Fine-Tuning Detection Criteria for Enhancing Anomaly Detection in Time Series

Edson Pinto Sobrinho[1], **Jéssica Souza**[1], **Janio Lima** [1], **Lucas Giusti** [1], **Eduardo Bezerra** [1], **Rafaelli Coutinho** [1], **Lais Baroni** [1], **Esther Pacitti**[2], **Fabio Porto**[3], **Kele Belloze**[1], **Eduardo Ogasawara**[1]

[1]CEFET/RJ - Federal Center for Technological Education of Rio de Janeiro

[2]University of Montpellier / INRIA

[3]LNCC - National Laboratory for Scientific Computing

{edson.sobrinho, jessica.souza, janio.lima, lucas.giusti}@aluno.cefet-rj.br,

{ebezerra, rafaelli.coutinho, lais.baroni, kelle.belloze}@cefet-rj.br,

esther.pacitti@lirmm.fr, fporto@lncc.br, eogasawara@ieee.org

Abstract. *Anomaly detection is the problem of identifying observations that do not conform to the typical ones in a time series. Detection methods implicitly define detection criteria, such as deviation measures, filter thresholds, and candidate anomaly selection strategies. Choosing inappropriate criteria results in inaccurate outputs, generating spurious alerts or missing events. Adjusting these criteria is essential for monitoring systems. To address this challenge, this paper explores the fine-tuning of deviation measures, filter thresholds, and candidate selection strategies. Experimental results show that the proper choice of criteria significantly improves anomaly detection performance, often with greater impact than changing the detection methods.*

1. Introduction

Anomaly detection (AD) is the problem of identifying observations that do not conform to the typical ones in a time series [Chandola et al., 2009]. Anomalies indicate problems in monitored systems, such as equipment failures, changes in user behavior, or shifts in environmental conditions [Ariyaluran Habeeb et al., 2019]. A central challenge in AD is designing methods that adapt to different data characteristics to minimize spurious alerts and maximize the detection of events [Lima et al., 2024; Ariyaluran Habeeb et al., 2019].

False positives (FP) lead to unnecessary interventions, while false negatives (FN), corresponding to missed anomalies, may cause losses or failures [Ibidunmoye et al., 2015]. Balancing FP and FN requires methodological design and a deep understanding of the data. Some detection measures, such as the F_1 score, aim to balance FP and FN. However, a method may present balanced results without reaching adequate levels of $\mathrm{precision}$ and recall for specific domains [Dixit et al., 2022; Cook et al., 2020; Ahmed et al., 2016].

Several decisions that influence detection results are often implicit in anomaly detection methods. These include the choice of deviation measures, the definition of filter thresholds, and the selection among candidate anomalies when multiple nearby observations are detected. Although not always explicitly discussed, these elements are critical for accurate detection [Ogasawara et al., 2025].

Deviation measures define how the difference between observed and expected values is computed. Different measures affect detection differently. In the literature, squared deviations are suited for data with a normal distribution. In contrast, absolute deviations, such as mean absolute error, are more robust when the data distribution deviates from normality [Gorard, 2005; Hodson, 2022].

Filter thresholds determine the point at which a deviation is considered sufficient to characterize an anomaly. Filters are often based on distributional assumptions, such as boxplot limits or thresholds derived from a Gaussian distribution [Han et al., 2022; Aggarwal, 2016]. Although filters are more explicitly addressed in detection methods, their role and sensitivity are not thoroughly explored.

Following that, after deviations are computed and filters are applied, groups of candidate anomalies may emerge. In such cases, it is necessary to decide which observations to select as anomalies. Selection strategies include choosing the first detected anomaly in a group or the one with the largest deviation [Lima et al., 2022, 2024]. The choice among candidates influences detection performance, but is usually predefined and lacks adaptation.

These detection criteria, deviation measures, filter thresholds, and candidate selection, are often applied in a fixed manner, without considering data characteristics or application goals [Lima et al., 2022; Souza et al., 2024; Talagala et al., 2020]. Using fixed criteria across different domains leads to performance degradation when data properties differ from those under which the method was developed.

To address these challenges, three research questions are proposed: (Q1) Does changing the deviation measure impact detection? (Q2) Are detections sensitive to changes in filter criteria? (Q3) Are detections sensitive to the choice of candidate anomaly when multiple nearby observations are detected?

This paper examines the fine-tuning of detection criteria to adapt anomaly detection methods. Techniques for deviation measurement, filter definition, and candidate selection are evaluated to answer the proposed research questions and address the observed gap in the literature. Experimental results demonstrate that adjusting detection criteria according to the application improves anomaly detection performance, often with greater impact than changing the detection methods.

Besides this introduction, the paper is organized into four additional sections. Section 2 provides background on anomaly detection. Section 3 presents the methodology for adapting detection methods. Section 4 describes the experiments and their results. Section 5 concludes the study.

2. Literature Review

2.1. Overview of Anomaly Detection in Time Series

Time series anomaly detection (AD) aims to identify observations or intervals that deviate from the expected distribution. Anomalies hold specific meanings within each domain

[Talagala et al., 2020; Ren et al., 2019], often reflecting deviations in trend, volatility, or other aspects of temporal dynamics [Ariyaluran Habeeb et al., 2019; Ogasawara et al., 2025].

AD methods employ diverse approaches, including distribution analysis, regression, and classification [Chandola et al., 2009; Truong et al., 2020]. A common task across these methods is defining filters beyond which observations are considered anomalous. Detection is generally based on deviations between observed and predicted values or between observed and model-adjusted values [Lima et al., 2024].

In prediction-based detection, methods forecast the next time series value and compare it with the actual observation. A high deviation indicates an anomaly. In contrast, in model-adjusted detection, the model is fitted directly to the observed series, and anomalies are identified based on deviations between the observations and the fitted values [Talagala et al., 2020; Zhang et al., 2020; Ren et al., 2019].

In both approaches, detection performance is sensitive to the chosen deviation measure and filter criterion. The methodology used to compute deviations directly affects detection effectiveness. Section 2.2 discusses deviation measures. Given a deviation measure, the filter criterion that separates anomalies from typical observations is discussed in Section 2.3. When multiple nearby candidate anomalies are detected, the strategy for selecting candidates is addressed in Section 2.4.

2.2. Deviation Measures Criterion

Deviation measures quantify the difference between observations and the values predicted or adjusted by AD methods [Han et al., 2022; Gorard, 2005]. These measures are generally categorized into two types: (i) absolute deviation (ABS) and (ii) squared deviation (SQD).

The ABS is computed as the absolute difference between an observation x_i and its expected value $\hat{x}_i$: $w_i = |x_i - \hat{x}_i|$. Alternatively, the squared deviation amplifies larger discrepancies by squaring the difference: $w_i = (x_i - \hat{x}_i)^2$. The choice of deviation measure influences detection results. Statistically, SQD aligns with the assumption of normally distributed errors, whereas ABS is more robust in the presence of heavy-tailed or noisy data [Gorard, 2005; Hodson, 2022].

Deviation measures are used as core components of AD methods, whether explicitly in deviation-based techniques [Aggarwal, 2016; Han et al., 2022], or internally within classification-based [Cauteruccio et al., 2021], clustering-based [Li et al., 2021], and statistical-based approaches [Salles et al., 2020; Lima et al., 2022].

2.3. Filter Criterion

Filter criteria define thresholds that separate typical observations from anomalies. They act as decision boundaries for flagging anomalies based on the residuals between observed and expected values [Ahmad et al., 2017; Hasani, 2017].

A widely adopted technique is the boxplot filter (BP), which uses the interquartile range (IQR) to set a threshold: $T_{\mathrm{BP}}(\omega) = Q_3(\omega) + 1.5 \times (Q_3(\omega) - Q_1(\omega))$, where $Q_1(\omega)$ and $Q_3(\omega)$ are the first and third quartiles of the residuals ω [Han et al., 2022; Lima et al., 2022].

When data distribution is approximately normal, thresholds can be defined based on Gaussian assumptions. Observations beyond three standard deviations from the mean are considered anomalies: $T_{GS}(Z_\omega) = 3, \quad \text{with} \quad Z_{\omega_i} = \frac{\omega_i - \mu(Z_\omega)}{\sigma(Z_\omega)}$, where μ and σ are the mean and standard deviation of the residuals.

The ratio threshold (RT) normalizes residuals by their maximum value and calculates a relative deviation: $\omega' = 1 - \frac{\omega}{\max(\omega)}$. An anomaly is flagged when $T_{\mathrm{RT}}(\omega) = \overline{\omega'} + \sigma_{\omega'}$, where $\overline{\omega'}$ and $\sigma_{\omega'}$ represent the mean and standard deviation of ω', respectively [Souza et al., 2024; Scharf and Demeure, 1991].

Extreme Value Theory (EVT) provides another approach by modeling the distribution of extreme deviations, emphasizing rare and significant anomalies [Aggarwal, 2016; Talagala et al., 2020].

2.4. Candidate Selection

When multiple candidate anomalies are detected in close temporal proximity, selecting the appropriate anomaly helps avoid redundant alarms [Lima et al., 2024; Salles et al., 2024]. Two common strategies for candidate selection are: (i) first detected anomaly (F), and (ii) highest deviation observed (H).

The F strategy selects the earliest detected anomaly, while the H strategy chooses the observation with the largest deviation from expected values [Lima et al., 2022, 2024]. The choice between these strategies affects performance metrics such as precision, recall, and F_1. Typically, F provides a better balance between precision and recall, while H increases precision at the cost of more false positives [Lima et al., 2024].

2.5. Related Work

Several studies, reviews, and surveys explore different aspects of anomaly detection (AD) [Darban et al., 2024; Wu and Keogh, 2023; Pang et al., 2021; Blázquez-García et al., 2021]. In parallel, the literature also focuses on benchmarking frameworks, time series datasets, and evaluation metrics [Paparrizos et al., 2022; Boniol et al., 2022; Wenig et al., 2022].

Talagala et al. [2020] propose a framework for online AD based on Extreme Value Theory (EVT), which dynamically adjusts thresholds in nonstationary environments. Lima et al. [2022] introduce FBIAD, a method that employs absolute deviations to assess time series behavior without squaring deviations. Comparative analyses by Moustati et al. [2024], Hodson [2022], and Gorard [2005] demonstrate that RMSE is suitable for normal distributions, while MAE provides robustness across diverse data types. However, despite these contributions, few studies systematically evaluate the effect of tuning deviation measures, filter criteria, and candidate selection strategies on AD performance.

3. Benchmark Method

An adaptable modular framework is employed to benchmark anomaly detection (AD) methods. The framework focuses on three critical components: Deviation Measure Criterion, Filter Criterion, and Candidate Selection Criterion, as illustrated in Figure 1. By incorporating interfaces that abstract these elements, the framework enables seamless integration and exploration of different detection criteria.

During execution, specialized techniques are selected and incorporated into the detection process. Although the base interface invocation remains unchanged, different behaviors are obtained by assigning specific implementations for deviation, filtering, and candidate selection, as detailed in Algorithm 1.

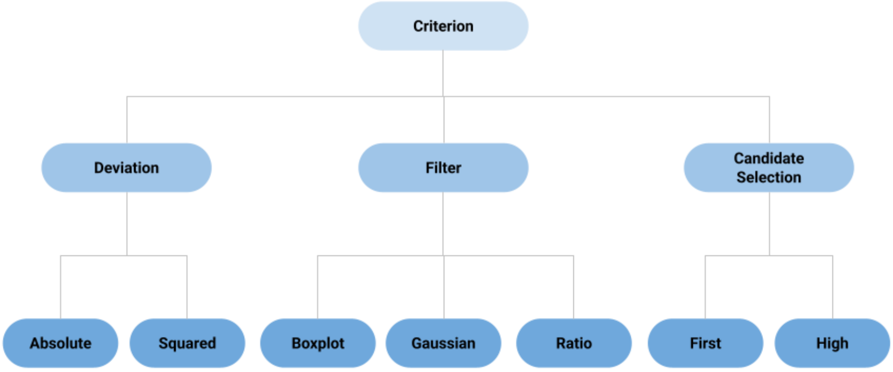

Figure 1. Deviation Measure Criterion, Filter Criterion, and Candidate Selection Criterion for AD

In Algorithm 1, the benchmark function receives a time series Y, an AD method adm, and interfaces for the Deviation Measure Criterion $devm$, Filter Criterion $filter$, and Candidate Selection Criterion $candidate$.

Algorithm 1 Benchmark Algorithm

```
 1: function benchmark($Y$, $adm$, $devm$, $filter$, $candidate$)
 2:     $model \leftarrow \text{setup}(adm, devm, filter, candidate)$
 3:     $model \leftarrow \text{fit}(model, Y)$                    ▷ Fit the model with data
 4:     $anomalies \leftarrow \text{detect}(model, Y)$             ▷ Detect anomalies
 5:     return $anomalies$
 6: end function
 7: function detect($model, Y$)
 8:     $\omega \leftarrow residuals(model, Y)$
 9:     $\dot{\omega} \leftarrow devm(model, \omega)$
10:     $canom \leftarrow filter(model, \dot{\omega})$
11:     $anom \leftarrow candidate(model, canom)$
12:     return $anom$
13: end function
```

The main steps are described in lines 2–4. Line 2 initializes the model using adm and the selected detection criteria. Line 3 fits the model to the time series Y. Line 4 applies the detection process based on the fitted model and selected criteria.

The detection process, detailed in lines 8–11, begins by computing the residuals ω from the model and the time series. It then computes deviations $\dot{\omega}$ based on the Deviation Measure Criterion. Line 10 identifies candidate anomalies $canom$ using the Filter Criterion. Finally, line 11 confirms anomalies $anom$ using the Candidate Selection Criterion.

The Deviation Measure Criterion ($devm$) quantifies the magnitude of deviations in the time series. The framework enables the selection of different computation strategies, as discussed in Section 2.2, including Absolute Deviation (ABS) and Squared Deviation (SQD). These measures provide insight into the residual variance and influence the detection outcome.

The Filter Criterion $(filter)$ defines thresholds that determine whether deviations are large enough to classify observations as anomalies. Several strategies are available, as detailed in Section 2.3, including Boxplot (BP), Gaussian Distribution (GD), and Ratio Threshold (RT). The selected filter affects the balance between false positives and false negatives.

When multiple nearby candidate anomalies are detected, the Candidate Selection Criterion $(candidate)$ determines which observations are confirmed as anomalies. The available strategies, described in Section 2.4, include First Anomaly (F) and Highest Deviation (H). Choosing an appropriate strategy influences the final precision and recall of the detection process.

4. Experimental Evaluation

This section analyzes the experimental results to assess the impact of varying the three main detection criteria: (i) Deviation Measure Criterion, (ii) Filter Criterion, and (iii) Candidate Selection Criterion. These criteria are evaluated using different anomaly detection methods across various types of time series, addressing the research questions proposed in Section 1.

4.1. Experimental Setup

Three datasets with distinct characteristics are selected to evaluate the proposed approach: Yahoo Labs, Numenta Anomaly Benchmark (NAB), and GECCO 2018 Challenge datasets.

The Yahoo Labs dataset contains hourly observations with manually labeled anomalies, comprising both real and synthetic time series. It includes five real time series (A1Benchmark set) related to Yahoo's service traffic and fifteen synthetic time series (A2Benchmark, A3Benchmark, and A4Benchmark sets) [Ogasawara et al., 2024].

The NAB dataset includes labeled time series for anomaly detection tasks, with both synthetic and real-world data. It is used to monitor cloud services such as CPU utilization, network traffic, and disk reads [Lima et al., 2022; Souza et al., 2024].

The GECCO 2018 Challenge dataset comprises 1,500 hourly observations related to water quality monitoring, with 72 labeled events across nine variables, including temperature, pH, redox potential, electrical conductivity, and turbidity [Lima et al., 2022; Souza et al., 2024].

Each anomaly detection method is evaluated under all combinations of:

- Deviation Measure: Absolute Deviation (ABS) and Squared Deviation (SQD),
- Filter Criterion: Boxplot (BP), Gaussian Distribution (GD), and Ratio Threshold (RT),
- Candidate Selection Criterion: First (F) and Highest Deviation (H).

The employed methods include statistical approaches (REMD [Souza et al., 2024], EMD, FBIAD [Lima et al., 2022], and ARIMA) and machine learning models (LSTM, ELM, Conv1D, and SVM) [Ogasawara et al., 2025]. These methods are chosen based on their relevance to AD tasks and availability in the Harbinger framework [Ogasawara et al., 2024].

Experiments are conducted in a computational environment with a Xeon processor (12 cores), 256 GB of RAM, and Ubuntu 22.04 LTS. The hyperparameters are set as follows: EMD and REMD use five trials with a noise parameter of 0.1. FBIAD employs a sliding window of 30 observations. LSTM and Conv1D are trained for 10,000 epochs. ELM uses the Purelin activation function, and SVM adopts the radial basis function (RBF) kernel.

Performance is evaluated using classification metrics derived from the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [Salles et al., 2024]. The derived measures are: (i) Precision, the proportion of correctly identified anomalies among all detections, (ii) Recall, the proportion of actual anomalies that are detected, and (iii) F_1 (the harmonic mean of precision and recall) [Han et al., 2022].

Following this setup, Section 4.2 presents an overall performance analysis. Then, results are discussed in detail for each dataset: Yahoo (Section 4.3), GECCO (Section 4.4), and Numenta (Section 4.5). A deep-dive study for a specific Yahoo series is presented in Section 4.6.

4.2. Overall Detection Analysis

This subsection analyzes the average performance of Precision, Recall, and F_1 across the three studied datasets. Table 1 presents results grouped by the three main detection criteria: Deviation Measures Criterion, Filter Criterion, and Candidate Selection Criterion.

As shown in Table 1, selecting ABS as the Deviation Measures Criterion and F as the Candidate Selection Criterion yields the best overall results across Precision, Recall, and F_1. Regarding the Filter Criterion, results vary according to the goal: RT yields better Precision, BP yields better Recall, and GD yields better F_1. The values in Table 1 represent the average of the metric values computed for all configurations that use each respective criterion.

Table 1. Overall Detection Analysis

	Deviation Measures		Filter Criterion			Candidate Criterion	
Measures	**ABS**	**SQD**	**BP**	**GD**	**RT**	**F**	**H**
Precision	**0.219**	0.217	0.070	0.239	**0.345**	**0.225**	0.210
Recall	**0.315**	0.290	**0.376**	0.331	0.200	**0.323**	0.281
F1	**0.201**	0.178	0.089	**0.246**	0.237	**0.199**	0.179

4.3. Detailed Analysis of Yahoo Dataset

This subsection presents a detailed analysis of the experimental results for the Yahoo dataset. Table 2 summarizes the performance of different detection methods combined with variations of the Deviation Measure Criterion, Filter Criterion, and Candidate Selection Criterion.

Table 2. Summary Results for Yahoo Dataset

Method	Criteria	Prec.	Recall	F1	Method	Criteria	Prec.	Recall	F1
REMD	ABS-RT-F	0.714	0.482	0.643	**LSTM**	ABS-GD-F	0.456	0.578	0.457
	ABS-RT-H	0.708	0.482	0.691		SQD-BP-F	0.039	0.657	0.070
	ABS-BP-H	0.133	0.629	0.245					
EMD	SQD-RT-H	0.205	0.092	0.103	**ELM**	ABS-GD-F	0.414	0.581	0.447
	SQD-RT-F	0.155	0.130	0.131		SQD-BP-F	0.039	0.653	0.069
	ABS-RT-H	0.138	0.170	0.112					
FBIAD	ABS-GD-H	0.582	0.523	0.597	**Conv1D**	SQD-RT-H	**0.785**	0.474	0.524
	ABS-RT-H	0.378	0.626	0.345		SQD-BP-F	0.038	0.599	0.066
ARIMA	ABS-RT-F	0.632	0.422	0.685	**SVM**	ABS-RT-F	0.630	0.432	0.470
	ABS-RT-H	0.526	0.361	**0.694**		SQD-GD-F	0.615	0.636	0.576
	ABS-BP-F	0.141	**0.758**	0.235		ABS-BP-F	0.111	0.701	0.172

REMD achieves a strong balance between $\mathrm{Precision}$ and Recall when combined with the ABS deviation measure, the RT filter criterion, and either F or H for candidate selection. Changing the filter from BP to RT increases $\mathrm{Precision}$ by approximately 81%, although Recall decreases by about 23%.

ARIMA, when combined with ABS and the RT filter, presents high F_1 scores. The configuration $\mathrm{ABS\text{-}RT\text{-}F}$ offers a balanced performance, while Conv1D achieves the highest $\mathrm{Precision}$ across all methods when paired with $\mathrm{SQD\text{-}RT\text{-}H}$. However, performance is highly sensitive to the selected detection criteria, reinforcing the importance of fine-tuning.

These results demonstrate that adapting detection criteria (deviation measure, filter, and candidate selection) significantly impacts performance on the Yahoo dataset, confirming the benefits of the proposed methodology.

4.4. Detailed Analysis of GECCO Dataset

This subsection presents a detailed analysis of the experimental results for the GECCO dataset. Table 3 summarizes the performance of different detection methods combined with variations of the Deviation Measure Criterion, Filter Criterion, and Candidate Selection Criterion.

Among the evaluated configurations, EMD combined with $\mathrm{SQD\text{-}RT\text{-}H}$ achieves the highest $\mathrm{Precision}$ (0.306). However, this result comes with a notably low Recall (0.014), highlighting a trade-off in which improved $\mathrm{Precision}$ reduces anomaly sensitivity.

FBIAD combined with $\mathrm{ABS\text{-}GD\text{-}F}$ achieves the highest F_1 score (0.160), balancing $\mathrm{Precision}$ and Recall. This result suggests that, for the GECCO dataset, jointly tuning the filter threshold and candidate selection improves detection performance. Conv1D and ELM present reasonable $\mathrm{Precision}$ but suffer from low Recall, indicating a tendency to miss anomalies.

These findings underscore the need to balance detection objectives according to application context. In environmental monitoring, achieving a trade-off between sensitivity and false alarms is critical for ensuring the relevance and reliability of detections.

Table 3. Summary Results for GECCO Dataset

Method	Criteria	Prec.	Recall	F1	Method	Criteria	Prec.	Recall	F1
REMD	SQD-BP-F	0.072	0.085	0.076	**LSTM**	SQD-BP-F	0.073	0.073	0.071
	ABS-GD-H	0.160	0.032	0.064		SQD-RT-H	0.207	0.017	0.029
	SQD-RT-H	0.127	0.012	0.101					
EMD	ABS-RT-F	0.125	**0.102**	0.084	**ELM**	SQD-BP-F	0.070	0.090	0.076
	SQD-RT-H	**0.306**	0.014	0.026		ABS-GD-H	0.178	0.043	0.066
FBIAD	SQD-BP-F	0.079	0.076	0.070	**Conv1D**	SQD-BP-H	0.060	0.090	0.067
	ABS-GD-F	0.269	0.054	**0.160**		ABS-BP-F	0.138	0.076	0.090
	ABS-GD-H	0.276	0.054	0.134		SQD-RT-H	0.206	0.022	0.032
ARIMA	SQD-BP-F	0.078	0.093	0.081	**SVM**	SQD-BP-F	0.051	0.079	0.060
	ABS-BP-F	0.110	0.074	0.089		ABS-BP-F	0.080	0.063	0.062
	SQD-RT-H	0.170	0.017	0.080		ABS-GD-H	0.152	0.035	0.049

4.5. Detailed Analysis of Numenta Dataset

This subsection presents a detailed analysis of the experimental results for the Numenta dataset. Table 4 summarizes the detection performance for different combinations of the Deviation Measure Criterion, Filter Criterion, and Candidate Selection Criterion.

Table 4. Summary Results for Numenta Dataset

Method	Criteria	Prec.	Recall	F1	Method	Criteria	Prec.	Recall	F1
REMD	ABS-GD-F	0.015	0.500	0.059	**LSTM**	SQD-RT-F	0.118	0.066	0.083
	SQD-BP-H	0.003	**1.000**	0.006		ABS-BP-F	0.015	0.489	0.028
EMD	ABS-BP-H	0.000	0.000	0.000	**ELM**	ABS-RT-H	0.438	0.198	0.263
						SQD-RT-H	**0.500**	0.182	0.263
						ABS-BP-H	0.023	0.604	0.042
FBIAD	ABS-RT-H	0.006	0.500	0.023	**Conv1D**	SQD-RT-H	0.389	0.134	0.196
	ABS-RT-F	0.006	0.500	0.023		ABS-BP-H	0.010	0.491	0.020
ARIMA	ABS-RT-H	0.300	0.108	**0.522**	**SVM**	ABS-RT-H	0.412	0.172	0.233
	SQD-BP-H	0.021	0.675	0.057		SQD-RT-H	0.471	0.157	0.231
						SQD-BP-H	0.011	0.607	0.022

Among the evaluated configurations, REMD combined with $\mathrm{SQD\text{-}BP\text{-}H}$ achieves perfect Recall (1.000), detecting all anomalies in the dataset. However, this result is accompanied by extremely low $\mathrm{Precision}$ (0.003), resulting in many false positives.

ARIMA combined with ABS and $\mathrm{RT\text{-}H}$ yields the highest F_1 score (0.522), indicating a balanced trade-off between $\mathrm{Precision}$ and Recall.

The ELM method, with $\mathrm{SQD\text{-}RT\text{-}H}$, achieves the highest $\mathrm{Precision}$ (0.500) while maintaining moderate Recall (0.182). This configuration demonstrates a good balance between minimizing false positives and maintaining detection sensitivity. Similarly, SVM with both $\mathrm{ABS\text{-}RT\text{-}H}$ and $\mathrm{SQD\text{-}RT\text{-}H}$ presents competitive performance, confirming the benefits of ratio-based filters and highest-deviation selection strategies.

These results reinforce the importance of aligning detection criteria with application requirements. In scenarios where detecting all anomalies is critical, such as security monitoring, high Recall is prioritized despite more false positives. Conversely, applications requiring precision and reliability, such as decision support systems, benefit from configurations that minimize false alarms.

4.6. Deep-dive Analysis of Yahoo Series

This subsection provides a detailed analysis of anomaly detection results for the Yahoo dataset, focusing on the time series A3Benchmark-TS43. Figure 2 illustrates the detection performance. Dark blue points represent ground truth anomalies, green points indicate true positives (TP), red points correspond to false positives (FP), and blue points denote false negatives (FN).

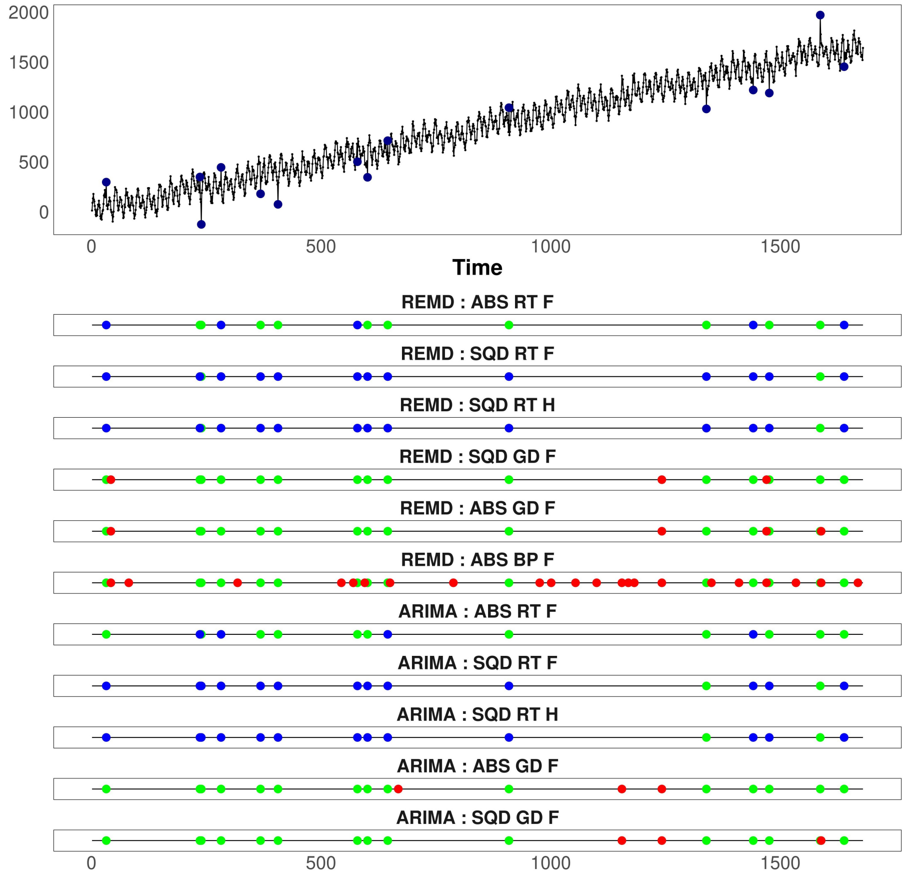

Figure 2. Best Combinations by Precision, Recall, and F_1 Score for A3Benchmark-TS43.

The REMD method, combined with ABS deviation and F candidate selection, presents strong performance, balancing Precision and Recall. Among the filter criteria, GD stands out, especially when paired with ABS and F selection. In contrast, configurations using SQD exhibit lower performance in both Precision and Recall, reinforcing the importance of deviation measure choice, particularly when data distributions deviate from normality.

The combination of GD filtering and ABS deviation proves effective for this series, supporting the use of Gaussian-based thresholds when data exhibit near-normal behavior. Regarding candidate selection, the F strategy, selecting the first detected anomaly, yields a better balance between identifying anomalies and avoiding redundant detections. On the other hand, the H strategy, selecting the point with the highest deviation, tends to increase false positives, lowering Precision. While it captures extreme deviations, it may overemphasize minor fluctuations near actual anomalies.

ARIMA, configured with ABS and F, also achieves robust performance, combining high Precision and Recall. This result confirms that tuning detection criteria can yield performance levels comparable to or even better than more complex models. Notably, when SQD is combined with GD and F, perfect Recall is achieved, though at the cost of very low Precision due to excessive false positives. This trade-off highlights the relevance of prioritizing Recall in contexts where missing anomalies is unacceptable.

The BP filter, when applied with REMD, maintains high Recall but significantly reduces Precision. This suggests that although BP is sensitive to outliers, it can trigger excessive false alarms in asymmetric data distributions. This deep-dive analysis illustrates the importance of carefully selecting deviation measures, filters, and candidate selection strategies for each specific use case. Visualizations such as Figure 2 provide valuable support for interpreting model behavior and guiding fine-tuning decisions before deployment.

5. Conclusion

This paper examines the effect of fine-tuning detection criteria, particularly deviation measures, filter thresholds, and candidate selection strategies, on anomaly detection performance in time series. Experimental results demonstrate that adjusting these criteria often improves detection accuracy, sometimes with greater impact than changing the detection method itself. By systematically evaluating combinations across multiple datasets and methods, the study highlights the sensitivity of detection outcomes to criterion selection.

The experimental analysis addresses the research questions posed in Section 1. Regarding Q1, changing the deviation measure influences detection results across datasets, with absolute deviation (ABS) often outperforming squared deviation (SQD) in balancing Precision and Recall. For Q2, adjusting the filter criterion significantly affects the trade-off between false positives and false negatives. The ratio-based threshold (RT) improves Precision in multiple scenarios. Concerning Q3, selecting the candidate anomaly within groups, either the first occurrence (F) or the point with the highest deviation (H), alters detection performance, with the F strategy frequently providing a better balance between metrics.

The results confirm that fine-tuning detection criteria is a key strategy for improving anomaly detection in real-world applications, especially when domain-specific characteristics influence time series behavior. Adjusting deviation measures, filter thresholds, and candidate selection strategies allows methods to adapt to varying distributions, noise levels, and anomaly types without altering the core detection algorithms.

Future work includes expanding the evaluation to additional datasets and anomaly types, exploring automated approaches to selecting detection criteria based on data characteristics, and integrating adaptive fine-tuning mechanisms into anomaly detection frameworks. This direction aims to improve robustness and reduce manual intervention in diverse operational environments.

References

Aggarwal, C. C. (2016). *Outlier Analysis*. Springer International Publishing.

Ahmad, S., Lavin, A., Purdy, S., and Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134 – 147.

Ahmed, M., Mahmood, A. N., and Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278 – 288.

Ariyaluran Habeeb, R. A., Nasaruddin, F., Gani, A., Targio Hashem, I. A., Ahmed, E., and Imran, M. (2019). Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management*, 45:289 – 307.

Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. (2021). A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3).

Boniol, P., Paparrizos, J., Kang, Y., Palpanas, T., Tsay, R. S., Elmore, A. J., and Franklin, M. J. (2022). Theseus: Navigating the Labyrinth of Time-Series Anomaly Detection. *Proceedings of the VLDB Endowment*, 15(12):3702 – 3705.

Cauteruccio, F., Cinelli, L., Corradini, E., Terracina, G., Ursino, D., Virgili, L., Savaglio, C., Liotta, A., and Fortino, G. (2021). A framework for anomaly detection and classification in Multiple IoT scenarios. *Future Generation Computer Systems*, 114:322 – 335.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3).

Cook, A. A., Misirli, G., and Fan, Z. (2020). Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal*, 7(7):6481 – 6494.

Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., and Salehi, M. (2024). Deep Learning for Time Series Anomaly Detection: A Survey. *ACM Computing Surveys*, 57(1):15:1–15:42.

Dixit, P., Bhattacharya, P., Tanwar, S., and Gupta, R. (2022). Anomaly detection in autonomous electric vehicles using AI techniques: A comprehensive survey. *Expert Systems*, 39(5).

Gorard, S. (2005). Revisiting A 90-year-old debate: The advantages of the mean deviation. *British Journal of Educational Studies*, 53(4):417 – 430.

Han, J., Pei, J., and Tong, H. (2022). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Cambridge, MA, 4th edition edition.

Hasani, Z. (2017). Robust anomaly detection algorithms for real-time big data: Comparison of algorithms. In *2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings*.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14):5481 – 5487.

Ibidunmoye, O., Hernández-Rodriguez, F., and Elmroth, E. (2015). Performance anomaly detection and bottleneck identification. *ACM Computing Surveys*, 48(1).

Li, J., Izakian, H., Pedrycz, W., and Jamal, I. (2021). Clustering-based anomaly detection in multivariate time series data. *Applied Soft Computing*, 100.

Lima, J., Salles, R., Porto, F., Coutinho, R., Alpis, P., Escobar, L., Pacitti, E., and Ogasawara, E. (2022). Forward and Backward Inertial Anomaly Detector: A Novel Time Series Event Detection Method. In *2022 International Joint Conference on Neural Networks (IJCNN)*, volume 2022-July, pages 1–8.

Lima, J., Tavares, L. G., Pacitti, E., Ferreira, J. E., Santos, I., Siqueira, I. G., Carvalho, D., Porto, F., Coutinho, R., and Ogasawara, E. (2024). Online Event Detection in

Streaming Time Series: Novel Metrics and Practical Insights. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8.

Moustati, I., Gherabi, N., and Saadi, M. (2024). Time-Series Forecasting Models for Smart Meters Data: An Empirical Comparison and Analysis. *Journal Europeen des Systemes Automatises*, 57(5):1419 – 1427.

Ogasawara, E., Castro, A., Mello, A., Paixão, E., Fraga, F., Lima, J., Souza, J., Baroni, L., Tavares, L., Borges, H., Salles, R., Carvalho, D., Bezerra, E., Coutinho, R., Pacitti, E., and Porto, F. (2024). harbinger: A Unified Time Series Event Detection Framework.

Ogasawara, E., Salles, R., Porto, F., and Pacitti, E. (2025). *Event Detection in Time Series*. Springer, 2025 edition.

Pang, G., Shen, C., Cao, L., and Van Den Hengel, A. (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2).

Paparrizos, J., Kang, Y., Boniol, P., Tsay, R. S., Palpanas, T., and Franklin, M. J. (2022). TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection. *Proceedings of the VLDB Endowment*, 15:1697 – 1711.

Ren, H., Xu, B., Wang, Y., Yi, C., Huang, C., Kou, X., Xing, T., Yang, M., Tong, J., and Zhang, Q. (2019). Time-series anomaly detection service at Microsoft. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3009 – 3017.

Salles, R., Escobar, L., Baroni, L., Zorrilla, R., Ziviani, A., Kreischer, V., Delicato, F., Pires, P., Maia, L., Coutinho, R., Assis, L., and Ogasawara, E. (2020). Um framework para integração e análise de métodos de detecção de eventos em séries temporais. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*. SBC.

Salles, R., Lima, J., Reis, M., Coutinho, R., Pacitti, E., Masseglia, F., Akbarinia, R., Chen, C., Garibaldi, J., Porto, F., and Ogasawara, E. (2024). SoftED: Metrics for soft evaluation of time series event detection. *Computers and Industrial Engineering*, 198.

Scharf, L. L. and Demeure, C. (1991). *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Publishing Company.

Souza, J., Pãixao, E., Fraga, F., Baroni, L., Alves, R. F. S., Belloze, K., Dos Santos, J., Bezerra, E., Porto, F., and Ogasawara, E. (2024). REMD: A Novel Hybrid Anomaly Detection Method Based on EMD and ARIMA. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8.

Talagala, P. D., Hyndman, R. J., Smith-Miles, K., Kandanaarachchi, S., and Muñoz, M. A. (2020). Anomaly Detection in Streaming Nonstationary Temporal Data. *Journal of Computational and Graphical Statistics*, 29(1):13 – 27.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167.

Wenig, P., Schmidl, S., and Papenbrock, T. (2022). TimeEval: A Benchmarking Toolkit for Time Series Anomaly Detection Algorithms. *Proceedings of the VLDB Endowment*, 15(12):3678 – 3681.

Wu, R. and Keogh, E. J. (2023). Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2421 – 2429.

Zhang, M., Guo, J., Li, X., and Jin, R. (2020). Data-driven anomaly detection approach for time-series streaming data. *Sensors (Switzerland)*, 20(19):1 – 17.