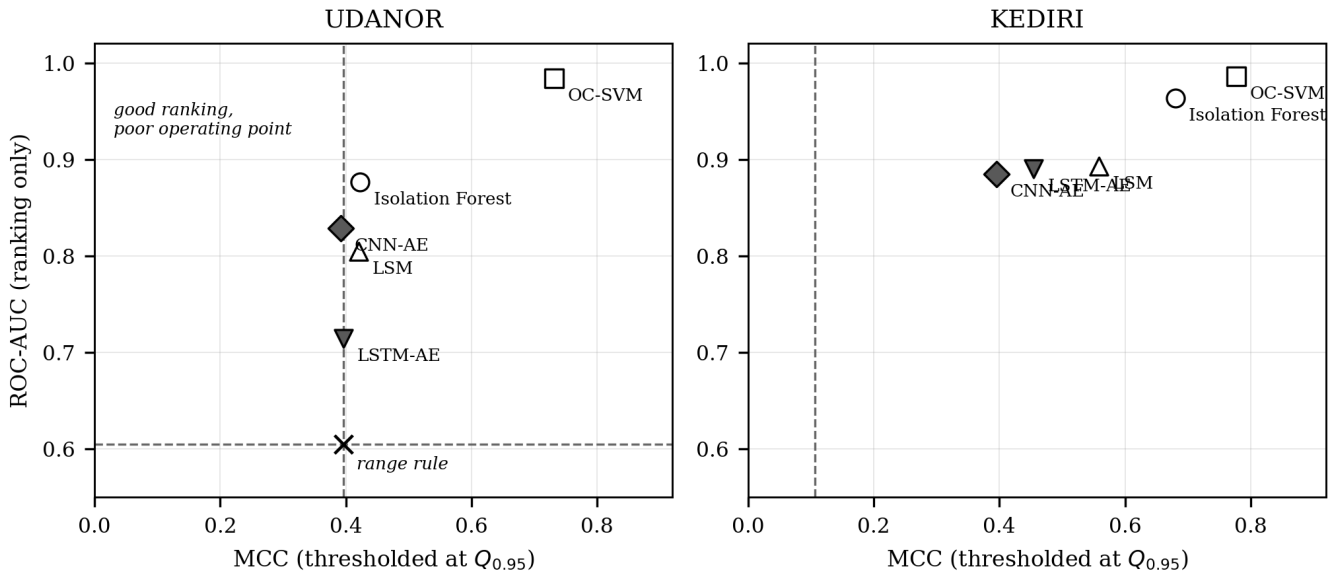

Fig. 3. Ranking quality versus thresholded performance (mean over 10 seeds). Detectors above the range-rule ROC line but left of its MCC line rank anomalies well yet are miscalibrated at the 95th-percentile operating point; CNN-AE on UDANOR is the clearest case.

TABLE V. PAIRWISE MCC COMPARISONS ON UDANOR (WILCOXON, HOLM-CORRECTED; n = 10 SEEDS)

A	B	mean A	mean B	Δ	d_z	95% CI	p (Wilc.)	p (Holm)	Diff.
IF	OC-SVM	0.4225	0.7310	−0.3085	−29.48	[−0.3164, −0.3003]	0.0020	0.0195	yes
IF	LSM	0.4225	0.4200	0.0025	0.07	[−0.0245, 0.0215]	0.1934	0.3750	no
IF	CNN-AE	0.4225	0.3918	0.0307	2.65	[0.0212, 0.0445]	0.0020	0.0195	yes
IF	LSTM-AE	0.4225	0.3959	0.0266	3.41	[0.0203, 0.0332]	0.0020	0.0195	yes
OC-SVM	LSM	0.7310	0.4200	0.3110	9.40	[0.2809, 0.3306]	0.0020	0.0195	yes
OC-SVM	CNN-AE	0.7310	0.3918	0.3392	30.60	[0.3296, 0.3506]	0.0020	0.0195	yes
OC-SVM	LSTM-AE	0.7310	0.3959	0.3351	47.73	[0.3289, 0.3406]	0.0020	0.0195	yes
LSM	CNN-AE	0.4200	0.3918	0.0282	0.84	[0.0072, 0.0588]	0.0078	0.0234	yes
LSM	LSTM-AE	0.4200	0.3959	0.0241	0.75	[0.0060, 0.0541]	0.0039	0.0195	yes
CNN-AE	LSTM-AE	0.3918	0.3959	−0.0041	−0.48	[−0.0119, 0.0000]	0.1875	0.3750	no

TABLE VI. PAIRWISE MCC COMPARISONS ON KEDIRI (WILCOXON, HOLM-CORRECTED; n = 10 SEEDS)

A	B	mean A	mean B	Δ	d_z	95% CI	p (Wilc.)	p (Holm)	Diff.
IF	OC-SVM	0.6799	0.7767	−0.0969	−2.15	[−0.1363, −0.0635]	0.0020	0.0195	yes
IF	LSM	0.6799	0.5581	0.1218	1.95	[0.0741, 0.1687]	0.0020	0.0195	yes
IF	CNN-AE	0.6799	0.3949	0.2850	3.51	[0.2075, 0.3606]	0.0020	0.0195	yes
IF	LSTM-AE	0.6799	0.4548	0.2251	3.55	[0.1795, 0.2715]	0.0020	0.0195	yes
OC-SVM	LSM	0.7767	0.5581	0.2186	4.92	[0.1777, 0.2546]	0.0020	0.0195	yes
OC-SVM	CNN-AE	0.7767	0.3949	0.3819	5.58	[0.3229, 0.4356]	0.0020	0.0195	yes
OC-SVM	LSTM-AE	0.7767	0.4548	0.3220	7.02	[0.2854, 0.3599]	0.0020	0.0195	yes
LSM	CNN-AE	0.5581	0.3949	0.1632	2.02	[0.0823, 0.2437]	0.0059	0.0195	yes
LSM	LSTM-AE	0.5581	0.4548	0.1033	1.64	[0.0552, 0.1508]	0.0098	0.0195	yes
CNN-AE	LSTM-AE	0.3949	0.4548	−0.0599	−0.73	[−0.1278, 0.0137]	0.0840	0.0840	no

degenerate convergence rather than a virtue: every initialisation converges to the same uninformative operating point. The LSM is the least stable detector on UDANOR (SD 0.0457; per-seed range 0.3957–0.5473), because the random reservoir draw is not re-optimised — a single favourable draw at seed 2048 reaches 0.5473, comfortably above Isolation Forest, while eight of ten draws sit at the trivial-baseline level. On KEDIRI, CNN-AE spans 0.2638–0.5581 (SD 0.0962). Single-seed results for reservoir and deep detectors on corpora of this size are therefore uninformative, and reporting the best of several seeds — a common practice — would have changed the ranking here.

F. Metric Selection

Eq. (12) fixes the F1 of a zero-information predictor at 0.4591 on UDANOR and 0.2354 on KEDIRI. Fig. 4 compares these lines against all detectors. On UDANOR the all-positive predictor's F1 of 0.4591 exceeds that of every detector except OC-SVM, including both autoencoders (0.3407, 0.3463), Isolation Forest (0.3823), the LSM (0.3794) and the range rule (0.3459). Its MCC is exactly zero. A study reporting F1 as its headline metric on this corpus would rank a detector that flags every single window above four of its five trained models.

The rank inversion is only avoided because OC-SVM performs well; had it not been included, the study's own results table would have been headed by a predictor carrying no information. On KEDIRI, at lower prevalence, the all-positive line is lower (0.2354) and all trained detectors clear it, so the pathology does not manifest. Its appearance is therefore prevalence-dependent and must be checked per corpus, not argued from precedent.

G. Computational Cost

Cost and accuracy are not in tension here (Fig. 5). On both corpora the two most accurate detectors are also among the cheapest: OC-SVM fits in 1.48 s and 0.12 s, and Isolation Forest in 0.22 s and 0.19 s, against 10.47 s and 14.66 s for the GPU-accelerated autoencoders on UDANOR. The LSM occupies a middle position (4.27 s and 0.54 s) while running as unoptimised NumPy on the CPU, and on KEDIRI it exceeds both GPU-trained autoencoders in accuracy at roughly one-eighth of the LSTM-AE's fitting time — the one result that supports the edge-deployment argument for reservoir computing.

As noted in Section III-J, these times span heterogeneous backends and should be read as orders of magnitude only. Two scaling caveats apply: OC-SVM training is between quadratic and cubic in the number of windows and will not extend to corpora orders of magnitude larger, whereas the autoencoders scale linearly; and the LSM's closed-form ridge readout requires only one pass, so its cost is dominated by the reservoir simulation and is trivially parallelisable.

V. Discussion

A. Why the Autoencoders Collapse on UDANOR

The recall values in Table III are the diagnostic clue: the range rule, both autoencoders, the LSM and Isolation Forest all recover between 20.5% and 23.7% of anomalous windows at near-perfect precision. Such close agreement between a two-number rule and a trained recurrent network indicates a shared constraint rather than five coincidences.

The constraint is contamination of the nominally normal class, introduced by the window-labelling rule Eq. (5). With $\rho = 0.20$, a window containing up to 19% anomalous samples is labelled normal. At $L = 32$ that is up to six samples with DO < 2.0 mg L^{-1} or pH outside [6.5, 8.5]. Since only normal windows are used for fitting and calibration, the "normal" training set demonstrably contains sub-threshold DO and out-of-range pH values.

Two consequences follow, and they explain the two families of detector at once.

For the range rule the consequence is exact. Its per-channel bounds are the extrema of the normal class, and because that class contains excursion samples, its lower DO bound lies well below 2.0 mg L^{-1}. The 95th percentile of its calibration scores is exactly zero, so the effective decision is "flag any window leaving the normal envelope". Its recall of 0.2091 therefore states a property of the data, not of the method: only about 21% of anomalous windows on UDANOR contain a value outside the envelope of the normal class. The other 79% are anomalous by duration — enough samples below 2.0 mg L^{-1} to satisfy ρ — while remaining inside the amplitude envelope of contaminated normal data. Detecting them requires exploiting the sustained character of the excursion, not its magnitude.

For the reconstruction-based detectors the consequence is analogous. Trained with mean squared error on contaminated windows, an autoencoder learns to reconstruct excursion-containing windows as normality. The upper tail of its calibration error distribution is thereby inflated, the 95th percentile rises, and only the most extreme windows clear it — approximately the same windows the envelope rule catches. The observed recall of 0.2054–0.2094 is thus not a failure of representation, as ROC-AUC 0.8286 for CNN-AE confirms: the score ranks well but the calibrated operating point sits far out on an inflated tail. Consistently, the realised false-positive rate on the test partition is near zero rather than the nominal 5%, indicating that the calibrated threshold does not transfer as intended across the chronological cut.

OC-SVM escapes because $\nu = 0.05$ explicitly permits 5% of fitting windows to lie outside the estimated support. Contaminated windows are exactly the ones the optimisation is free to exclude, so the fitted boundary is tighter than the contaminated envelope, and the resulting score separates duration-driven anomalies that lie inside that envelope. In this reading, OC-SVM's advantage on UDANOR is not superior representational power but built-in robustness to label contamination — a hypothesis that a sweep over ν and over ρ, jointly, would test directly.

We state that experiment precisely, because the present study argues the mechanism from converging evidence — the matched recall values, the near-zero realised false-positive rate, and the reversal on KEDIRI — rather than establishing it by ablation. The decisive design is factorial: the window-label contamination ratio $\rho \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$ crossed with the OC-SVM rejection fraction $\nu \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$, with the range rule and both autoencoders recomputed at every ρ so that the trivial reference moves with the label definition. The account given here makes three predictions that design can falsify. First, as ρ falls towards a strictly clean normal class the shared recall ceiling near 0.21 should rise for the range rule and for both autoencoders, because the amplitude envelope of the normal class contracts. Second, OC-SVM's margin over the autoencoders should narrow in the same limit, because the contamination its ν absorbs is progressively removed. Third, at fixed $\rho = 0.20$ the MCC of OC-SVM should be non-monotone in ν, peaking near the fraction of fitting windows that actually contain anomalous samples, and degrading once ν discards clean support vectors. If OC-SVM instead retains its full advantage at $\rho \to 0$, or if its MCC proves flat in ν, the contamination account is wrong and a representational

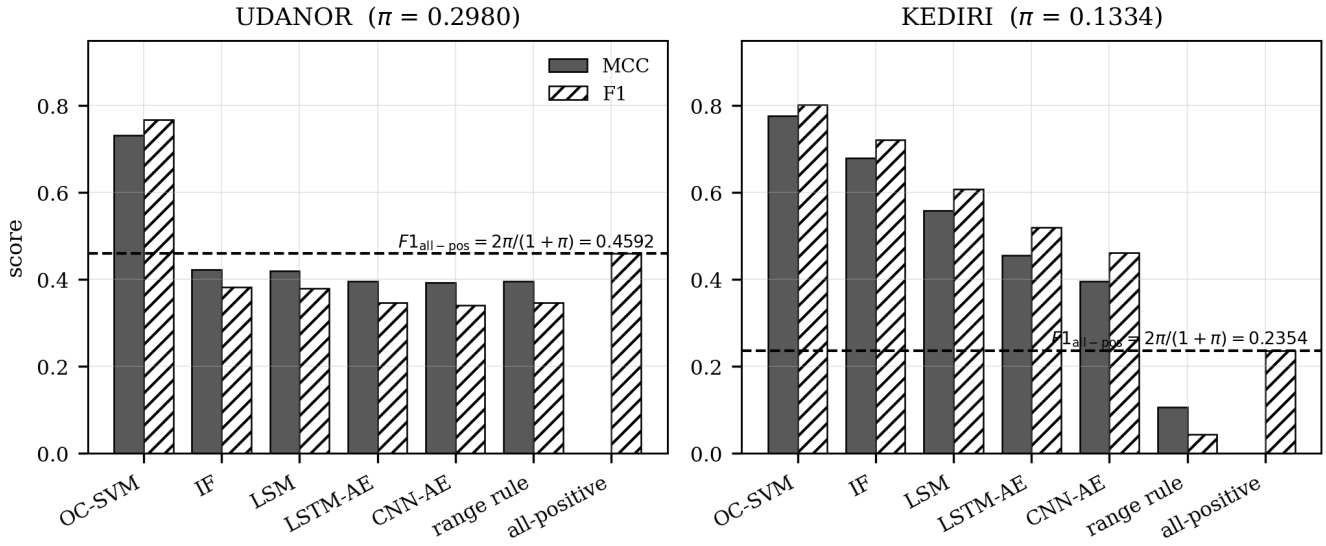

Fig. 4. F1 versus MCC for all detectors, with the zero-information line F1 = $2\pi/(1+\pi)$ (Eq. (12)). On UDANOR the all-positive predictor's F1 exceeds that of four of five trained detectors while its MCC is zero.

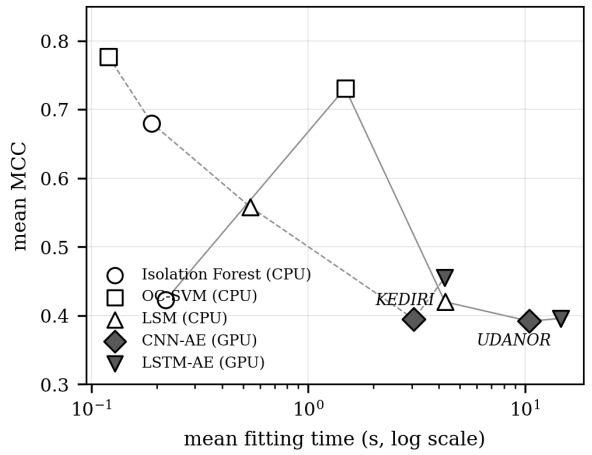

Fig. 5. Mean MCC versus mean fitting time (log scale). Marker shape denotes compute backend; times are not comparable across backends (Section III-J).

explanation must be preferred. We flag this sweep as the most valuable single extension of the present study, and note that its cost is modest: 25 configurations at the fitting times of Table III.

KEDIRI supports the same account from the opposite direction. There the range rule's recall falls to 0.0223: the normal envelope covers 98% of the anomalous region, so amplitude carries almost no information at all. Yet every trained detector performs well, with the LSM at 0.5581 and OC-SVM at 0.7767. On a corpus whose normal manifold is tight relative to sensor noise, temporal shape is learnable even when amplitude is uninformative — which is exactly the regime in which sequence models should help, and where our results show that they do.

B. What the Benchmark Actually Measures

Because the labels satisfy $Y = h(g(x))$ with g a per-sample threshold rule on the model inputs and h the monotone aggregation Eq. (5), the attainable MCC is 1 on both corpora. Every number in Tables III and IV is therefore a measurement of threshold recovery from unlabelled, contaminated normal data, not of anomaly detection in the sense a farm operator would understand. Read that way the results remain informative and, we argue, more honestly so: OC-SVM recovers the analyst's rule best; deep reconstruction recovers it worst; and on one of the two corpora a two-number rule recovers it as well as a trained LSTM.

What the benchmark cannot support is the inference usually drawn from such tables — that the winning detector will identify water-quality events on a new farm. That inference requires labels not derived from the input features. Three routes are available, in increasing order of cost: injecting synthetic anomalies of types the labelling rule does not describe (drift, stuck sensors, phase-shifted diel cycles) and testing whether detectors trained on rule-labelled normality recover them; labelling from an external channel, so that a detector using only DO and pH is evaluated against, say, verified mortality or feeding-response records; or expert annotation of a held-out period. We regard the first as the minimum a follow-up study should provide and the second as the standard the field should adopt.

C. Threats to Validity

We list these explicitly because several are severe and because the study's argument depends on candour about them.

1) Definitional leakage: By design and, we contend, unavoidably given available corpora. It is the object of study, not an accident, but it bounds every claim as set out above. One consequence deserves emphasis. Because the labels are husbandry thresholds rather than expert annotations, anomaly types the rule does not describe — sensor drift, biofouling,

TABLE VII. REPORTING PROTOCOL FOR RULE-LABELLED ANOMALY-DETECTION STUDIES

#	Requirement	Failure it prevents
1	State whether the labelling rule uses variables that are model inputs; if so, declare the attainable ceiling explicitly	Reporting threshold recovery as anomaly detection
2	Report an all-positive baseline and, for threshold-derived labels, a per-channel range rule	Rewarding models that do not exceed a two-number rule
3	Use MCC as the headline metric; report F1 only alongside $2\pi/(1+\pi)$	Rank inversion by a zero-information predictor (Section IV-F)
4	Report both ranking metrics (ROC/PR-AUC) and thresholded metrics	Confounding representational failure with threshold miscalibration
5	Split chronologically and within each entity, never randomly across pooled records	Temporal and pond-identity leakage
6	Calibrate thresholds on held-out normal data never used for fitting	Rewarding memorisation in autoencoders
7	Report the window-labelling rule, ρ, and the resulting contamination of the normal class	Unexplained recall ceilings (Section V-A)
8	Report $\geq$10 seeds with dispersion; never report best-of-n	Reservoir and deep results that do not reproduce (Section IV-E)
9	Fix data-quality exclusions before any modelling and publish the criteria	Outcome-dependent data selection
10	State hardware per detector when reporting time	Backend-confounded efficiency claims

stuck probes, phase-shifted diel cycles, behavioural change preceding disease — are absent from the ground truth and therefore invisible to the evaluation, whether or not a detector flags them. Nothing reported here is evidence that any detector will recognise a real-world excursion the rule does not encode; establishing that requires the externally validated labels set out in Section V-B.

2) Pond exclusion: The integrity screen removed 8 of 11 UDANOR ponds and 69.3% of its records. The criteria were fixed before modelling, but they are heuristic and no independent verification of probe failure exists. Results on UDANOR describe three ponds, not eleven.

3) Concentration of discriminative signal in KEDIRI: Four of six test partitions are entirely normal and contribute no discriminative information. All 426 positive test windows originate from two ponds (the high-stocking-density and shallow ponds). KEDIRI results should therefore be read as a two-pond result with four negative controls, and the cross-corpus consistency of the ranking is correspondingly weaker evidence than the pond count suggests.

4) Label dimensionality on KEDIRI: The un-ionised ammonia criterion never fires and pH violations are almost entirely nested within DO violations (Table II), so the KEDIRI label is effectively a univariate hypoxia criterion while the models receive three channels. One input channel is uninformative about the target by construction.

5) Confounded corpus contrast: The two corpora differ simultaneously in record count, channel count, window length and stride, management regime, acquisition protocol and label dimensionality. The opposite verdicts they return on the trivial range rule (Section IV-B) therefore cannot be attributed to any one of these factors, and the mechanistic account of Section V-A rests on that unresolved contrast. A controlled follow-up should vary one factor at a time; the two cheapest manipulations are to re-window UDANOR at $L = 16$ with stride 4 so that the window geometry matches KEDIRI, and to restrict the UDANOR label to its DO criterion so that both corpora carry an effectively univariate target. Until at least one such variable is isolated, the cross-corpus comparison should be read as descriptive.

6) Corpus provenance: The near-uniform record counts across the six KEDIRI ponds (7,104–7,186) reflect a controlled acquisition protocol. Authors must state precisely how these series were obtained — continuous logging, resampling, or simulation of pond dynamics — since the interpretation of the tight normal manifold in Section V-A depends on it.

7) Fixed test partition: All statistics quantify seed reproducibility, not sampling variability over ponds, seasons or farms. No claim of generalisation to unseen farms is made or supported.

8) No hyperparameter search: Every detector runs at one fixed configuration. This is a comparison of default-configured detectors; a tuned autoencoder might well clear the trivial baseline on UDANOR, and our result should not be read as a claim about the family's ceiling.

9) Heterogeneous timing backends: Neural detectors used a GPU, all others a CPU. Time comparisons are indicative only.

10) Single window length and stride per corpus: L and ρ were fixed a priori; both plausibly affect the contamination mechanism of Section V-A, which we have argued for from converging evidence rather than demonstrated by ablation.

D. Recommendations for Practice

From the above we distil a reporting protocol for rule-labelled aquaculture anomaly-detection studies (Table VII). Each item corresponds to a specific failure this study either exhibited or narrowly avoided.

The protocol is written for new studies, but most of the literature it is meant to correct is already published, and raw per-pond logs are frequently unavailable — to readers, to reviewers, and often to the original authors. We therefore grade

compliance into three tiers by what remains accessible, so that a retrospective audit is possible without re-running any pipeline.

1) Tier 1, paper alone: Items 1, 3, 7 and 10 require no data at all. One sentence stating whether the labelling rule reads variables that are also model inputs discharges item 1; stating the window-labelling rule and ρ discharges item 7; naming the hardware per detector discharges item 10; and item 3 is discharged by publishing the prevalence π of each partition, from which any reader computes the zero-information line F1 = $2\pi/(1+\pi)$ and confirms that its MCC is zero. Publishing π per partition is the single highest-value disclosure in this protocol and costs one number per corpus.

2) Tier 2, released windows and labels but no raw logs: Items 2, 4 and 8 become feasible. Both trivial baselines are computable post hoc without retraining anything, since the all-positive baseline follows from π alone and the range rule needs only the per-channel extrema of the normal training windows; ranking and thresholded metrics can be recomputed from stored scores; and additional seeds require no new data.

3) Tier 3, raw per-pond logs: Only items 5, 6 and 9 require them, because within-entity chronological splitting, held-out threshold calibration and a pre-registered integrity screen all act on the raw series. Work that cannot reach this tier is not thereby invalidated, but its splitting and calibration must be reported as unverifiable.

For auditing published work we regard Tier 1 as the minimum: the labelling rule, ρ, and π per partition. Those three disclosures are enough to determine whether a reported F1 exceeds the zero-information line, which is the failure with the widest reach in the literature surveyed here. Where even these cannot be supplied, the reported metrics should be treated as unaudited rather than as comparable to figures obtained under the full protocol.

VI. Conclusion

We benchmarked five semi-supervised anomaly detectors and two deliberately trivial baselines over ten seeds on two aquaculture water-quality corpora, under a protocol designed to eliminate temporal, pond-identity, normalisation and threshold-calibration leakage. OC-SVM was the strongest detector on both corpora (MCC 0.7310 and 0.7767) and also among the cheapest, and the ranking OC-SVM $>$ Isolation Forest $\geq$ LSM $>$ LSTM-AE $\geq$ CNN-AE was stable across corpora that differ substantially in size, channel count and noise. Neither deep detector was best on either corpus. That stability is a consistency check on two pond subsets — three ponds retained by the integrity screen on UDANOR, and the two KEDIRI ponds that supply every positive test window — obtained under statistics that measure seed reproducibility on fixed partitions; it is not evidence of generalisation to unseen farms or seasons.

The study's more consequential findings are negative. On the public corpus, an LSTM autoencoder (0.3959) and a per-channel range rule storing two numbers per channel (0.3957) are indistinguishable, and a convolutional autoencoder is beaten by that rule — while the same autoencoder attains ROC-AUC 0.8286 against the rule's 0.6045, showing that the deficit lies in the calibrated operating point rather than in the learned representation. We traced the shared recall ceiling near 0.21 to contamination of the nominally normal class by the ratio-based window label, which simultaneously widens the amplitude envelope and inflates the percentile threshold, and we argued that OC-SVM escapes it because its ν parameter permits contaminated points to fall outside the estimated support. Independently, we showed that a zero-information all-positive predictor attains F1 = 0.4591 at the public corpus's prevalence, exceeding four of five trained detectors while attaining MCC exactly zero.

Underlying all of this is a structural point. When anomaly labels are generated by thresholding the variables the detector receives as input, the label is a deterministic function of the input, perfect classification is attainable in closed form, and the benchmark measures threshold recovery rather than anomaly detection. This condition is widespread in the aquaculture literature and is rarely declared. Making it visible costs almost nothing — two trivial baselines and one sentence about the attainable ceiling — and materially changes how a results table should be read.

Future work should, in order of priority: sweep ρ jointly with ν under the factorial design of Section V-A, whose three predictions can falsify the contamination account; train and calibrate on windows with zero anomalous samples to remove contamination; isolate one at a time the corpus properties that co-vary between UDANOR and KEDIRI — window geometry, channel count, label dimensionality — so that the divergent competitiveness of the trivial rule can be attributed to a cause; evaluate against labels that are not functions of the input features, beginning with injected anomaly types the husbandry rule does not describe and progressing to externally validated events; and extend the corpus to more ponds and full production cycles so that generalisation, rather than seed reproducibility, becomes measurable.

Declaration on Generative AI

During the preparation of this work the authors used a generative AI assistant for language refinement and for drafting assistance in structuring the manuscript. The authors designed the study, implemented the pipeline, generated and verified all numerical results, formulated the arguments and conclusions, and reviewed and edited all AI-assisted text. The authors take full responsibility for the content of this publication.

Acknowledgment

This research was funded by the Directorate of Research and Community Service, Directorate General of Research and Development, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, grant number SP DIPA-139.04.1.693320/2026.

Data and Code Availability

The data and code that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Food and Agriculture Organization of the United Nations, *The State of World Fisheries and Aquaculture 2024*. Rome: FAO, 2024.

[2] C. E. Boyd, *Water Quality for Pond Aquaculture*, Research and Development Series No. 43. Auburn, AL: International Center for Aquaculture and Aquatic Environments, Auburn University, 1998.

[3] C. E. Boyd, *Water Quality: An Introduction*, 3rd ed. Cham: Springer, 2020.

[4] S. Zhao, S. Zhang, J. Liu, H. Wang, J. Zhu, D. Li, and R. Zhao, "Application of machine learning in intelligent fish aquaculture: A review," *Aquaculture*, vol. 540, 736724, 2021.

[5] R. Wu and E. J. Keogh, "Current time series anomaly detection benchmarks are potentially misleading and why," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2421–2429, 2023.

[6] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon, "Towards a rigorous evaluation of time-series anomaly detection," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, 2022, pp. 7194–7201.

[7] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 6, 2020.

[8] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, 13, 2021.

[9] K. Emerson, R. C. Russo, R. E. Lund, and R. V. Thurston, "Aqueous ammonia equilibrium calculations: Effect of pH and temperature," *J. Fish. Res. Board Can.*, vol. 32, no. 12, pp. 2379–2383, 1975.

[10] Badan Standardisasi Nasional, *SNI 7550:2009 — Produksi ikan nila (Oreochromis niloticus Bleeker) kelas pembesaran di kolam air tenang*. Jakarta: BSN, 2009.

[11] C. N. Udanor, N. I. Ossai, E. O. Nweke, et al., "An internet of things labelled dataset for aquaponics fish pond water quality monitoring system," *Data in Brief*, vol. 43, 108400, 2022.

[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*, 2008, pp. 413–422.

[13] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[14] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[15] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. MLSDA Workshop on Machine Learning for Sensory Data Analysis*, 2014, pp. 4–11.

[16] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based encoder-decoder for multi-sensor anomaly detection," in *Proc. ICML Anomaly Detection Workshop*, 2016.

[17] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[18] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, 38, 2021.

[19] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint* arXiv:1901.03407, 2019.

[20] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural Computation*, vol. 14, no. 11, pp. 2531–2560, 2002.

[21] H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks," GMD Report 148, German National Research Center for Information Technology, 2001.

[22] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.

[23] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[24] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, e0118432, 2015.

[25] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012.

[26] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[28] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] A. Paszke, S. Gross, F. Massa, et al., "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[31] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[32] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised Anomaly Detection on multivariate time series," in *Proc. 26th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2020, pp. 3395–3404.