Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

AI-Driven Predictive Analytics for Early Detection of Anomalies and Emerging Patterns in Large-Scale Data

Syed Muhammad Raza Zaidi
Department of Computer Systems Engineering, NED University of Engineering and Technology, Karachi, Sindh, Pakistan Email: mraza28067@gmail.com

Sara Irfan
Department of Data Science, National University of Computer and Emerging Sciences (FAST-NUCES), Islamabad, Pakistan. sarairfan581@gmail.com

Ahsan Ikram
Department of Data Sciences, University of Engineering and Technology, Lahore, Pakistan. ahsanikram789@gmail.com

ABSTRACT

Organisations now collect records and sensor readings faster than analysts can inspect them, and the events that matter most, such as fraud, intrusions, equipment faults and the first signs of a change in behaviour, are rare and usually unlabelled. We studied how well unsupervised machine learning detects these events, and at what cost, in two settings. For static records we compared seven detectors from five methodological families (Isolation Forest, ECOD, k-nearest neighbours, LOF, one-class SVM, PCA and an autoencoder) and two rank-averaged ensembles on ten public benchmark datasets ranging from 1,831 to 567,498 records, with five repetitions each. For ordered data we built a two-stage detector that combines the residuals of an autoregressive forecast with a cumulative-sum (CUSUM) statistic, and compared it with four streaming baselines on eight labelled series from the Numenta Anomaly Benchmark. Isolation Forest was the most reliable single detector, with a mean ROC-AUC of 0.880, a mean average precision of 0.541 and the best mean rank (1.9); the differences between methods were significant (Friedman $\chi^2 = 39.2$, $p < 0.001$). A rank-averaged ensemble of four diverse detectors ranked second (3.6) but did not beat Isolation Forest (Wilcoxon $p = 0.014$), and removing its weakest member raised its mean AUC from 0.834 to 0.855. Proximity-based methods failed on clustered anomalies and their run time grew roughly quadratically, reaching 31 s for 80,000 records against 0.5 s for Isolation Forest. In the streaming experiments the residual–CUSUM detector found all 21 labelled anomaly windows, earlier than any baseline (mean delay 28% of the window length), at a cost of 1.95 false-alarm events per 1,000 points. The results suggest that a fast tree-based detector for records, paired with forecast residuals and change accumulation for streams, is a sound default for large-scale monitoring, and that ensembles should be built with care because they inherit the failures of their members.

Keywords: Anomaly Detection; Predictive Analytics; Isolation Forest; Outlier

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Ensembles; Streaming Data; Change Detection; CUSUM; Big Data; Benchmark Evaluation

Introduction

Most of the data that modern organisations collect is never looked at by a person. Card transactions, network connections, readings from industrial sensors and server metrics arrive in volumes that rule out manual review, and the few records that matter are hidden among millions that do not. Anomaly detection, the task of finding observations that do not conform to expected behaviour, is the usual answer to this problem [1, 2]. It underpins fraud screening [3, 4], intrusion detection [5], predictive maintenance and the monitoring of cloud infrastructure [6, 7], and it is often the first analytical step applied to a new data source [8].

Two features make the task hard at scale. The first is the lack of labels. Anomalies are rare, varied and often new, so supervised classifiers are rarely an option and most deployed systems rely on unsupervised detectors that score how unusual each observation is [1, 9]. The second is time. In many applications the question is not only whether a record is odd but whether the process that generates the data has started to behave differently. A slow rise in machine temperature, a shift in demand or a gradual increase in latency are *emerging patterns*: no single point may look extreme, yet the accumulated evidence signals a change that an analyst would want to know about early [10–12].

A large number of algorithms address one or other of these problems, from isolation-based and density-based detectors [13, 14] to deep reconstruction models [9, 15] and change-point methods [16, 17]. Choosing among them is not straightforward. Benchmark studies agree that no detector wins everywhere [18–20], and time-series benchmarks have been criticised for trivial anomalies, mislabelled windows and evaluation protocols that flatter weak methods [21, 22]. Practitioners therefore face a real decision problem: which detector, or which combination of detectors, gives reliable early warning on large data at a cost they can afford?

Problem statement

The problem addressed in this paper is the lack of an integrated, reproducible assessment that answers this decision problem for both static records and ordered data. Accuracy, robustness to different anomaly types, computational cost and timeliness are usually reported in separate studies, under different protocols, so a practitioner cannot tell which trade-offs are real. Rank-based ensembles are widely recommended as a safe choice when labels are absent [23, 24], but how much they gain over the best single detector, and how they behave when one member fails, has rarely been measured directly.

Research gap

Our reading of the literature, summarised in Section 2 and in Table 1, points to four gaps. First, large benchmark studies of tabular detectors [18, 20] report accuracy in detail but say little about run time as data grow, although cost decides what can be deployed on millions of records. Second, outlier ensembles are motivated mainly by

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

theory and by results on small collections [23–25]; there is little evidence on how an ensemble degrades when one of its members fails catastrophically on a particular anomaly type. Third, studies of streaming anomalies tend to measure point-wise or point-adjusted scores [22, 26], which reward detectors that fire somewhere inside a long anomaly window but ignore how early they fire and how many false alarms they raise elsewhere. Fourth, static outlier detection and the detection of gradual change are usually studied apart, even though monitoring systems need both [7, 10].

Objectives and research questions

The aim of this study is to build and evaluate a two-layer predictive analytics framework for the early detection of anomalies and emerging patterns, using only public data so that every result can be reproduced. We ask four questions. RQ1: which unsupervised detectors give the most accurate and most consistent ranking of anomalies across datasets of different size, dimensionality and contamination? RQ2: does a rank-averaged ensemble of diverse detectors outperform the best single detector, and how sensitive is it to its members? RQ3: how do run time and accuracy trade off as the number of records grows? RQ4: can a detector that combines forecast residuals with a cumulative-sum statistic recognise emerging patterns in streams earlier than standard baselines, and at what false-alarm cost?

In answering these questions the paper makes the following contributions. We report a controlled comparison of seven detectors and two ensembles on ten benchmark datasets, with five repetitions, non-parametric significance tests [27] and an explicit scalability experiment. We analyse an ensemble by leave-one-out ablation and show how a single failing member lowers its accuracy. We propose a simple residual–CUSUM fusion for streaming data and evaluate it with event-level measures of recall, false alarms and detection delay rather than point-adjusted scores [22]. Finally, we turn the findings into practical guidance for choosing detectors in large-scale monitoring.

The rest of the paper is organised as follows. Section 2 reviews related work and states the research gap in detail. Section 3 describes the framework, data and evaluation protocol. Section 4 presents the results, Section 5 discusses them, and Section 6 concludes.

Literature Review

Definitions and taxonomies

Hawkins' classical notion of an outlier, an observation that deviates so much from the others that it seems to have been generated by a different mechanism [28], underlies most modern definitions. Chandola *et al.* [1] distinguished point, contextual and collective anomalies, a taxonomy that is still useful: a fraudulent transaction is usually a point anomaly, an unusually high temperature at night is contextual, and a run of slightly elevated readings is collective. Hodge and Austin [2] and Aggarwal [8] grouped detection methods by their underlying assumption, whether statistical, distance-based, density-based, clustering-based or model-based. More recent reviews add deep learning approaches and show that many of them can be understood as variants of the same shallow principles [9, 15].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Shallow detectors for static records

Distance-based detectors score a point by its distance to its *k*-th nearest neighbour or by the mean distance to its *k* neighbours [29]. The local outlier factor (LOF) compares the local density of a point with that of its neighbours, so it can find anomalies in regions of different density [14]. Both methods are intuitive and often accurate, but they need all pairwise neighbourhoods and therefore scale poorly. Isolation Forest takes a different route: it grows random trees that split the data on random features and thresholds, and scores each point by how quickly it becomes isolated, which needs no distance computation and runs in near-linear time [13]. Extensions such as the extended isolation forest remove the axis-parallel bias of the original splits [30]. Boundary methods such as the one-class support vector machine learn a region that contains most of the data [31], while projection methods based on principal component analysis flag points that are poorly reconstructed from the main components [32]. More recently, ECOD estimated the tail probability of each feature from its empirical distribution and combined them into a parameter-free score [33], following the copula-based COPOD [34].

Deep and reconstruction-based detectors

Autoencoders learn to compress and reconstruct normal data and use the reconstruction error as an anomaly score [35]. Recurrent and encoder–decoder variants extend this idea to sequences [36], and forecasting models such as LSTMs have been used to detect anomalies in spacecraft telemetry from the size of their prediction errors [6]. Reviews of deep anomaly detection report that these models are powerful on images, text and high-dimensional signals, but that on tabular data they often fail to beat well-tuned shallow detectors, while costing far more to train [9, 15, 20].

Outlier ensembles

Because no single detector is best on all data, combining several has long been recommended. Lazarevic and Kumar built ensembles by running LOF on random feature subsets [25]. Zimek *et al.* argued that unsupervised ensembles need both accurate and diverse members and that scores must be made comparable before they are combined, typically by converting them to ranks or probabilities [23]. Aggarwal and Sathe analysed outlier ensembles through a bias–variance lens and showed that averaging reduces variance but can also dilute the signal of the strongest member [24]. Libraries such as PyOD made such combinations easy to build [37], but systematic evidence on when an ensemble beats its best member, and on how it behaves when a member fails, remains limited.

Streaming anomalies and emerging patterns

In ordered data an observation is anomalous relative to its recent context, and the change in the generating process is often more important than any single value. Classical statistical process control addresses this with the cumulative-sum (CUSUM) chart, which accumulates small deviations until they cross a threshold and is optimal for detecting persistent shifts in the mean [12, 16]. The machine learning literature treats

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

such changes as concept drift and has developed adaptive-window detectors such as ADWIN [17] and many drift-adaptation strategies [10, 11]. Forecasting-based detectors predict the next value with an autoregressive or seasonal model and flag large residuals [38, 39]. The Numenta Anomaly Benchmark (NAB) provided labelled real-world streams and a scoring scheme that rewards early detection [40], and Ahmad *et al.* used it to evaluate a hierarchical temporal memory detector [7].

Benchmarks and evaluation pitfalls

Campos *et al.* showed that the conclusions of outlier detection studies depend strongly on the choice of datasets, parameters and measures [18], and Goldstein and Uchida [19], Domingues *et al.* [41] and Emmott *et al.* [42] reported similar sensitivity. ADBench, the largest tabular benchmark so far, compared 30 algorithms on 57 datasets and found that no unsupervised method is statistically better than the others on all data [20]. For time series, Wu and Keogh argued that many popular benchmarks contain trivial or mislabelled anomalies [21], Kim *et al.* showed that the widespread point-adjustment protocol lets even random scores look competitive [22], and Schmidl *et al.* [26] and Lai *et al.* [43] called for clearer definitions and broader evaluations. For imbalanced problems, precision–recall measures are more informative than ROC curves alone [44, 45], and comparisons across many datasets should rely on rank-based tests such as the Friedman and Nemenyi procedures [27].

Synthesis and research gap

Table 1 summarises representative studies and what each leaves open. Taken together, the literature offers many detectors and several careful benchmarks, but it rarely combines accuracy, cost and robustness in one controlled design, rarely tests ensembles against their own best member, and rarely evaluates streaming detectors by how early they warn and how often they raise false alarms. The framework in Section 3 was designed to address these points within a single reproducible protocol.

Table 1. Representative studies on anomaly detection and the questions they leave open.

Study	Scope	Main finding	Open question for this work
Liu *et al.*, 2008 [13]	Isolation Forest	Near-linear time, strong accuracy on benchmark data	Behaviour against newer detectors and inside ensembles
Breunig *et al.*, 2000 [14]	Local outlier factor	Finds anomalies in regions of varying density	Cost and failure modes on large, clustered data
Zimek *et al.*, 2014 [23]; Aggarwal and Sathe, 2015 [24]	Outlier ensembles	Diversity and score normalisation are essential	Gain over the best member; effect of a failing member
Campos *et al.*, 2016 [18]	12 detectors, 23 datasets	Conclusions depend on data and measures	Run time as data grow

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Study	Scope	Main finding	Open question for this work
Han *et al.*, 2022 [20]	30 detectors, 57 datasets (ADBench)	No unsupervised method dominates	Cost–accuracy trade-off; ensemble sensitivity
Li *et al.*, 2023 [33]	ECOD	Parameter-free, interpretable, fast	Robustness across anomaly types
Lavin and Ahmad, 2015 [40]; Ahmad *et al.*, 2017 [7]	Streaming (NAB)	Early detection matters; labelled real streams	Simple statistical baselines vs learned models
Kim *et al.*, 2022 [22]; Wu and Keogh, 2023 [21]	Time-series evaluation	Point adjustment and flawed labels inflate results	Event-level recall, false alarms and delay
This study	7 detectors + 2 ensembles, 10 datasets; 5 streaming detectors, 8 series	See Sections 4 and 5	Integrated accuracy, cost, robustness and timeliness

Methodology

Framework overview

The framework has two layers that share a pre-processing stage and an evaluation stage (Figure 1). The first layer scores static records, each independently of time, and is evaluated on ten tabular benchmark datasets. The second layer scores ordered observations and is designed to pick up both isolated spikes and slower, persistent changes that we refer to as emerging patterns [10, 12]. Every detector is unsupervised: labels are used only to evaluate scores, never to fit or tune a model, which matches the setting in which these systems are usually deployed [1, 18].

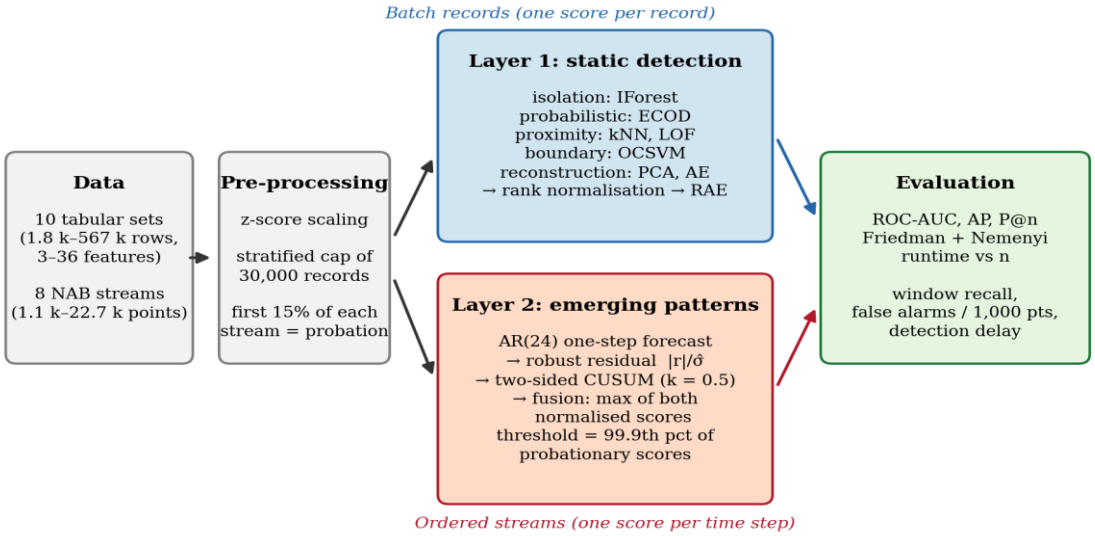


Figure 1. The two-layer predictive analytics framework. Layer 1 scores static records

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

with seven detectors from five families and combines four of them by rank averaging (RAE). Layer 2 scores ordered streams by fusing autoregressive forecast residuals with a two-sided CUSUM statistic [16]. Both layers share pre-processing and are evaluated with rank-based statistics and event-level measures [22, 27].

Datasets

For the first layer we used ten datasets from the ODDS collection [46] in the versions distributed with ADBench [20], chosen to span three orders of magnitude in size, 3 to 36 features and anomaly rates from 0.17% to 31.6% (Table 2, Figure 2). They come from medicine (Cardio, Cardiotocography, Thyroid, Mammography), remote sensing (Satimage-2, Satellite), a synthetic signal (Waveform), an industrial control process (Shuttle), network intrusion (HTTP, from KDD Cup 1999 [5]) and payment fraud (European credit-card transactions [4]). The two largest sets also have very different anomaly structures: HTTP attacks form dense clusters of near-identical records, whereas fraudulent transactions are scattered.

Table 2. Benchmark datasets for static anomaly detection.

Dataset	Domain	Records	Features	Anomalies (%)
Cardio	Medical (fetal heart rate)	1,831	21	9.61
Cardiotocography	Medical (fetal heart rate)	2,114	21	22.04
Thyroid	Medical (thyroid function)	3,772	6	2.47
Waveform	Synthetic signal	3,443	21	2.90
Satimage-2	Remote sensing	5,803	36	1.22
Satellite	Remote sensing	6,435	36	31.64
Mammography	Medical imaging	11,183	6	2.32
Shuttle	Industrial control	49,097	9	7.15
HTTP (KDDCup99)	Network intrusion	567,498	3	0.39
Credit-card fraud	Financial transactions	284,807	29	0.17

Sources: ODDS [46] via ADBench [20]; HTTP from KDD Cup 1999 [5]; credit-card data from Dal Pozzolo *et al.* [4].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

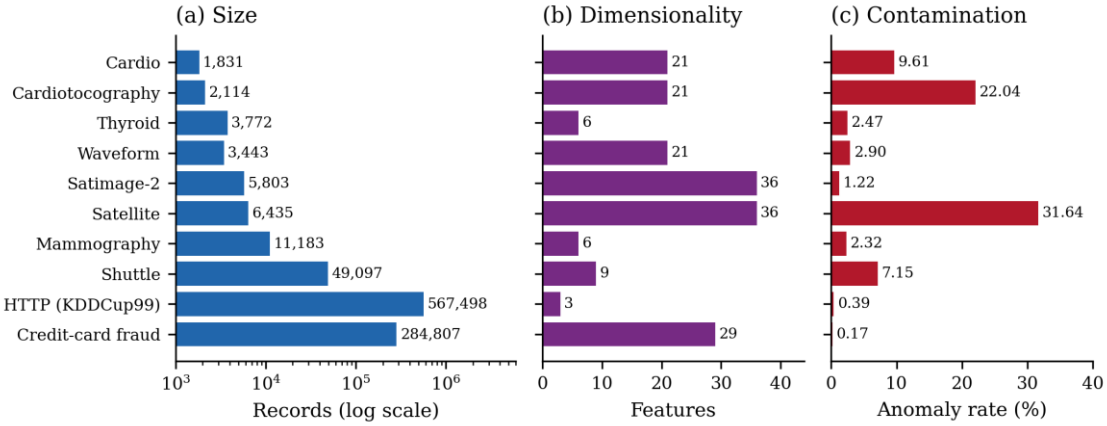


Figure 2. Characteristics of the benchmark datasets: (a) number of records on a log scale, (b) number of features and (c) proportion of anomalies. Sets larger than 30,000 records were analysed as stratified subsamples of 30,000 (Section 3.3).

For the second layer we used eight real series from NAB [40]: ambient temperature in an office, CPU utilisation of a misconfigured auto-scaling group, CPU utilisation and request latency of Amazon EC2 instances, temperature of an industrial machine before a failure, New York taxi demand, keystroke timing of a rogue agent and road traffic speed. The series contain between 1,127 and 22,695 points and 21 labelled anomaly windows after the probationary period. NAB has known weaknesses, including generous window labels [21], which we take into account when interpreting the results.

Pre-processing

All features were standardised to zero mean and unit variance [47]. To keep the comparison fair to methods with quadratic cost, datasets with more than 30,000 records were reduced to a stratified random subsample of 30,000 that preserved the anomaly rate; a new subsample was drawn for each of the five repetitions (seeds 0–4). The cost of each method on the full data is examined separately in the scalability experiment. For each stream the first 15% of points formed a probationary period, following NAB practice [40]: models were fitted and thresholds set on this period only, and alarms were counted only afterwards.

Static detectors

We selected seven detectors that represent the main families described in Section 2 and are widely used in practice [20, 37]. Isolation Forest (IForest) [13] builds 100 random trees on subsamples of 256 records. If $h(\mathbf{x})$ is the path length needed to isolate $\mathbf{x}$ and $c(\psi)$ is the average path length of an unsuccessful search in a binary tree of ψ points, the score is

$$s(\mathbf{x}) = 2^{-\mathrm{E}[h(\mathbf{x})] / c(\psi)} \quad (1)$$

so that points isolated after few splits receive scores close to 1. ECOD [33] computes, for each feature j, the left and right empirical tail probabilities of the observed value, $\hat{F}_j(x_j)$ and $1 - \hat{F}_j(x_j)$, and combines them as

$$O(\mathbf{x}) = \max\{ -\Sigma_j \log \hat{F}_j(x_j),\ -\Sigma_j \log[1 - \hat{F}_j(x_j)],\ -\Sigma_j \log T_j(x_j) \} \quad (2)$$

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

where T_j selects the left or right tail according to the sign of the skewness of feature j. The k-nearest-neighbour detector (kNN) uses the mean distance to the five nearest neighbours [29], and LOF uses 20 neighbours to compare local reachability densities [14]. The one-class SVM (OCSVM) with a radial basis kernel and $\nu = 0.1$ [31] was fitted on a random sample of 3,000 records to keep its cost manageable and then applied to all records. The PCA detector keeps the components that explain 90% of the variance and scores each record by its squared reconstruction error [32]. The autoencoder (AE) is a fully connected network with three hidden layers (sizes $d/2$, $d/4$ and $d/2$, with minimum widths of 8 and 4), trained with Adam [48] for 60 epochs on up to 20,000 records to reconstruct its input; its score is the mean squared reconstruction error [35].

Rank-averaged ensemble (RAE)

Raw scores from different detectors live on different scales, so they must be normalised before they are combined [23]. We converted each detector's scores to normalised ranks, $u_m(\mathbf{x}) = \text{rank}(s_m(\mathbf{x})) / n$, and averaged them over a set M of members:

$$\text{RAE}(\mathbf{x}) = (1/|M|) \sum_{m \in M} u_m(\mathbf{x}) \quad (3)$$

The RAE ensemble was defined before any results were seen, with one member from each of four families: IForest (isolation), ECOD (probabilistic), kNN (proximity) and AE (reconstruction), following the advice that ensemble members should be accurate and diverse [23, 24]. For comparison we also report the average of all seven detectors (All-7). To measure each member's contribution we repeated the RAE with each member left out in turn.

3.6 Streaming detection of emerging patterns

Let x_t be the observation at time t, standardised with the mean and standard deviation of the probationary period. The proposed detector first fits a ridge-regularised autoregressive model of order 24 on the probationary period [38, 39] and computes one-step residuals

$$r_t = x_t - \sum_{i=1}^{24} \varphi_i \, x_{t^*-*i} \quad (4)$$

which are scaled by a robust estimate $\hat{\sigma} = 1.4826 \times \text{MAD}$ of the probationary residuals [49]. Large values of $|r_t|/\hat{\sigma}$ indicate sudden deviations. To capture changes that build up gradually, the scaled residuals $z_t = r_t/\hat{\sigma}$ feed a two-sided CUSUM statistic [12, 16]:

$$C^+_t = \max(0, C^+_{t^*-1} + z_t - k), \quad C^-_t = \max(0, C^-_{t^*-1} - z_t - k) \quad (5)$$

with reference value $k = 0.5$, so that the statistic grows only when deviations persist in one direction. Each component is divided by its own threshold τ, the 99.9th percentile of its values during the probationary period, and the fused score

$$S_t = \max(|r_t| / (\hat{\sigma}\, \tau_{\text{res}}), \ \max(C^+_t, C^-_t) / \tau_{\text{cusum}}) \quad (6)$$

raises an alarm when $S_t > 1$. We compared this AR–CUSUM fusion with four baselines that use the same threshold rule: a rolling z-score over the previous 96 points; an Isolation Forest fitted on lag vectors of 12 previous values plus the current value [13]; the AR residual alone; and the CUSUM statistic alone.

Evaluation protocol

Static detectors were evaluated with the area under the ROC curve (ROC-AUC), average precision (AP), which is more informative when anomalies are rare [44, 45], and precision at n (P@n), the precision among the n highest-scoring records when n

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

equals the number of true anomalies [18]. Each value is the mean of five repetitions. Methods were ranked on each dataset and compared with the Friedman test [50] followed by the Nemenyi post-hoc test [51], whose critical difference for $k = 9$ methods and $N = 10$ datasets at $\alpha = 0.05$ is

$$\mathrm{CD} = q_\alpha \sqrt{k(k+1)/(6N)} = 3.102 \times \sqrt{(90/60)} = 3.80 \qquad (7)$$

as recommended by Demšar [27]; selected pairs were also compared with the Wilcoxon signed-rank test [52]. Scalability was measured as wall-clock time on random subsamples of the credit-card data of 2,000 to 80,000 records. Streaming detectors were evaluated at the level of events rather than points, to avoid the bias of point adjustment [22]. Consecutive alarms less than ten points apart were merged into one alarm event. A labelled window counted as detected if any alarm event overlapped it; alarm events that overlapped no window counted as false alarms; and the detection delay was the time from the start of the window to the first alarm, expressed as a fraction of the window length. Point-wise ROC-AUC and AP against the window labels are reported for completeness.

Implementation

All experiments were written in Python with scikit-learn [47], NumPy and SciPy, and run on a single CPU core with 3 GB of memory. ECOD and the rank-averaging ensemble were implemented from their published descriptions [23, 33]; as a check, our ECOD reached a ROC-AUC of 0.900 on Mammography, close to the value of about 0.91 listed for the PyOD implementation [37]. The complete experiment, including all five repetitions and the streaming study, ran in less than 15 minutes.

Results

Detection accuracy on static data

Table 3 and Figure 3 give the ROC-AUC of every detector on every dataset. Isolation Forest had the highest mean ROC-AUC (0.880) and was best on five of the ten datasets, including Shuttle (0.998) and HTTP (1.000). ECOD (0.838), the RAE ensemble (0.834) and OCSVM (0.833) formed a second group, followed by All-7 (0.819), AE (0.768), PCA (0.695), kNN (0.687) and LOF (0.582). Results were stable across repetitions; the largest standard deviation for Isolation Forest on any dataset was 0.029. Two datasets were hard for every method: no detector exceeded 0.75 on Cardiotocography or Satellite, where the anomaly class covers 22% and 32% of the records and is therefore not rare in the sense the detectors assume [1, 42].

Table 3. ROC-AUC (mean of five repetitions). The best value on each dataset is in bold.

Dataset	IForest	ECOD	kNN	LOF	OCSVM	PCA	AE	RAE	All-7
Cardio	**0.932**	0.881	0.643	0.547	0.874	0.727	0.810	0.861	0.836
Cardiotocography	0.686	0.648	0.481	0.524	**0.741**	0.529	0.560	0.600	0.612

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Dataset	IForest	ECOD	kNN	LOF	OCSVM	PCA	AE	RAE	All-7
Thyroid	**0.978**	0.921	0.957	0.674	0.927	0.585	0.845	0.947	0.901
Waveform	0.718	**0.737**	0.729	0.708	0.657	0.536	0.554	0.722	0.702
Satimage-2	0.994	0.884	0.923	0.542	**0.995**	0.796	0.845	0.975	0.932
Satellite	**0.696**	0.496	0.657	0.546	0.544	0.648	0.638	0.673	0.643
Mammography	0.859	**0.900**	0.833	0.720	0.763	0.622	0.712	0.855	0.845
Shuttle	**0.998**	0.994	0.638	0.530	0.894	0.591	0.874	0.936	0.908
HTTP (KDDCup99)	**1.000**	0.991	0.060	0.358	0.993	0.996	0.922	0.835	0.879
Credit-card fraud	0.937	0.930	**0.952**	0.672	0.943	0.919	0.922	0.941	0.928
Mean	0.880	0.838	0.687	0.582	0.833	0.695	0.768	0.834	0.819
Mean rank	1.9	3.8	5.1	7.9	4.0	7.1	6.6	3.6	5.0

IForest, Isolation Forest; ECOD, empirical-CDF outlier detection; kNN, *k*-nearest neighbours; LOF, local outlier factor; OCSVM, one-class SVM; PCA, principal component reconstruction; AE, autoencoder; RAE, rank-averaged ensemble of IForest, ECOD, kNN and AE; All-7, rank average of all seven detectors. Lower mean rank is better.



Figure 3. ROC-AUC heat map for nine detectors on ten datasets (mean of five

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

repetitions). Bold values mark the best detector on each dataset; the white line separates single detectors from ensembles. Proximity-based detectors (kNN, LOF) collapse on HTTP, whose attacks form dense clusters.

The failures of the proximity methods are worth noting. On HTTP, kNN reached a ROC-AUC of only 0.060 and LOF 0.358, both far below random, whereas Isolation Forest, PCA and OCSVM all exceeded 0.99. The attacks in HTTP are thousands of near-identical connections, so each attack record has close neighbours of the same kind and looks perfectly normal to a distance-based score. This is the masking effect described in the literature on clustered anomalies [8, 13], and it shows why a detector that ranks well on average can still be unsafe for a particular application.

Precision on rare anomalies

Because the anomaly rate is below 3% in six of the ten datasets, AP and P@*n* are closer to what an analyst experiences when reviewing the top of an alert list [45]. Table 4 shows that the advantage of Isolation Forest is larger on these measures: its mean AP of 0.541 was 0.19 higher than that of the next detector, OCSVM (0.353), and its mean P@*n* was 0.533. The ensembles lost more ground here than on ROC-AUC (RAE 0.287, All-7 0.224), because rank averaging pulls the highest scores of the best member towards the middle when other members disagree [24]. Figure 4 illustrates the effect: on Cardio and Thyroid the RAE curve lies between those of its members, while on Mammography and the credit-card data ECOD gives clearly the highest precision at low recall.

Table 4. Average precision (mean of five repetitions) and summary of P@*n*.

Dataset	IForest	ECOD	kNN	LOF	OCSVM	PCA	AE	RAE	All-7
Cardio	**0.563**	0.464	0.250	0.155	0.407	0.268	0.365	0.407	0.333
Cardiotocography	**0.427**	0.373	0.278	0.276	0.413	0.279	0.317	0.352	0.328
Thyroid	**0.518**	0.220	0.262	0.075	0.250	0.030	0.179	0.249	0.121
Waveform	0.057	0.062	**0.106**	0.076	0.052	0.040	0.038	0.066	0.069
Satimage-2	**0.924**	0.080	0.317	0.032	0.895	0.220	0.189	0.342	0.180
Satellite	**0.657**	0.343	0.508	0.376	0.460	0.519	0.479	0.552	0.507
Mammography	0.194	**0.422**	0.156	0.091	0.173	0.105	0.093	0.203	0.195
Shuttle	**0.982**	0.912	0.164	0.122	0.381	0.122	0.462	0.554	0.426

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Dataset	IForest	ECOD	kNN	LOF	OCSVM	PCA	AE	RAE	All-7
HTTP (KDDCup99)	**0.962**	0.297	0.020	0.045	0.363	0.598	0.103	0.033	0.037
Credit-card fraud	0.130	**0.287**	0.109	0.007	0.138	0.050	0.041	0.113	0.046
Mean AP	0.541	0.346	0.217	0.125	0.353	0.223	0.227	0.287	0.224
Mean rank (AP)	2.0	4.0	5.4	7.7	4.0	6.3	6.3	3.8	5.5
Mean P@*n*	0.533	0.328	0.233	0.146	0.347	0.203	0.237	0.291	0.235

P@*n*: precision among the *n* highest-scoring records, with *n* equal to the number of true anomalies [18]. Abbreviations as in Table 3.

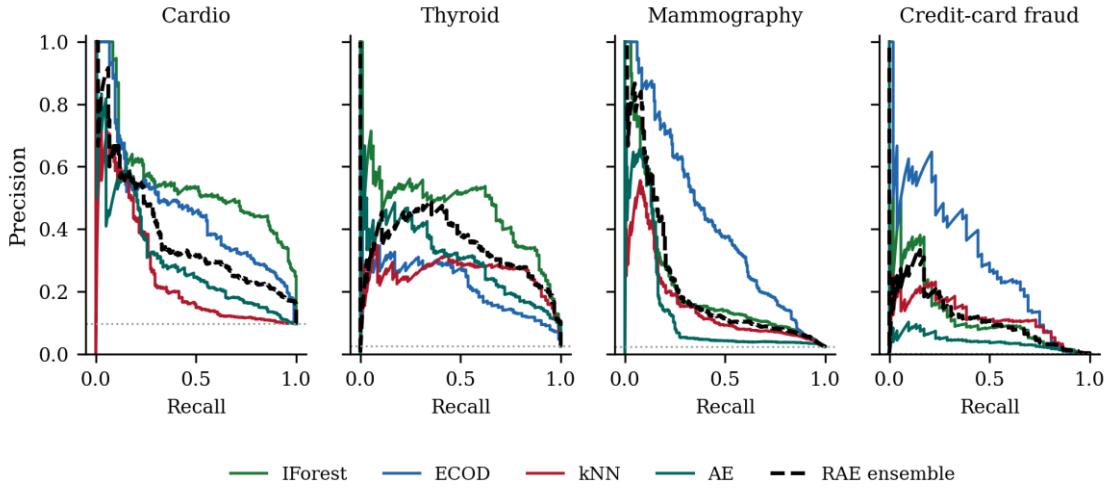


Figure 4. Precision–recall curves on four datasets (first repetition) for the four RAE members and the ensemble. The dotted line marks the anomaly rate, the precision of a random ranking [45].

Statistical comparison

The Friedman test rejected the hypothesis that all nine methods perform alike, both for ROC-AUC ($\chi^2 = 39.2$, $p = 4.5 \times 10^{-6}$) and for AP ($\chi^2 = 31.4$, $p = 1.2 \times 10^{-4}$) [27, 50]. The critical difference diagrams in Figure 5 show the mean ranks. With a critical difference of 3.80, Isolation Forest (mean rank 1.9) was significantly better than LOF (7.9), PCA (7.1) and AE (6.6) for ROC-AUC, while its differences from RAE (3.6), ECOD (3.8) and OCSVM (4.0) did not reach significance under the conservative Nemenyi correction [51]. A direct paired comparison was clearer: Isolation Forest beat RAE on eight of ten datasets (Wilcoxon $W = 4$, $p = 0.014$) [52], whereas RAE and All-7 did not differ significantly ($p = 0.105$).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

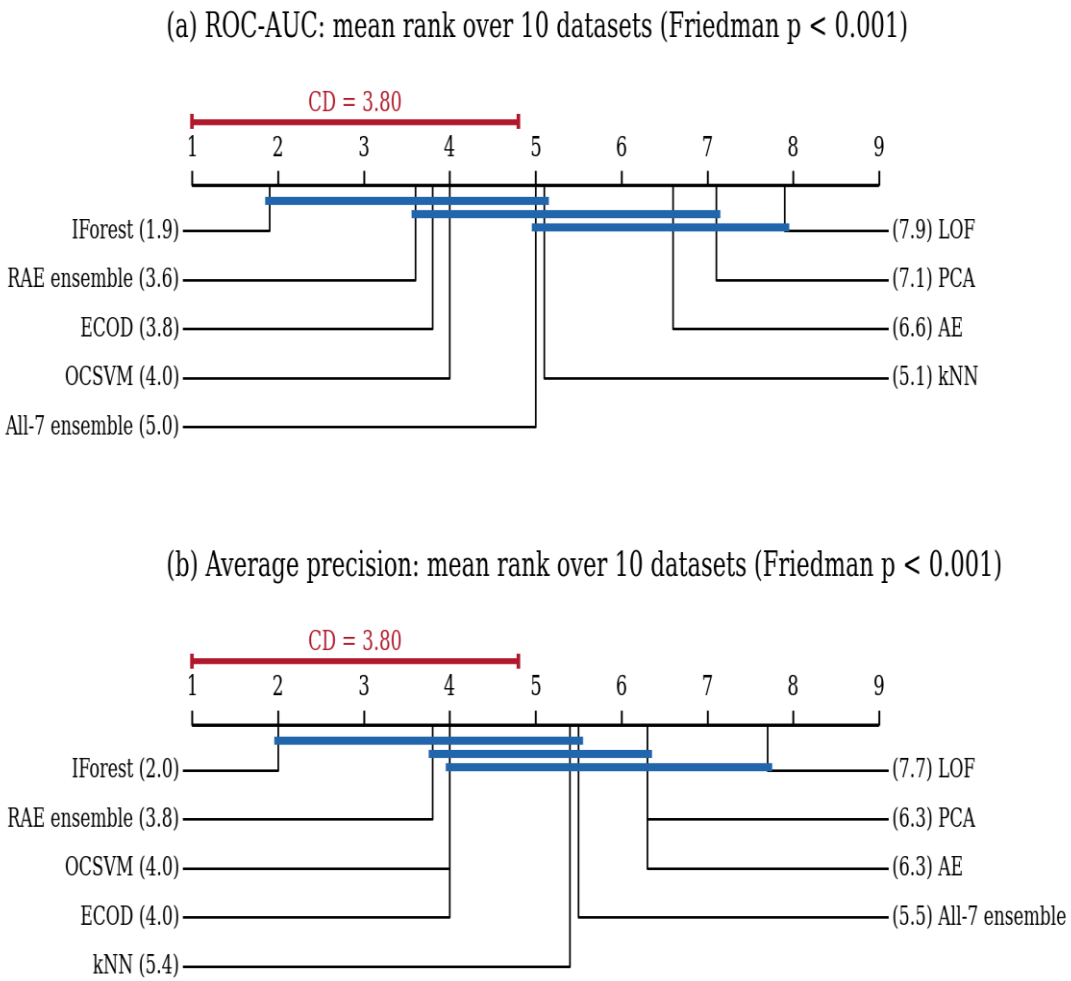


Figure 5. Critical difference diagrams (Nemenyi test, $\alpha = 0.05$) for (a) ROC-AUC and (b) average precision [27, 51]. Methods joined by a horizontal blue bar are not significantly different; lower ranks are better.

Ensemble ablation

Leaving each member out of the RAE ensemble in turn shows how much each contributed (Figure 6a). Removing Isolation Forest lowered the mean ROC-AUC from 0.834 to 0.808 and removing ECOD lowered it to 0.811, so these two members carried most of the ensemble's accuracy. Removing the autoencoder made almost no difference (0.835). Removing kNN, however, *raised* the mean ROC-AUC to 0.855, mainly because it eliminated the collapse on HTTP described in Section 4.1. An ensemble therefore does not simply average out a bad member; when one member ranks the anomalies at the bottom, it actively drags them down [23, 24]. Figure 6b places all methods by their mean ROC-AUC and mean AP and shows that the ensembles sit well below Isolation Forest on precision.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

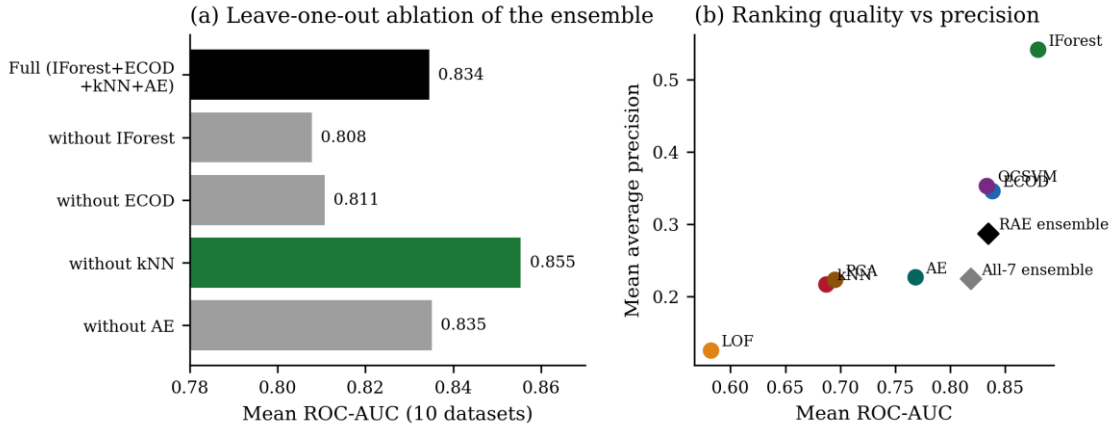


Figure 6. (a) Mean ROC-AUC of the RAE ensemble with each member left out in turn; removing kNN improves the ensemble. (b) Mean ROC-AUC against mean average precision for all methods across ten datasets.

Scalability and cost

Figure 7a shows how run time grew with the number of records on the credit-card data. Isolation Forest, ECOD and PCA scaled close to linearly and needed 0.52 s, 0.89 s and 0.07 s for 80,000 records. The cost of OCSVM (1.5 s) was held down by fitting on a fixed sample, and the autoencoder levelled off at about 2.8 s because it was trained on at most 20,000 records. The two neighbour-based detectors grew roughly quadratically: their time rose from 0.04 s at 2,000 records to 31.2 s at 80,000, a factor of about 780 for a forty-fold increase in data. Figure 7b combines cost and accuracy. Isolation Forest is the only method that is both among the fastest and the most accurate, whereas the RAE ensemble pays for all of its members, including the expensive neighbour search, without matching Isolation Forest's accuracy [13, 20].

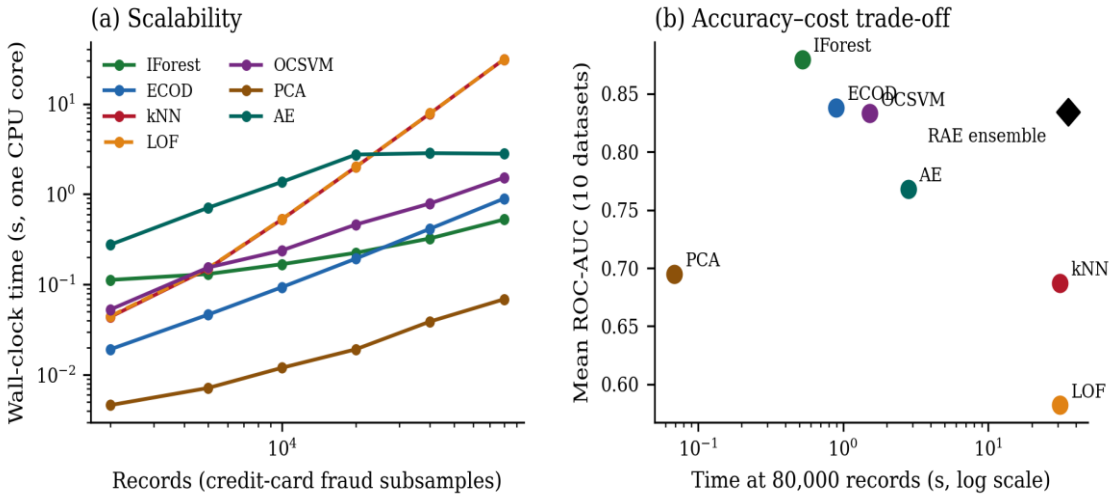


Figure 7. (a) Wall-clock time against number of records on subsamples of the credit-card data (single CPU core, log–log scale). (b) Mean ROC-AUC across the ten datasets against run time at 80,000 records; the ensemble time is the sum of its members.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Early detection of emerging patterns in streams

Table 5 summarises the streaming results over the eight series and 21 labelled windows, and Table 6 breaks the detections down by series. The AR–CUSUM fusion detected all 21 windows and did so earliest, with a mean delay of 28.3% of the window length. CUSUM alone detected 20 windows with fewer false alarms (1.53 per 1,000 points), which shows that the accumulation step does most of the work in recognising slow changes [12, 16]. The AR residual alone found only 14 windows: it reacted to abrupt jumps but missed both windows on the ambient-temperature and rogue-agent series, where the departure from normal behaviour builds up over many steps. The windowed Isolation Forest had the highest point-wise ROC-AUC (0.681) but raised the most false alarms (2.81 per 1,000 points), many of them on the strongly seasonal taxi series. The rolling *z*-score raised the fewest false alarms but missed 8 of the 21 windows and was the slowest to react.

Table 5. Event-level streaming results on eight NAB series (58,990 evaluated points, 21 anomaly windows).

Detector	Windows detected	Recall	False-alarm events	False alarms per 1,000 pts	Mean delay (% of window)	Point ROC-AUC	Point AP
Rolling z-score	13 / 21	0.619	42	0.71	43.0	0.512	0.154
Windowed IForest	18 / 21	0.857	166	2.81	38.9	0.681	0.359
AR residual	14 / 21	0.667	85	1.44	34.8	0.604	0.221
CUSUM	20 / 21	0.952	90	1.53	32.5	0.602	0.287
AR–CUSUM fusion (proposed)	21 / 21	1.000	115	1.95	28.3	0.597	0.284

Alarms closer than ten points were merged into one event. Delay is measured from the start of a window to the first alarm inside it and averaged over detected windows. Point-wise measures are given for reference only, because they are known to be unreliable for windowed labels [21, 22].

Table 6. Windows detected and false-alarm events (in parentheses) per series.

Series	Points	Windows	Rolling z	Win. IForest	AR resid.	CUSUM	AR–CUSUM
Ambient temperature	7,267	2	1 (14)	2 (15)	0 (2)	2 (4)	2 (5)
CPU, auto-scaling group	18,050	1	0 (2)	1 (47)	1 (27)	1 (29)	1 (42)
EC2 CPU utilisation	4,032	1	1 (3)	1 (10)	1 (7)	1 (0)	1 (0)

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Series	Points	Windows	Rolling z	Win. IForest	AR resid.	CUSUM	AR–CUSUM
EC2 request latency	4,032	3	3 (9)	3 (21)	3 (27)	3 (33)	3 (38)
Machine temperature	22,695	3	2 (7)	2 (2)	3 (13)	2 (10)	3 (13)
NYC taxi demand	10,320	5	1 (0)	5 (67)	2 (2)	5 (13)	5 (14)
Rogue agent key hold	1,882	2	2 (5)	1 (0)	0 (2)	2 (0)	2 (0)
Traffic speed	1,127	4	3 (2)	3 (4)	4 (5)	4 (1)	4 (3)

Source: Numenta Anomaly Benchmark [40]. The first 15% of each series was used for fitting and thresholds and is excluded.

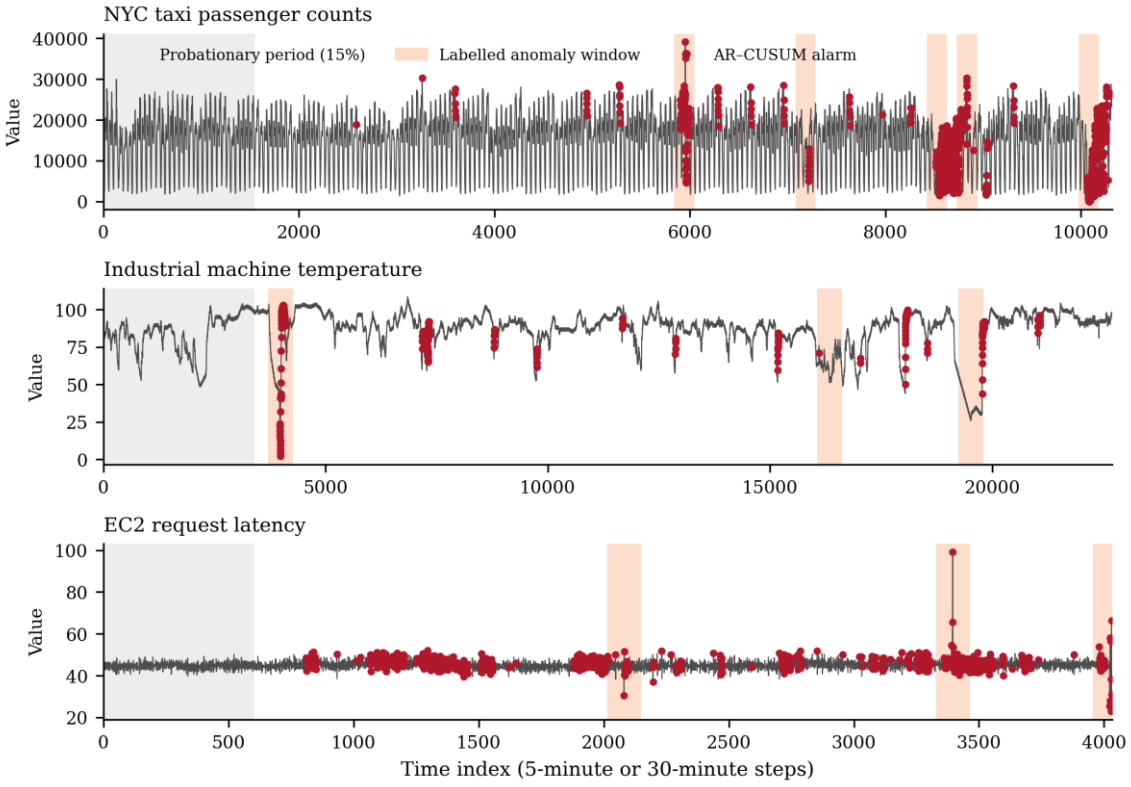


Figure 8. Three NAB series with labelled anomaly windows (shaded) and the alarms raised by the AR–CUSUM fusion (red points). The grey region is the probationary period used to fit the model and set the threshold [40]. Alarms outside the shaded windows count as false alarms.

Figure 8 shows typical behaviour. On the taxi series the detector fired inside all five windows, which cover holidays and a snowstorm, but also on several ordinary peaks. On the machine-temperature series it caught the slow decline that preceded the first failure well before the window ended, but it also responded to shorter dips that the labels do not mark. Figure 9 sets out the trade-off between recall, false alarms and timeliness: the fusion moves the detector towards full recall and earlier warning at a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

moderate increase in false alarms over CUSUM alone [7, 40].

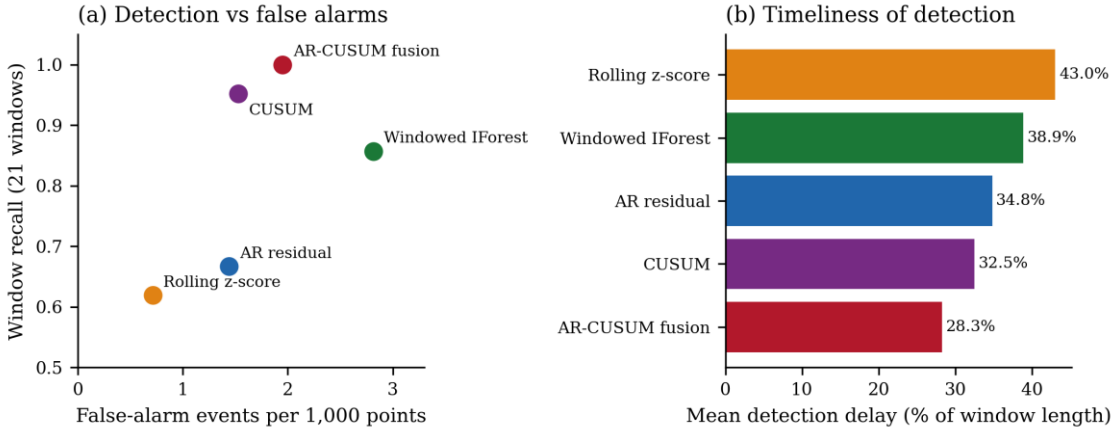


Figure 9. Streaming detectors compared on (a) window recall against false-alarm events per 1,000 points and (b) mean detection delay as a percentage of the anomaly window. The upper-left corner of (a) and short bars in (b) are desirable.

Discussion

Why a simple tree-based detector remains hard to beat

The clearest result of this study is that Isolation Forest, a method published in 2008 [13], was the most accurate and among the cheapest detectors on the ten datasets. This agrees with large benchmarks that place it in the leading group of unsupervised methods [18, 20]. Its strength comes from the way it scores points. Isolation depends on how separable a point is from the rest, not on how close it is to its neighbours, so it is not fooled when anomalies cluster together, as they do in HTTP. Because each tree is grown on a small subsample, the method also tolerates irrelevant features and contaminated training data better than density estimators do [13, 30]. ECOD was a strong second at an even lower cost and was the best choice on Mammography and the credit-card data, where anomalies show up as extreme values in individual features [33]. Its weakness on Satellite (0.496) reflects the opposite case, in which anomalies are unusual combinations of ordinary values that no single tail can reveal.

Ensembles inherit the failures of their members

Rank averaging is often presented as a safe default when no labels are available to pick a detector [23, 24]. Our ensemble did achieve the second-best mean rank, which supports that view, but it never matched Isolation Forest on average and lost most ground on precision, the measure that matters when analysts review a short list of alerts. The ablation explains why. An average of ranks treats each member as equally trustworthy, so a member that inverts the ranking on one dataset, as kNN did on HTTP, pulls the true anomalies towards the middle. Diversity helps only when the members fail in different places and none fails badly [24]. In practice this argues for ensembles built from members that are each robust on their own, or for combination rules that limit the influence of a member that disagrees strongly with the others, rather than for adding as many detectors as possible [23, 42].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Cost as a design constraint

Large-scale analytics is constrained as much by computation as by accuracy. The quadratic growth of the neighbour-based methods means that a data set ten times larger takes about a hundred times longer, so at a million records these methods would need more than an hour on one core while Isolation Forest and ECOD would still finish in seconds. Approximate neighbour search and distributed implementations can narrow this gap [33, 37], but our results suggest that for tabular monitoring the extra cost buys little: kNN and LOF were among the least accurate detectors on average. The autoencoder, trained on a capped sample, was affordable but less accurate than the shallow methods, which is consistent with reports that deep detectors offer little advantage on low-dimensional tabular data [9, 15, 20].

Detecting emerging patterns early

In the streaming experiments the most important ingredient was accumulation. A detector that looks only at the size of the current residual responds to spikes but misses changes that unfold over many steps, and many real failures start that way: a machine heats up slowly, a misconfigured service degrades gradually. The CUSUM statistic adds up small deviations in a consistent direction, which is why it detected 20 of the 21 windows against 14 for the residual alone [12, 16]. Fusing the two gave full recall and the shortest delays, because sharp anomalies are caught by the residual term and slow ones by the accumulated term. The price was a moderate rise in false alarms, many of them on series with strong daily or weekly cycles that a 24-lag autoregression captures only partly [39]. The windowed Isolation Forest shows the opposite trade-off: it ranks points well overall but, without a model of the expected next value, it flags many ordinary peaks. For operational use these alarm rates would need to be matched to the cost of investigating an alert, and seasonal forecasting models would be the natural first improvement [38, 39].

Practical recommendations

Table 7 turns the findings into guidance for practitioners who must choose detectors for large data without labels. The recommendations are deliberately conservative: they favour methods that were accurate across all datasets and cheap enough to rerun often, over methods that win only on particular data.

Table 7. Practical guidance derived from the experiments.

Situation	Recommended choice	Evidence in this study
Default for large tabular data, no labels	Isolation Forest	Best mean ROC-AUC (0.880) and AP (0.541); 0.52 s for 80,000 records
Anomalies expected as extreme values in single features (e.g. fraud amounts)	ECOD, alone or alongside Isolation Forest	Best AP on Mammography and credit-card fraud; parameter-free; under 1 s for 80,000 records

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3(2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Situation	Recommended choice	Evidence in this study
Anomalies that may form dense clusters (e.g. coordinated attacks)	Avoid kNN and LOF; prefer isolation or projection methods	kNN 0.060 and LOF 0.358 on HTTP versus ≥ 0.99 for IForest, PCA and OCSVM
Combining detectors	Rank-average only members that are individually robust; check by leave-one-out	Removing kNN raised RAE from 0.834 to 0.855
Streams with gradual degradation	Forecast residuals plus CUSUM accumulation	21/21 windows detected; shortest delay (28.3% of window)
Streams where false alarms are expensive	CUSUM alone, or raise the fusion threshold	20/21 windows with 1.53 false alarms per 1,000 points

Limitations

Several limitations should be kept in mind. The benchmark datasets are widely used but small by industrial standards, and some of their anomaly labels are derived from minority classes rather than from genuine faults, which may favour certain detectors [18, 42]. Large sets were analysed as subsamples of 30,000 records so that every method could run, so the accuracy results describe detection at that scale; the scalability experiment covers cost on larger samples but not accuracy. We used default or literature-recommended hyperparameters, because tuning without labels is itself an open problem, and different settings could change the ranking of individual methods [18, 20]. The streaming study used eight NAB series, whose windows are wide and sometimes imprecisely placed [21]; our event-level measures reduce but do not remove this problem. Finally, all experiments ran on a single CPU core, so the timing results compare methods with one another rather than estimate the speed of optimised production systems.

Future work

Three extensions follow directly from these results. The first is to replace uniform rank averaging with combination rules that weight or exclude members according to their agreement with the others, and to test whether such rules keep the robustness of an ensemble without its loss of precision [23, 24]. The second is to add seasonal forecasting and adaptive thresholds to the streaming layer and to evaluate it on the UCR anomaly archive, which was built to avoid the weaknesses of earlier benchmarks [21, 26]. The third is to run the framework on distributed infrastructure with data of tens of millions of records, where the cost differences reported here would decide which methods remain practical [33, 37].

Conclusion

We set out to identify which unsupervised methods give reliable early warning of anomalies and emerging patterns in large data, and at what cost. Across ten benchmark datasets, Isolation Forest was the most accurate and among the fastest detectors, with a

Online ISSN 3007-3197
Print ISSN 3007-3189

mean ROC-AUC of 0.880 and the best mean rank, and ECOD was a strong, cheap alternative for anomalies that show up in individual features [13, 33]. A rank-averaged ensemble was robust but did not beat its best member, and it was damaged by a member that failed on clustered anomalies; ensembles should therefore be assembled and checked with care [23, 24]. For streams, combining forecast residuals with cumulative-sum accumulation detected every labelled anomaly window and did so earlier than the baselines, at a moderate false-alarm cost [16, 40]. These findings support a simple architecture for large-scale predictive monitoring: a fast tree-based detector for records, a residual-and-accumulation detector for streams, and evaluation that looks at precision, cost and timeliness rather than at a single accuracy figure.

References

[1] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. *ACM Comput Surv*. 2009;41(3):15.

[2] Hodge VJ, Austin J. A survey of outlier detection methodologies. *Artif Intell Rev*. 2004;22(2):85–126.

[3] Bolton RJ, Hand DJ. Statistical fraud detection: a review. *Stat Sci*. 2002;17(3):235–255.

[4] Dal Pozzolo A, Caelen O, Johnson RA, Bontempi G. Calibrating probability with undersampling for unbalanced classification. In: *Proc IEEE Symp Ser Comput Intell (SSCI)*; 2015. p. 159–166.

[5] Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: *Proc IEEE Symp Comput Intell Secur Def Appl (CISDA)*; 2009. p. 1–6.

[6] Hundman K, Constantinou V, Laporte C, Colwell I, Soderstrom T. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In: *Proc 24th ACM SIGKDD Int Conf Knowl Discov Data Min*; 2018. p. 387–395.

[7] Ahmad S, Lavin A, Purdy S, Agha Z. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*. 2017;262:134–147.

[8] Aggarwal CC. *Outlier Analysis*. 2nd ed. Cham: Springer; 2017.

[9] Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, Kloft M, et al. A unifying review of deep and shallow anomaly detection. *Proc IEEE*. 2021;109(5):756–795.

[10] Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. *ACM Comput Surv*. 2014;46(4):44.

[11] Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: a review. *IEEE Trans Knowl Data Eng*. 2019;31(12):2346–2363.

[12] Basseville M, Nikiforov IV. *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs: Prentice Hall; 1993.

[13] Liu FT, Ting KM, Zhou ZH. Isolation forest. In: *Proc 8th IEEE Int Conf Data Mining (ICDM)*; 2008. p. 413–422.

[14] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: *Proc ACM SIGMOD Int Conf Manag Data*; 2000. p. 93–104.

[15] Pang G, Shen C, Cao L, van den Hengel A. Deep learning for anomaly

Online ISSN 3007-3197
Print ISSN 3007-3189

detection: a review. *ACM Comput Surv*. 2021;54(2):38.

[16] Page ES. Continuous inspection schemes. *Biometrika*. 1954;41(1/2):100–115.

[17] Bifet A, Gavaldà R. Learning from time-changing data with adaptive windowing. In: *Proc SIAM Int Conf Data Min (SDM)*; 2007. p. 443–448.

[18] Campos GO, Zimek A, Sander J, Campello RJGB, Micenková B, Schubert E, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min Knowl Discov*. 2016;30(4):891–927.

[19] Goldstein M, Uchida S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One*. 2016;11(4):e0152173.

[20] Han S, Hu X, Huang H, Jiang M, Zhao Y. ADBench: anomaly detection benchmark. In: *Advances in Neural Information Processing Systems 35 (NeurIPS 2022), Datasets and Benchmarks Track*; 2022. p. 32142–32159.

[21] Wu R, Keogh EJ. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *IEEE Trans Knowl Data Eng*. 2023;35(3):2421–2429. doi:10.1109/TKDE.2021.3112126

[22] Kim S, Choi K, Choi HS, Lee B, Yoon S. Towards a rigorous evaluation of time-series anomaly detection. In: *Proc AAAI Conf Artif Intell*. 2022;36(7):7194–7201. doi:10.1609/aaai.v36i7.20680

[23] Zimek A, Campello RJGB, Sander J. Ensembles for unsupervised outlier detection: challenges and research questions. *ACM SIGKDD Explor Newsl*. 2014;15(1):11–22.

[24] Aggarwal CC, Sathe S. Theoretical foundations and algorithms for outlier ensembles. *ACM SIGKDD Explor Newsl*. 2015;17(1):24–47.

[25] Lazarevic A, Kumar V. Feature bagging for outlier detection. In: *Proc 11th ACM SIGKDD Int Conf Knowl Discov Data Min*; 2005. p. 157–166.

[26] Schmidl S, Wenig P, Papenbrock T. Anomaly detection in time series: a comprehensive evaluation. *Proc VLDB Endow*. 2022;15(9):1779–1797.

[27] Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res*. 2006;7:1–30.

[28] Hawkins DM. *Identification of Outliers*. London: Chapman and Hall; 1980.

[29] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: *Proc ACM SIGMOD Int Conf Manag Data*; 2000. p. 427–438.

[30] Hariri S, Kind MC, Brunner RJ. Extended isolation forest. *IEEE Trans Knowl Data Eng*. 2021;33(4):1479–1489.

[31] Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Comput*. 2001;13(7):1443–1471.

[32] Jolliffe IT. *Principal Component Analysis*. 2nd ed. New York: Springer; 2002.

[33] Li Z, Zhao Y, Hu X, Botta N, Ionescu C, Chen GH. ECOD: unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Trans Knowl Data Eng*. 2023;35(12):12181–12193. doi:10.1109/TKDE.2022.3159580

[34] Li Z, Zhao Y, Botta N, Ionescu C, Hu X. COPOD: copula-based outlier detection. In: *Proc 20th IEEE Int Conf Data Mining (ICDM)*; 2020. p. 1118–1123.

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

[35] Sakurada M, Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In: *Proc MLSDA 2014 Workshop on Machine Learning for Sensory Data Analysis*; 2014. p. 4–11.

[36] Malhotra P, Ramakrishnan A, Anand G, Vig L, Agarwal P, Shroff G. LSTM-based encoder-decoder for multi-sensor anomaly detection. arXiv:1607.00148; 2016.

[37] Zhao Y, Nasrullah Z, Li Z. PyOD: a Python toolbox for scalable outlier detection. *J Mach Learn Res*. 2019;20(96):1–7.

[38] Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th ed. Hoboken: Wiley; 2015.

[39] Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*. 3rd ed. Melbourne: OTexts; 2021.

[40] Lavin A, Ahmad S. Evaluating real-time anomaly detection algorithms – the Numenta Anomaly Benchmark. In: *Proc 14th IEEE Int Conf Mach Learn Appl (ICMLA)*; 2015. p. 38–44.

[41] Domingues R, Filippone M, Michiardi P, Zouaoui J. A comparative evaluation of outlier detection algorithms: experiments and analyses. *Pattern Recognit*. 2018;74:406–421.

[42] Emmott A, Das S, Dietterich T, Fern A, Wong WK. A meta-analysis of the anomaly detection problem. arXiv:1503.01158; 2015.

[43] Lai KH, Zha D, Xu J, Zhao Y, Wang G, Hu X. Revisiting time series outlier detection: definitions and benchmarks. In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021), Datasets and Benchmarks Track*; 2021.

[44] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: *Proc 23rd Int Conf Mach Learn (ICML)*; 2006. p. 233–240.

[45] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.

[46] Rayana S. *ODDS Library*. Stony Brook (NY): Stony Brook University, Department of Computer Sciences; 2016. Available from: http://odds.cs.stonybrook.edu

[47] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.

[48] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *Int Conf Learn Represent (ICLR)*; 2015. arXiv:1412.6980.

[49] Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Am Stat Assoc*. 1993;88(424):1273–1283.

[50] Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc*. 1937;32(200):675–701.

[51] Nemenyi PB. *Distribution-free multiple comparisons* [PhD thesis]. Princeton (NJ): Princeton University; 1963.

[52] Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1945;1(6):80–83