# Advances in Time-Series Anomaly Detection:

Algorithms, Benchmarks, and Evaluation Measures

John Paparrizos The Ohio State University; Aristotle University paparrizos.1@osu.edu

Paul Boniol Inria, ENS, PSL University Paul.boniol@inria.fr

Qinghua Liu The Ohio State University liu.11085@osu.edu Themis Palpanas Université Paris Cité; IUF themis@mi.parisdescartes.fr













- Part 1: Introduction, Motivation and Foundations
- Part 2: Taxonomy of Anomaly Detection Methods
- Part 3: Evaluation Measures
- Part 4: Anomaly Detection Benchmarks
- Part 5: Automated Solutions for Anomaly Detection
- Part 6: Conclusion and Open Problems

# Part 1: Introduction, Motivation and Foundations

#### **Energy Production**



Edf.fr: tinyurl.com/yc7x5xje

#### Astrophysics



Virgo: https://www.virgo-gw.eu/

#### Medicine



tinyurl.com/39dx2us4

#### Volcanology



tinyurl.com/ybcttmfz

Energy Production	
Secondary circuit sensor measurements	

#### Astrophysics



Virgo: https://www.virgo-gw.eu/

#### Medicine



tinyurl.com/39dx2us4

#### Volcanology



tinyurl.com/ybcttmfz



#### Medicine



tinyurl.com/39dx2us4

#### Volcanology



tinyurl.com/ybcttmfz



#### Volcanology



tinyurl.com/ybcttmfz











*Large-scale time series database* 

#### **Energy Production**



Edf.fr: tinyurl.com/yc7x5xje

Large-scale time series database





Edf.fr: tinyurl.com/yc7x5xje

Large-scale time series database



Large-scale time series database



• Time series T (example : number of taxi passengers in New York City)



• Time series T (example : number of taxi passengers in New York City)





 Anomaly: rare point or sequence (of a given length) potentially non-desired







KDD 2025 | 03/08/2025 | 22



KDD 2025 | 03/08/2025 | 23

# Introduction: Outline

• Time series (example : number of taxi passengers in New York City)



# Introduction: Outline

• Time series (example : number of taxi passengers in New York City)



# Introduction: Outline

• Time series (example : number of taxi passengers in New York City)





















50






### Foundations: Type of anomalies



# Part 2: Taxonomy of Anomaly Detection Methods



**Time Series** 









KDD 2025 | 03/08/2025 | 44







[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. H VLDB Endow. 15, 9 (May 2022), 1779–1797.

ConInd[5] S-H-ESD [62] SH-ESD+[138] FAST-MCD[115] MA[18] EWMA [65] SARIMA [52] Kalman Filter [52] ANODE [96] AR [18] MGDD [126] **Statistics PCI**[157] ARMA [18] MedianMethod [10] pEWMA [25] Holt-Winter's [1] EWMA-STR [162] **DSPOT**[122] **ARIMA**[65] RePAD [76] AMD Segmentation [153] Holt's [65]

[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc.KDD 2025 | 03/08/2025 | 48VLDB Endow. 15, 9 (May 2022), 1779–1797.



[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. KDD 2025 | 03/08/2025 | 49 VLDB Endow. 15, 9 (May 2022), 1779–1797.



[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. KDD 2025 | 03/08/2025 | 50 VLDB Endow. 15, 9 (May 2022), 1779–1797.



[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc.KDD 2025 | 03/08/2025 | 51VLDB Endow. 15, 9 (May 2022), 1779–1797.



[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc.KDD 2025 | 03/08/2025 | 52VLDB Endow. 15, 9 (May 2022), 1779–1797.



[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc.KDD 2025 | 03/08/2025 | 53VLDB Endow. 15, 9 (May 2022), 1779–1797.

RobustPCA[101] Eros-SVMs [74] k-Means [151] XGBoosting [34] KNN[110]	SR [112] DWT-MLEAD [134] L-HMM [127] U-GMM-HMM [68]
NetworkSVM [160] MS-SVDD [149] sequenceMiner [23] AOSVM [48]	
RUSBoost [54] OC-KFD [114] PhaseSpace-SVM [85] NoveltySVR [86]	Signal Analysis SmartSifter [152] LaserDBN [100]
SLADE-TS [141]	Online DWT- VET [111] CLASSAL Stochastic
PCA[121] S-SVM[11]	MLEAD [133] GLA [84] Stochastic
Hybrid K-Means [140] Random Forest Regressor [165]	EM-HMM [105] Learning EDBN [107]
SLADE-MTS [142] PCC [121] Normalizing Flow [116]   Hybrid KNN [124] Hybrid KNN [124] EncDec-AD [88]	
HBOS[47] LSIN STOMP[164] DeepISTM[31] SSA[155] VAE C	A-based LAMP [166] FuzzyDNBC [136]
	DAE [117] TCN-AE [135] HMAD [49]
Series ZGraph [16] DeepNAP [72] LSTM-VAE [106] MAD-GAN [77] OmniAnomaly [125]	
GrammarViz[120] TwoFinger [90] CoalESN [99] Torsk [60]	
KnorrSeq2[102] Left STAMPi[156] STORN[123] Donut[150] Deep I	Learning PAD[33] DeepAnT[94] ConInd[5] <i>DeepAnT</i> [94] S-H-ESD[62]
TSBitmap [144] DADS [119] MSCRED [159]	TTM [14c] T-lementer [c4] LSTM-AD [89] FAST-MCD [115] SH-ESD+ [138]
HOT SAX [70] DissimilarityAlgo [6] RADM [40] SP. CNN [112]	VELC[158] MA[18] EWMA[65] SARIMA[52]
MoteESN[30]	$\frac{\partial GAN[8]}{AE[117]} = \frac{AE[17]}{Bagel[79]} = \frac{AE[17]}{AE[17]} = \frac{Bagel[79]}{AE[17]} = \frac{Bageel[79]}{AE[17]} = Bageel[79$
NorM[14] Data Mining MTAD CAT[161] Nument	ANODE [96] Kalman Filter [52]
Image-embe	AR [18]
BoehmerGraph [13] VALMOD [82] PST [128]	MGDD[126] Statistics PCI[157]
TARZAN[71] MERLIN [97] STAMP[156] MCOD [73] CBLOF [59] ARMA [18]   Isolation Forest [83] EIF [58] EIF [58] MedianMethod [10]	
NormA-SJ[15] ILOF [108] DAD [154] LOCI/aLOCI [103] Subsequence	e IF [83] Subsequence LOF [22] Holt-Winter's [1]
NormA-smpl[15] SurpriseEncoding[26] COPOD[80]	
IF-LOF[36] Outlier Detection ARIMA[65] DSPOT[122] RePAD[76]	
SCRIMP++ [163] Ensemble GI [43] Hybrid Isolation Forest [91] COF [130] BLOF [59]	GeckoFSM [118] DBStream [55] LOF [22] DILOF [95] AMD Segmentation [153] Holt's [65]

[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc.KDD 2025 | 03/08/2025 | 54VLDB Endow. 15, 9 (May 2022), 1779–1797.KDD 2025 | 03/08/2025 | 54

By inputs...

**TSAD** Methods



KDD 2025 | 03/08/2025 | 56



#### Anomaly Detection methods: A taxonomy By inputs... **TSAD** Methods **Supervised** Semi-supervised Unsupervised - Normal examples - Normal examples Training Training - Anomaly examples dataset dataset Time Series T Time Series T ightarrow Time Series T

**TSAD** Methods


























# Anomaly Detection methods: *A taxonomy* By methods...



# Anomaly Detection methods: *A taxonomy* By methods...



KDD 2025 | 03/08/2025 | 79

### Anomaly Detection methods: *A taxonomy*

By time...



# Anomaly Detection methods: *A taxonomy* By time...



# Anomaly Detection methods: *A taxonomy* By time...





















#### Matrix Profile [6] (MP)

Compute the distance to the nearest neighbor (using the MASS algorithm z-norm Euclidean distance computation) and use it as anomaly score





#### Matrix Profile [6] (MP)

Compute the distance to the nearest neighbor (using the MASS algorithm z-norm Euclidean distance computation) and use it as anomaly score





#### Matrix Profile [6] (MP)

Compute the distance to the nearest neighbor (using the MASS algorithm z-norm Euclidean distance computation) and use it as anomaly score





The matrix Profile is computed as follows:  $S_T = \left[ NN(T_{0,\ell}), NN(T_{1,\ell}), \dots, NN(T_{|T|-\ell,\ell}) \right]$ 

#### Matrix Profile [6] (MP)

Compute the distance to the nearest neighbor (using the MASS algorithm z-norm Euclidean distance computation) and use it as anomaly score





#### Matrix Profile [6] (MP)

Compute the distance to the nearest neighbor (using the MASS algorithm z-norm Euclidean distance computation) and use it as anomaly score





Matrix Profile [6] (MP)

Compute the distance to the nearest neighbor (using the MASS algorithm z-norm Euclidean distance computation) and use it as anomaly score





#### NormA [10]

Distance-based approach that summarize the time series into a weighted set of subsequences and use the distance to them as anomaly score



[10] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. The VLDB Journal 30, 6 (Nov 2021), 909–931.



NormA [10]

Distance-based approach that summarize the time series into a weighted set of subsequences and use the distance to them as anomaly score



[10] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. The VLDB Journal 30, 6 (Nov 2021), 909–931.



NormA [10]

Distance-based approach that summarize the time series into a weighted set of subsequences and use the distance to them as anomaly score



[10] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. The VLDB Journal 30, 6 (Nov 2021), 909–931.



#### NormA [10]

Distance-based approach that summarize the time series into a weighted set of subsequences and use the distance to them as anomaly score



[10] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. The VLDB Journal 30, 6 (Nov 2021), 909–931.

KDD 2025 | 03/08/2025 | 100



#### NormA [10]

Distance-based approach that summarize the time series into a weighted set of subsequences and use the distance to them as anomaly score



[25] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J. Franklin. 2021. SAND: streaming subsequence anomaly detection. Proc. VLDB Endow. 14, 10 (June 2021), 1717–1729.





















[11] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422 KDD 2025 | 03/08/2025 | 111







#### Isolation Forest [11]

Density-based approach that split the space randomly and using the depth of the trees to identify anomalies








Each **node** is an ensemble of similar subsequences.

#### Series2Graph [13]

Density-based approach that convert the time series into a graph and detect unusual trajectories





Each **node** is an ensemble of similar subsequences.

Each **edge** is associated to a weight *w* that corresponds to the number of times a subsequence move from one node to another.

#### Series2Graph [13]

Density-based approach that convert the time series into a graph and detect unusual trajectories



Univariate

subsequence







[26] Schneider, J., Wenig, P. & Papenbrock, T. Distributed detection of sequential anomalies in univariate time series. The VLDB Journal **30**, 579–602 (2021).









[13] Paul Boniol and Themis Palpanas. 2020. Series2Graph: graph-based subsequence anomaly detection for time series. Proc. VLDB Endow. 13, 12 (August 2020), 1821–1834



subsequence



[13] Paul Boniol and Themis Palpanas. 2020. Series2Graph: graph-based subsequence anomaly detection for time series. Proc. VLDB Endow. 13, 12 (August 2020), 1821–1834

KDD 2025 | 03/08/2025 | 121



GraphAn [28] An interactive tool to dive into the computation steps of Series2Graph :

[28] Paul Boniol, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2020. GraphAn: graph-based subsequence anomaly detection. Proc. VLDB Endow. 13, 12 (August 2020), 2941–2944.

















[15] Pankaj Malhotra, Lovekesh Vig, Gautam Shro, and Puneet Agarwal. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. (2015).

KDD 2025 | 03/08/2025 | 130



[16] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed. 2019. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. IEEE Access 7 (2019), 1991–2005.













[16] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed. 2019. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. IEEE Access 7 (2019), 1991–2005.

















#### AutoEncoders [17] (AE)

Neural Network composed of an encoder (that reduce the dimensionality) and decoder that reconstruct the time series. The objective is to minimize the reconstruction error.



[17] Mayu Sakurada and Takehisa Yairi. 2014. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis (Gold Coast, Australia QLD, Australia) (MLSDA'14).

#### Part 3: Evaluation Measures

#### Evaluation measures: A general overview



#### Evaluation measures: A general overview



# Evaluation measures: Threshold-based

Threshold-based Evaluation Measures:



# Evaluation measures: Threshold-based

Threshold-based Evaluation Measures:



# Evaluation measures: Threshold-based

Threshold-based Evaluation Measures:


Threshold-based Evaluation Measures:



Threshold-based Evaluation Measures:



Threshold-based Evaluation Measures:



Threshold-based Evaluation Measures:

- Precision:  $\frac{TP}{TP+FP}$ - Recall (true positive rate):  $\frac{TP}{TP+FN}$ - False positive rate:  $\frac{FP}{FP+TN}$ - E-score:  $\frac{(1+\beta^2)*Precision}{FP+TN}$ 

--score: 
$$\frac{\beta^2 * Precision + Recall}{\beta^2 * Precision + Recall}$$



How do we set the threshold?



How do we set the threshold?



How do we set the threshold?























[21] Jesse Davis and Mark Goadrich. 2006. The Relationship between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06).

Labeling can be an issue for time series [22]:



Labeling can be an issue for time series [22]:



Labeling can be an issue for time series [22]:

 Misalignment can lead to significant changes of accuracy values.



Labeling can be an issue for time series [22]:

- Misalignment can lead to significant changes of accuracy values.
- This is a real issue because of:
  - Different Labeling strategies between domains and applications
  - Methods that produce misaligned anomaly scores.



#### **Existing solutions:**

- Range Precision and Recall [23]:

- 
$$Recall_T(R, P) = \frac{\sum_{i=1}^{N_T} Recall_T(R_i, P)}{N_T}$$

- $Recall_T(R_i, P) = \alpha * ExistenceR(R_i, P) + (1 \alpha) * OverlappingR(R_i, P)$
- $Precision_T(R, P) = \frac{\sum_{i=1}^{N_p} Precision_T(R, P_i)}{N_p}$
- $Precision_T(R, P_i) = CardinalityFactor(P_i, R) * \sum_{j=1}^{N_T} w(P_i, P_i \cap R_j, \delta)$
- Functions  $w(), \delta()$  are tunable functions to represent the overlap size and position respectively.



Anomaly Score

**Anomaly Score** 

(a) Lag impact on accuracy measures





(a) Lag impact on accuracy measures

[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.

(b) Noise impact on the accuracy measures



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.

(c) Normal – abnormal ratio impact on accuracy measures



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.

(c) Normal – abnormal ratio impact on accuracy measures



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.



#### Summary:

- Volume Under the Surface [22] (VUS):
- Modify the labels with buffer regions at the beginning and at the end of an anomaly
- We vary the buffer size (as well as the threshold) and we obtain a surface
- We use the volume under the surface (VUS) as accuracy





#### How is it computed?

$$AUC\text{-}ROC = \frac{1}{2} \sum_{k=1}^{N} \Delta_{TPR}^{k} * \Delta_{FPR}^{k}$$
  
with: 
$$\begin{cases} \Delta_{FPR}^{k} &= FPR(Th_{k}) - FPR(Th_{k-1}) \\ \Delta_{TPR}^{k} &= TPR(Th_{k-1}) + TPR(Th_{k}) \end{cases}$$

$$AUC-PR = \frac{1}{2} \sum_{k=1}^{N} \Delta_{Precision}^{k} * \Delta_{Recall}^{k}$$
  
with: 
$$\begin{cases} \Delta_{Recall}^{k} = Recall(Th_{k}) - Recall(Th_{k-1}) \\ \Delta_{Precision}^{k} = Precision(Th_{k-1}) + Precision(Th_{k}) \end{cases}$$

#### How is it computed?

$$\begin{split} AUC\text{-}ROC &= \frac{1}{2} \sum_{k=1}^{N} \Delta_{TPR}^{k} * \Delta_{FPR}^{k} \\ \text{with:} & \begin{cases} \Delta_{FPR}^{k} &= FPR(Th_{k}) - FPR(Th_{k-1}) \\ \Delta_{TPR}^{k} &= TPR(Th_{k-1}) + TPR(Th_{k}) \end{cases} \end{split}$$

$$AUC-PR = \frac{1}{2} \sum_{k=1}^{N} \Delta_{Precision}^{k} * \Delta_{Recall}^{k}$$
  
with: 
$$\begin{cases} \Delta_{Recall}^{k} = Recall(Th_{k}) - Recall(Th_{k-1}) \\ \Delta_{Precision}^{k} = Precision(Th_{k-1}) + Precision(Th_{k}) \end{cases}$$

$$\begin{split} &VUS\text{-}ROC = \frac{1}{4} \sum_{w=1}^{L} \sum_{k=1}^{N} \Delta^{(k,w)} * \Delta^{w} \\ &\text{with:} \begin{cases} \Delta^{(k,w)} &= \Delta^{k}_{TPR_{\ell_{w}}} * \Delta^{k}_{FPR_{\ell_{w}}} + \Delta^{k}_{TPR_{\ell_{w-1}}} * \Delta^{k}_{FPR_{\ell_{w-1}}} \\ \Delta^{k}_{FPR_{\ell_{w}}} &= FPR_{\ell_{w}}(Th_{k}) - FPR_{\ell_{w}}(Th_{k-1}) \\ \Delta^{k}_{TPR_{\ell_{w}}} &= TPR_{\ell_{w}}(Th_{k-1}) + TPR_{\ell_{w}}(Th_{k}) \\ \Delta^{w} &= |\ell_{w} - \ell_{w-1}| \end{cases} \\ &VUS\text{-}PR = \frac{1}{4} \sum_{w=1}^{L} \sum_{k=1}^{N} \Delta^{(k,w)} * \Delta^{w} \\ &\text{with:} \begin{cases} \Delta^{(k,w)} &= \Delta^{k}_{Pr_{\ell_{w}}} * \Delta^{k}_{Re_{\ell_{w}}} + \Delta^{k}_{Pr_{\ell_{w-1}}} * \Delta^{k}_{Re_{\ell_{w-1}}} \\ \Delta^{k}_{Re_{\ell_{w}}} &= Recall_{\ell_{w}}(Th_{k}) - Recall_{\ell_{w}}(Th_{k-1}) \\ \Delta^{w} &= |\ell_{w} - \ell_{w-1}| \end{cases} \end{split}$$

#### How is it computed?



$$\begin{split} & VUS\text{-}ROC = \frac{1}{4} \sum_{w=1}^{L} \sum_{k=1}^{N} \Delta^{(k,w)} * \Delta^{w} \\ & \text{with:} \begin{cases} \Delta^{(k,w)} &= \Delta^{k}_{TPR_{\ell_{w}}} * \Delta^{k}_{FPR_{\ell_{w}}} + \Delta^{k}_{TPR_{\ell_{w-1}}} * \Delta^{k}_{FPR_{\ell_{w-1}}} \\ \Delta^{k}_{FPR_{\ell_{w}}} &= FPR_{\ell_{w}}(Th_{k}) - FPR_{\ell_{w}}(Th_{k-1}) \\ \Delta^{k}_{TPR_{\ell_{w}}} &= TPR_{\ell_{w}}(Th_{k-1}) + TPR_{\ell_{w}}(Th_{k}) \\ \Delta^{w} &= |\ell_{w} - \ell_{w-1}| \end{cases} \\ & VUS\text{-}PR = \frac{1}{4} \sum_{w=1}^{L} \sum_{k=1}^{N} \Delta^{(k,w)} * \Delta^{w} \\ & \text{with:} \begin{cases} \Delta^{(k,w)} &= \Delta^{k}_{Pr_{\ell_{w}}} * \Delta^{k}_{Re_{\ell_{w}}} + \Delta^{k}_{Pr_{\ell_{w-1}}} * \Delta^{k}_{Re_{\ell_{w-1}}} \\ \Delta^{k}_{Re_{\ell_{w}}} &= Recall_{\ell_{w}}(Th_{k}) - Recall_{\ell_{w}}(Th_{k-1}) \\ \Delta^{k}_{Pr_{\ell_{w}}} &= Precision_{\ell_{w}}(Th_{k-1}) + Precision_{\ell_{w}}(Th_{k}) \\ \Delta^{w} &= |\ell_{w} - \ell_{w-1}| \end{cases} \end{split}$$

[22] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (2022), 2774–2787.

#### How is it computed?

Time Complexity: O(NT)

With:

- *T*: the time series length
- N: the number of thresholds

with:  $\begin{cases} \Delta_{Recall}^{k} = Recall(Th_{k}) - Recall(Th_{k-1}) \\ \Delta_{Precision}^{k} = Precision(Th_{k-1}) + Precision(Th_{k}) \end{cases}$ 

 $VUS\text{-}ROC = \frac{1}{4} \sum_{w=1}^{L} \sum_{k=1}^{w} \Delta^{(k,w)} * \Delta^{w}$  $(\Delta^{(k,w)} = \Delta^k_{TPR_{\ell_w}} * \Delta^k_{FPR_{\ell_w}} + \Delta^k_{TPR_{\ell_{w-1}}} * \Delta^k_{FPR_{\ell_{w-1}}}$ Time Complexity: O(NLT)With: T: the time series length *N*: the number of thresholds *L*: the number of buffer lengths  $Pr_{\ell_W} = Re_{\ell_W} = Pr_{\ell_W-1} = Re_{\ell_W-1}$  $= Recall_{\ell_{\mathcal{W}}}(Th_{k}) - Recall_{\ell_{\mathcal{W}}}(Th_{k-1})$  $\Delta^{k}_{Re_{\ell_{W}}} \\ \Delta^{k}_{Pr_{\ell_{W}}}$  $= Precision_{\ell_{w}}(Th_{k-1}) + Precision_{\ell_{w}}(Th_{k})$  $= |\ell_{w} - \ell_{w-1}|$ 

#### How is it computed?










A solution?



A solution?



Evaluation measures: VUS

#### A solution?





### Part 4: Anomaly Detection Benchmarks

#### **HEX/UCR** [18]

Set of 250 time series with labels.

Details

- The labels have been manually checked and are reliable
- Each time series contains only 1 labeled anomaly

HEX/UCR [18]	TimeEval [5]
Set of 250 time series with labels.	Set of 976 time series with labels.
Details	Details
<ul> <li>The labels have been manually checked and are reliable</li> </ul>	<ul> <li>New synthetic benchmark GutenTag used to tune parameters</li> </ul>
<ul> <li>Each time series contains only 1 labeled anomaly</li> </ul>	<ul> <li>Only Time series with low contamination rate (&lt; 0.1)</li> </ul>
	- Time series with at least one methods above 0.8 AUC-ROC

HEX/UCR [18]	TimeEval [5]	TSB-UAD [19]
Set of 250 time series with labels.	Set of 976 time series with labels.	Set of 2000 time series with labels.
Details	Details	Details
<ul> <li>The labels have been manually checked and are reliable</li> </ul>	<ul> <li>New synthetic benchmark GutenTag used to tune parameters</li> </ul>	<ul> <li>Collected as proposed in the literature (no filtering based on contamination, size or label quality)</li> </ul>
<ul> <li>Each time series contains only 1 labeled anomaly</li> </ul>	<ul> <li>Only Time series with low contamination rate (&lt; 0.1)</li> <li>Time series with at least one methods above 0.8 AUC-ROC</li> </ul>	<ul> <li>Artificial and synthetic data generation methods for reliable labels</li> </ul>





**Observations on TimeEval [5]:** 

[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. VLDB Endow. 15, 9 (May 2022), 1779–1797.

	Methods				AUC-ROC
	Sub-LOF [22] GrammarViz [120] DWT-MLEAD [134] VALMOD [82] SAND [17] Left STAMPi [156] Series2Graph [16]	2 % 3 % 0 % 1 % 5 % 2 % 0 %	0 % 0 % 0 % 1 % 0 %	0 % 0 % 0 % 11 % 22 % 1 % 5 %	
Unsupervised	ARIMA [65] PCI [157] STOMP [164] STAMP [156] Triple ES [1] NumentaHTM [3] NormA-SJ [15] Sub-IF [83] MedianMethod [10] SR [112] PS-SVM [85] PST [128] SSA [155] HOT SAX [70]	7 % 0 % 2 % 4 % 15 % 0 % 10 % 0 % 0 % 12 % 0 % 12 % 0 %	0 % 0 % 0 % 0 % 0 % 1 % 0 % 0 % 0 % 4 % 0 %	0 % 0 % 0 % 9 % 0 % 3 % 0 % 0 % 0 % 0 %	
emi-supervised	TSBitmap [144] DSPOT [122] FFT [111] S-H-ESD [62] Donut [150] RForest [21] IE-CAE [44] XGBoosting [34] OceanWNN [143] Bagel [79] SR-CNN [112] TARZAN [71]	24 % 0 % 6 % 0 % 0 % 1 % 12 % 0 % 0 % 0 %	1 % 0 % 0 % 0 % 0 % 0 % 0 % 0 % 0 % 0 %	1 % 0 % 0 % 49 % 2 % 1 % 10 % 2 % 1 % 18 %	

#### **Observations on TimeEval [5]:**

 Distance-based and Density-based methods have a better accuracy (AUC-ROC) than forecasting and reconstruction-based approaches

[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. VLDB Endow. 15, 9 (May 2022), 1779–1797.

	Meth	ods			AUC-ROC
supervised	Weth           Sub-LOF [22]           GrammarViz [120]           DWT-MLEAD [134]           VALMOD [82]           SAND [17]           Left STAMPi [156]           Series2Graph [16]           ARIMA [65]           PCI [157]           STOMP [164]           STAMP [156]           Triple ES [1]           NumentaHTM [3]           NormA-SJ [15]           Sub-IF [83]	Ods 2 % 3 % 0 % 1 % 5 % 2 % 0 % 7 % 0 % 2 % 4 % 15 % 0 % 10 % 0 %	0 % 0 % 9 % 0 % 0 % 0 % 0 % 0 % 0 % 0 %	0 % 0 % 11 % 22 % 1 % 5 % 0 % 0 % 0 % 0 % 3 % 0 %	AUC-ROC
I Un	MedianMethod [10] SR [112] PS-SVM [85] PST [128] SSA [155] HOT SAX [70] TSBitmap [144] DSPOT [122] FFT [111] S-H-ESD [62]	0 % 0 % 12 % 0 % 1 % 24 % 0 % 6 % 0 % 0 %	0 % 0 % 4 % 0 % 1 % 0 % 0 % 0 %	0 % 0 % 0 % 0 % 1 % 0 % 0 % 49 %	
Semi-supervised	Donut [150] RForest [21] IE-CAE [44] XGBoosting [34] OceanWNN [143] Bagel [79] SR-CNN [112] TARZAN [71]	1 % 12 % 0 % 0 % 19 % 22 % 0 %	1 % 0 % 0 % 0 % 0 % 0 %	2 % 0 % 1 % 0 % 10 % 2 % 1 % 18 %	

#### **Observations on TimeEval [5]:**

- Distance-based and Density-based methods have a better accuracy (AUC-ROC) than forecasting and reconstruction-based approaches
- Semi-supervised methods are not outperforming Unsupervised approaches

[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. VLDB Endow. 15, 9 (May 2022), 1779–1797.

Meth	ods			AUC-ROC
Meth Sub-LOF [22] GrammarViz [120] DWT-MLEAD [134] VALMOD [82] SAND [17] Left STAMPi [156] Series2Graph [16] ARIMA [65] PCI [157] STOMP [164] STAMP [156] Triple ES [1] NumentaHTM [3] NormA-SJ [15] Sub-IF [83] MedianMethod [10] SR [112] PS-SVM [85] PST [128] SSA [155] HOT SAX [70] TSBitmap [144] DSPOT [122]	Ods 2 % 3 % 0 % 1 % 5 % 2 % 0 % 7 % 0 % 2 % 4 % 0 % 15 % 0 % 10 % 0 % 12 % 0 % 12 % 0 % 1 % 5 % 2 % 0 % 1 % 0 % 1 % 0 % 1 % 0 % 1 % 0 % 0 % 1 % 0 % 0 % 0 % 0 % 0 % 0 % 0 % 0	0 % 0 % 9 % 0	0% 0% 11% 22% 1% 0% 0% 0% 0% 0% 0% 0% 0	AUC-ROC
FFT [111] S-H-ESD [62]	0 % 0 %	0 % 0 %	0 % 49 %	: ∰
Donut [150] RForest [21] IE-CAE [44] XGBoosting [34] OceanWNN [143] Bagel [79] SR-CNN [112]	1 % 12 % 0 % 0 % 19 % 22 %	1 % 0 % 0 % 0 % 0 %	2 % 0 % 1 % 0 % 10 % 2 % 1 %	
	Meth Sub-LOF [22] GrammarViz [120] DWT-MLEAD [134] VALMOD [82] SAND [17] Left STAMPi [156] Series2Graph [16] ARIMA [65] PCI [157] STOMP [164] STAMP [156] Triple ES [1] NumentaHTM [3] NormA-SJ [15] Sub-IF [83] MedianMethod [10] SR [112] PS-SVM [85] PST [128] SSA [155] HOT SAX [70] TSBitmap [144] DSPOT [122] FFT [111] S-H-ESD [62] Donut [150] RForest [21] IE-CAE [44] XGBoosting [34] OceanWNN [143] Bagel [79] SR-CNN [112] TARZAN [71]	Sub-LOF [22]       2         GrammarViz [120]       3         DWT-MLEAD [134]       0         VALMOD [82]       1         SAND [17]       5         Series2Graph [16]       0         ARIMA [65]       7         PCI [157]       0         STOMP [164]       2         STOMP [164]       2         STOMP [156]       4         Triple ES [1]       10         NumentaHTM [3]       0         NormA-SJ [15]       10         SR [112]       0         PST [128]       0         SSA [155]       1         FSSA [155]       1         PST [128]       0         SSA [155]       1         FFT [111]       0         SSA [155]       1         FFT [111]       0         SH-ESD [62]       0         Donut [150]       1         FFT [111]       0         SH-ESD [62]       0         Donut [150]       1         RForest [21]       12         RForest [21]       12         KGBoosting [34]       0         SR-CNN [143]       0	Sub-LOF [22]       2 %       0 %         GrammarViz [120]       3 %       0 %         DWT-MLEAD [134]       0 %       0 %         VALMOD [82]       1 %       9 %         SAND [17]       5 %       1 %         Left STAMPi [156]       2 %       0 %         Series2Graph [16]       0 %       0 %         ARIMA [65]       7 %       0 %         STOMP [164]       2 %       0 %         STAMP [156]       4 %       0 %         STAMP [156]       4 %       0 %         STAMP [156]       1 %       0 %         STAMP [156]       1 %       0 %         NumentaHTM [3]       0 %       0 %         NormA-SJ [15]       10 %       0 %         SR [112]       0 %       0 %         PST [128]       0 %       0 %         PST [128]       0 %       0 %         SSA [155]       1 %       0 %         FFT [111]       0 %       0 %         SPOT [122]       6 %       0 %         SPOT [122]       6 %       0 %         SSA [155]       1 %       1 %         DSPOT [122]       6 %       0 %	Methods         Sub-LOF [22]       2 %       0 %       0 %         GrammarViz [120]       3 %       0 %       0 %         DWT-MLEAD [134]       0 %       0 %       1 %         VALMOD [82]       1 %       9 %       11 %         SAND [17]       5 %       1 %       22 %         Left STAMPi [156]       2 %       0 %       0 %         Series2Graph [16]       0 %       0 %       0 %         PCI [157]       0 %       0 %       0 %         STOMP [164]       2 %       0 %       0 %         STAMP [156]       4 %       0 %       0 %         NumentaHTM [3]       0 %       0 %       0 %         NormA-SJ [15]       10 %       1 %       3 %         Sub-IF [83]       0 %       0 %       0 %         SR [112]       0 %       0 %       0 %         PS-SVM [85]       12 %       0 %       0 %         PST [128]       0 %       0 %       0 %         SSA [155]       1 %       0 %       0 %         PST [128]       0 %       0 %       0 %         DSPOT [122]       6 %       0 %       0 %      <

#### **Observations on HEX/UCR [18]:**

 Distance-based methods have a better accuracy (AUC-ROC) than forecasting and distribution-based approaches

[18] R. Wu and E. Keogh, "Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress" in IEEE Transactions on Knowledge & Data Engineering, vol. 35, no. 03, pp. 2421-2429, 2023.



#### **Observations on HEX/UCR [18]:**

 Distance-based methods have a better accuracy (AUC-ROC) than forecasting and distribution-based approaches

[18] R. Wu and E. Keogh, "Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress" in IEEE Transactions on Knowledge & Data Engineering, vol. 35, no. 03, pp. 2421-2429, 2023.



#### **Observations on TSB-UAD [19]:**

- Distance-based methods have a better accuracy (AUC-ROC) than forecasting-based methods.
- Isolation Forest (distribution-based and not proposed for time series) have also a strong accuracy
- AutoEncoder (AE) is also very accurate.

[19] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. Proc. VLDB Endow. 15, 8 (April 2022), 1697–1711.



#### **Observations on TSB-UAD [19]:**

- Distance-based methods have a better accuracy (AUC-ROC) than forecasting-based methods.
- Isolation Forest (distribution-based and not proposed for time series) have also a strong accuracy
- AutoEncoder (AE) is also very accurate.

[19] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. Proc. VLDB Endow. 15, 8 (April 2022), 1697–1711.



[19] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. Proc. VLDB Endow. 15, 8 (April 2022), 1697–1711.

**Observations on TSB-UAD [19]:** 



AUC-ROC

AUC-ROC

KDD 2025 | 03/08/2025 | 209

### Anomaly Detection methods:

Experimental evaluation

[19] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. Proc. VLDB Endow. 15, 8 (April 2022), 1697–1711.

### Anomaly Detection methods:

Experimental evaluation

#### **Observations on TSB-UAD [19]:**

- Forecasting methods (LSTM and CNN) are very accurate for point anomalies
- But have poor performances on sequencebased anomalies.



[19] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael



LSTM

OCSVM

0.0

0.2

KDD 2025 | 03/08/2025 | 211

0.4

0.6

AUC-ROC

0.8

1.0

1.0

0.8

LSTM

MP

0.0

0.2

0.4

0.6

AUC-ROC

### Anomaly Detection methods: Experimental evaluation

#### **Observations on TSB-UAD [19]:**

The ratio of normal/abnormal points has a strong impact on the methods ranking.

J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. Proc. VLDB Endow. 15, 8 (April 2022), 1697–1711.





[27] Paul Boniol, John Paparrizos, Yuhao Kang, Themis Palpanas, Ruey S. Tsay, Aaron J. Elmore, and Michael J. Franklin. 2022. Theseus: navigating the labyrinth of time-series anomaly detection. Proc. VLDB Endow. 15, 12 (August 2022), 3702–3705.













## Benchmark Practice: Dataset Construction



## Benchmark Practice: Dataset Construction



## Benchmark Practice: Dataset Construction


#### Benchmark Practice: Dataset Construction

Category	Split	# TS	Avg Length	Avg Anomaly Length	Avg # Anomalies	Anomaly Ratio
	All	870	38814.1	179.5	39.7	2.4%
TSB-AD-U	Eval	350	51886.7	321.3	46.6	4.5%
	Tuning	48	47143.3	185.9	82.6	3.5%
	All	200	107760.4	582.6	71.1	5.1%
TSB-AD-M	Eval	180	108826.7	591.2	67.7	5.0%
	Tuning	20	98164.1	504.7	101.1	5.7%







**TSB-AD-U** 

VUS-PR Ranking Sub-PCA W KShapeAD 2 W POLY Series2Graph 4 5 MOMENT (FT) **MOMENT (ZS)** 6 7 **KMeansAD** USAD 8 9 Sub-KNN MatrixProfile 10 11 SAND 12 CNN

① Top-performing methods been overlooked for many years



**TSB-AD-U** 



① Top-performing methods been overlooked for many years

② Performance of time-series foundation models shows promise



**TSB-AD-M** 

9				
	CNN			
2	OmniAnomaly			
3	PCA			
4	LSTMAD			
5	USAD AutoEncoder KMeansAD			
6				
7				
8	CBLOF			
9	MCD			
10OCSVM11Donut				
		2	RobustPCA	
	4 5 6 7 8 9 0 11 2			

VUS-PR Rankina

③ Neural-network-based methods strive in multivariate cases



**TSB-AD-M** 

CNN OmniAnomaly 2 PCA LSTMAD 4 **USAD** 5 AutoEncoder 6 7 **KMeansAD** CBLOF 8 MCD 9 10 OCSVM 11 Donut **RobustPCA** 12

🕨 VUS-PR Ranking

③ Neural-network-based methods strive in multivariate cases

④ Simpler architectures generally outperform more complex designs



1070 Curated Time Series

40 TSAD Algorithms

10 Evaluation Measures



[27] Liu, Q. and Paparrizos, J., 2024. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. Advances in Neural Information Processing Systems, 37, pp.108231-108261.

# Part 5: Automated Solutions for Anomaly Detection

# Automated Solutions: Background

#### Motivation:

- No one-size-fits-all model: How can we automatically identify the best anomaly detector given a time series?
- No comprehensive evaluation benchmark



[29] Maroua Bahri, Flavia Salutari, Andrian Putina, and Mauro Sozio: AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics 14(2): 113-126 (2022).
[41] Qinghua Liu, Seunghak Lee, and John Paparrizos: TSB-AutoAD: Towards Automated Solutions for Time-Series Anomaly Detection. VLDB 2025.

KDD 2025 | 03/08/2025 | 231

# Automated Solutions: Background

#### Motivation:

- No one-size-fits-all model: How can we automatically identify the best anomaly detector given a time series?
- No comprehensive evaluation benchmark

#### Challenge:

- Lack of labeled data
- Absence of universal objective function

[29] Maroua Bahri, Flavia Salutari, Andrian Putina, and Mauro Sozio: AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics 14(2): 113-126 (2022).
[41] Qinghua Liu, Seunghak Lee, and John Paparrizos: TSB-AutoAD: Towards Automated Solutions for

Time-Series Anomaly Detection. VLDB 2025.



# Automated Solutions: Taxonomy



#### (a) Model Selection:

Selecting the best anomaly detector from a predefined candidate model set.

- (a.1) Meta-learning-based
- (a.2) Internal Evaluation

# Automated Solutions: Taxonomy



# Automated Solutions: Taxonomy



(c) Model Generation Constructing of a completely new model based on the candidate set, which can then operate as an anomaly detector to produce scores

Definition: Using insights from historical labeled datasets to select the best model for new data

- Classification: MSAD
- Regression: SATzilla, UReg, CFact
- Nearest Neighbor: ARGOSMART
- Other Optimization: ISAC, MetaOD





Image from [29]: Model Selection Pipeline.

[29] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, Themis Palpanas. 2023. Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. Proceedings of the VLDB Endowment 16(11): 3418-3432.

KDD 2025 | 03/08/2025 | 237



Performance measures: F-score, AUC-PR, VUS-PR ...

Time Series For Training Candidate Model Set Performance Matrix

 $D_3$ 

0.9

0.7

0.6

...



[29] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, Themis Palpanas. 2023. Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. Proceedings of the VLDB Endowment 16(11): 3418-3432.

KDD 2025 | 03/08/2025 | 239



Time Series For Training Candidate Model Set Performance Matrix

Label



[32] Lin Xu, Frank Hutter, Holger H Hoos, Kevin Leyton-Brown. 2008. SATzilla: portfolio-based algorithm selection for SAT. Journal of Artificial Intelligence Research 32: 565-606.



Time Series For Training Candidate Model Set Performance Matrix

Label

Definition: Evaluate the effectiveness of a model without any reliance on external information

- Stand-alone: Clustering Quality, EM&MV, Synthetic anomaly injection
- Collective: Model Centrality, Rank Aggregation





#### Image from [28]: Internal Evaluation workflow.



#### Image from [28]: Internal Evaluation workflow.



Image from [28]: Internal Evaluation workflow.



Image from [28]: Internal Evaluation workflow.

# Automated Solutions: Model Ensembling

Definition: Integrate predictions from the candidate model set

- Full: OE
- Selective: SELECT, HITS, IOE, AutoTSAD



# Automated Solutions: Model Generation

Definition: Creating an entirely new model tailored to a specific dataset based on the predefined model set.

- Full: OE
- Selective: SELECT, HITS, IOE, AutoTSAD





Accuracy overview of automated solutions for time-series anomaly detection in terms of accuracy. Methods are arranged from left to right based on their performance, with the highest accuracy (measured by VUS-PR) on the left.



(a) illustration of the relationshipbetween VUS-PR and averagedetection time across the benchmark



(a) Meta-learning (ID) vs. Others



(b) Meta-learning (OOD) vs. Others





(C) ID vs. OOD



KDD 2025 | 03/08/2025 | 254

# Part 6: Conclusion and Open Problems

### Open Problems: Context-aware



### Open Problems: Context-aware



### Open Problems: Context-aware


## **Open Problems:** Foundation Model



# **Open Problems:** Foundation Model

Emerging trends in developing time series foundation model

	Sundial (ICML25)	Time-MOE (ICLR25)	Moirai (ICML24)	MOMENT (ICML24)	TimesFM (ICML24)	Chronos (TMLR24)	Lag-Llama (NIPSw23)
Architecture	Decoder	Decoder	Encoder	Encoder	Decoder	EncDec	Decoder
Model Size	32M 128M 444M	113M 453M 2.4B	14M 91M 311M	40M 125M 385M	17M 70M 200M	46M 200M 710M	200M
Context Length	<=2880	<=4096	<=5000	=512	<=512	<=512	<=1024
Objectives	Flow Matching	Huber Loss	Log Likelihood	MSE	MSE	Cross Entropy	Log Likelihood
Task	F	F	F	CLF, AD, F	F	F	F

# **Open Problems:** Foundation Model

Emerging trends in developing time series foundation model

Lack of benchmarking and adaptation for AD task

	Sundial (ICML25)	Time-MOE (ICLR25)	Moirai (ICML24)	MOMENT (ICML24)	TimesFM (ICML24)	Chronos (TMLR24)	Lag-Llama (NIPSw23)
Architecture	Decoder	Decoder	Encoder	Encoder	Decoder	EncDec	Decoder
Model Size	32M 128M 444M	113M 453M 2.4B	14M 91M 311M	40M 125M 385M	17M 70M 200M	46M 200M 710M	200M
Context Length	<=2880	<=4096	<=5000	=512	<=512	<=512	<=1024
Objectives	Flow Matching	Huber Loss	Log Likelihood	MSE	MSE	Cross Entropy	Log Likelihood
Task	F	F	F	CLF, AD, F	F	F	F

# Open Problems: Agentic Design

Please find the anomaly point in the time series data.



Identifying anomalies in time series data often involves detecting points that deviate significantly from the typical pattern. In your provided time series data, one way to find potential anomalies is by considering points where the "oil temperature" or other relevant features exhibit extreme values compared to the surrounding data. **Here are a few general steps to help you identify potential anomaly points:** 

1.Visual Inspection. (1) Plot the time series data for "oil temperature" and other features. (2) Look for points that stand out from the general pattern.

2.Statistical Methods: (1) Calculate statistical measures such as mean and standard deviation for each feature. (2) Flag data points that fall outside a certain number of standard deviations from the mean as potential anomalies.



	Ľ,
ď	<u></u> )
-1	
1	

Just show me the results but not methods.



Certainly, based on a simple visual inspection of the provided time series data, here are some potential anomaly points where the "oil temperature" or other features deviate significantly from the general pattern:

1. Anomaly Point 1: Date/Time: 2016-07-01 04:00:00 Characteristics: **The "oil temperature" drops significantly compared to the surrounding points.** 

2.Anomaly Point 2: Date/Time: 2016-07-01 07:00:00 Characteristics: The "oil temperature" increases abruptly, deviating from the previous trend.

Interpretability !

A time series anomaly detector with natural language interface

Image from [36]: Human interaction with ChatGPT for TSAD tasks.

# Open Problems: Agentic Design



Multi-modal time series datasets

Image from [42]: The demonstration of the Time-MQA with context enhancement.

[42] Kong, Y., Yang, Y., Hwang, Y., Du, W., Zohren, S., Wang, Z., ... & Wen, Q. (2025). Time-mqa: Time series multi-task question answering with context enhancement. arXiv preprint arXiv:2503.01875.

### If you are interested in anomaly detection in time series...

#### A review on outlier/anomaly detection in time series data

ANE BLÁZQUEZ-GARCÍA and ANGEL CONDE, Ikerian Technology Research Centre, Basque Research and Technology Alliance (BRTA), Spain

USUE MORI, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain

JOSE A. LOZANO, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain and Basque Center for Applied Mathematics (BCAM), Spain

Recent advances in technology have brought major breakthroughs in data collection, enabling a large amount of data to be gathered over time and thus generating time series. Mining this data has become an important task for researchers and practitioners in the past few yaras, including the detection of outliers or anomalies that may represent errors or events of interest. This review aims to provide a structured and comprehensive state-of-the-art on outlier detection techniques in the context of time series. To this end, a taxonomy is presented based on the main aspects that characterize an outlier detection technique.

Additional Key Words and Phrases: Outlier detection, anomaly detection, time series, data mining, taxonomy, software

#### 1 INTRODUCTION

Recent advances in technology allow us to collect a large amount of data over time in diverse research areas. Observations that have been recorded in an orderly fashion and which are correlated in time constitute a time series. Time series data mining aims to extract all meaningful knowledge from this data, and several mining taska (e.g., classification, clustering, forecasting, and outlier detection) have been considered in the literature [Esling and Agon 2012; Fu 2011; Ratanamhatana et al. 2010].

Outlier detection has become a field of interest for many researchers and practitioners and is now one of the main tasks of time series data mining. Outlier detection has been studied in a variety of application domains such as credit card fraud detection, intrusion detection in cybersecurity, or fault diagnosis in industry. In particular, the analysis of outliers in time series data examines anomalous behaviors across time [Cupta et al. 2014a]. In the first study on this topic, which was conducted by Fox [1972], too types of outliers in mix-static time series were defined type I, which affects a single observation, and type II, which affects both a particular observation and the subsequent observations. This work was first extended to four outlier types [Tsay 1988], and then to the case of multivariate time series was been proposed in the literature. However, to this day, there is still no consensus on the terms used [Carreño et al. 2019]; for example, outlier observations are often referred to as anomalies, discordant observations, discords, exceptions, aberrations, surprises, precultarities or contaminants.

Authori addresses: Ane Blirquer-Garcia, abharquer@dwrlan.ee; Arqel Conde, aconde@dwrlan.es, Karlan Technology Researd Centre, Bauyee Researds and Technology Allance (BRTA), P. JM. Arizmendiarrieta, 2., Arzasate/Mondragón, 20500, Spain; Usue Mori, usue.mori@dehu.es, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Bauge County (UPVIDU), Manael de Lardizabal, 1, 20018,

A. Blazquez-Garcia et al. ACM Computing Survey (2021) [24]

### If you are interested in anomaly detection in time series...

#### Dive into Time-Series Anomaly Detection: A Decade Review

#### A review on outlier/anomaly detection in time series data

ANE BLÁZQUEZ-GARCÍA and ANGEL CONDE, Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), Spain

USUE MORI, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain

JOSE A. LOZANO, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain and Basque Center for Applied Mathematics (BCAM), Spain

Recent advances in technology have brought major breakthroughs in data collection, enabling a large amount of data to be gathered over time and thus generating time series. Mining this data has become an important task for researchers and practitioners in the past for years, including the detection of outliers or anomalies that may represent errors or events of interest. This review aims to provide a structured and comprehensive state-of-the-art on outlier detection techniques in the context of time series. To this end, a taxonomy is presented based on the main aspects that characterize on outlier detection technique.

Additional Key Words and Phrases: Outlier detection, anomaly detection, time series, data mining, taxonomy, software

#### 1 INTRODUCTION

Recent advances in technology allow us to collect a large amount of data over time in diverse research areas. Observations that have been recorded in an orderly fashion and which are correlated in time constitute a time series. Time series data mining aims to extract all meaningful knowledge from this data, and several mining tasks (e.g., classification, clastering, forecasting, and outlier detection) have been considered in the literature [Ealing and Agon 2012; Fu 2011; Ratamanihatane et al. 2010].

Outlier detection has become a field of interest for many researchers and practitioners and is now one of the main tasks of time series data mining. Outlier detection has been studied in a variety of application domains such as credit card fraud detection, intrusion detection in cybersecurity, or fault diagnosis in industry. In particular, the analysis of outliers in time series data examines anomalous behaviors across time [Cupta et al. 2014a]. In the first study on this topic, which was conducted by Fox [1972], two types of outliers in univariate time series were defined: type I, which affects a single observation; and type II, which affects both a particular observation and the subsequent observations. This work was first extended to four outlier types [Tasy 1988], and then to the case of multivariate time series [Tasy et al. 2000]. Since then, many definitions of the term outlier and numerous detection methods have been proposed in the literature. However, to this day, there is still no consensus on the terms used [Carreho et al. 2019]; for example, outlier observations are often referred to as anomalies, discordant observations, discords, exceptions, aberrations, surprises, precultarities or contaminants.

Authors' addresses: Ane Bläzquez-García, ablazquez@ikerlan.ex; Angel Conde, aconde@ikerlan.ex; Berlan Technology Research Centre, Basque Research and Technology Alliance (BKTA), P: JM. Arizmendiarrieta, 2, Arrastate/Mondragón, 20500, Spain; Usue Mori, usue mori@bnues, Intelligent Systems (Group (USG), Department of Computer Sintere and Artificial Intelligence, University of the Basque County (UPV/BIN), Manuel de Lardinaba I, 2500

### A. Blazquez-Garcia et al. ACM Computing Survey (2021) [24]

PAUL BONIOL, Inria, DI ENS, PSL, CNRS, France QINGHUA LIU, The thin State University, USA MINGYI HUANG, The Ohio State University, USA THEMIS PALPANAS, Université Paris Cité; JUF, France JOHN PAPARRIZOS, The Ohio State University, USA

Recent advances in data collection technology, accompanied by the ever-rising volume and velocity of streaming data, underscore the vital need for time series analytics. In this regard, time-series anomaly detection has been an important activity, entailing various applications in fields such as cybers security, financial markets, law enforcement, and health care. While traditional literature on anomaly detection centered on statistical measures, the increasing number of machine laming algorithms increative stars and summarizes structured, general characterization of the research methods for time-series anomaly detection. This survey groups and summarizes anomaly detection existing solutions under a process-centric taxonomy in the time series context. In addition to giving an original categorization of anomaly detection methods, we also perform a meta-analysis of the literature and outling general trends in time-series anomaly detection estende.

#### ACM Reference Format:

Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Papartizos. 2024. Dive into Time-Series Anomaly Detection: A Decade Review. In Proceedings of Makes use to enter the correct conference till from your rights confirmation email (Conference acronym '20, ACM, New York, NY, USA, 5) Ispace. https://doi.org/2000/2000/20002000

#### 1 Introduction

A wide range of cost-effective sensing, networking, storage, and processing solutions enable the collection of enormous amounts of measurements over time [10<sup>3–111</sup>, 122, 137, 138, 141, 143, 179, 181, 186]. Recording these measurements results in an ordered sequence of re4-wiled data points commonly referred to a site *mesrek*. Nore generic terms, such as *data series* or *data sequences*, have also been used to refer to cases where the ordering of data relies on a dimension other than time (e.g., the angle in data from astronomy, the mass in data from spectrometry, or the position in data from biology) [176]. Analytical tasks over time series data are necessary virtually in every scientific discipline and their corresponding industries [11, 64, 62, 78, 161, 182, 190–192, 201]. Including in astronomy [4, 102, 264], biology [11–13, 64], economics [26, 74, 148, 155, 213, 221, 240], energy sciences [6, 9, 158], engineering [112, 162, 203, 243, 248], environmental sciences [77, 48, 100, 101, 164, 207, 247], medicine [52, 199, 206], neuroscience [21, 19], and social sciences [26, 160]. The analysis of time series has become increasingly prevalent for understanding a multitude of natural or human-made processes [157, 188]. Unfortunately, inherent complecities in the data generation of these

P. Boniol et al. Arxiv (2025) [28]

### If you are interested in anomaly detection in time series...

#### A review on outlier/anomaly detection in time series data

ANE BLÁZQUEZ-GARCÍA and ANGEL CONDE, Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), Spain

USUE MORI, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain

JOSE A. LOZANO, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain and Basque Center for Applied Mathematics (BCAM), Spain

Recent advances in technology have brought major breakthroughs in data collection, enabling a large amount of data to be gathered over time and thus generating time series. Mining this data has become an important task for researchers and practitioners in the past for years, including the detection of outliers or anomalies that may represent errors or events of interest. This review aims to provide a structured and comprehensive state-of-the-art on outlier detection techniques in the context of time series. To this end, a taxonomy is presented based on the main aspect that characterize an outlier detection technique.

Additional Key Words and Phrases: Outlier detection, anomaly detection, time series, data mining, taxonomy, software

#### 1 INTRODUCTION

Recent advances in technology allow us to collect a large amount of data over time in diverse research areas. Observations that have been recorded in an orderly fashion and which are correlated in time constitute a time series. Time series data mining aims to extract all meaningful knowledge from this data, and several mining tasks (e.g., classification, classtering, forecasting, and outlier detection) have been considered in the literature [Ealing and Agon 2012; Fu 2011; Ratanamhatana et al. 2010].

Outlier detection has become a field of interest for many researchers and practitioners and is now one of the main tasks of time series data mining. Outlier detection has been studied in a variety of application domains such as credit card fraud detection, intrusion detection in cybersecurity, or fault diagnosis in industry, in particular, the analysis of outliers in time series data examines anomalous behaviors across time [Copta et al. 2014a]. In the first study on this topic, which was conducted by Fox [1972], two types of outliers in nuivariate time series were defined: type I, which affects a single observation; and type I, which affects both a particular observation and the subsequent observations. This work was first extended to four outlier types [Tasy 1988], and then to the case of multivariate time series [Tasy 41, 2000]. Since then, many definitions of the term outlier and numerous detection methods have been proposed in the literature. However, to this day, there is still no consensus on the terms used [Carreho et al. 2019], for example, outlier observations are often referred to as anomalies, discordant observations, discords, exceptions, aberrations, surprises, precultarities or contaminants.

Authors' addresses: Ane Bläzquez-García, ablazquez@ikerlan.es; Angel Conde, aconde@ikerlan.es; Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), P: JM. Arizmendiarrieta, 2, Arrastate/Mondragón, 2050, Spain; Usue Mort, usue morigêneus, Findigent Systems (Group (USU), Department of Computer Science and Artificial Intelligence, University of the Basque County (UPV/ENIV). Manuel de Lardinaba I; 2009

### A. Blazquez-Garcia et al. ACM Computing Survey (2021) [24]

#### Dive into Time-Series Anomaly Detection: A Decade Review

PAUL BONIOL, Inria, DI ENS, PSL, CNRS, France QINGHUA LIU, The Ohio State University, USA MINGYI HUANG, The Ohio State University, USA THEMIS PALPANAS, Université Paris Cité; IUF, France

JOHN PAPARRIZOS, The Ohio State University, USA

Recent advances in data collection technology, accompanied by the ever-rising volume and velocity of atreaming data, undernove the vital need for time series analytics. In this regard, time-series anomaly detection has been an important activity, entailing various applications in fields work as velow security. financial markets, low enforcement, and health care. While traditional literature on anomaly detection is centered on attaitistical measures, the increasing number of machine learning algorithms in recent years call for a structured, general characterization of the research methods for time-series anomaly detection. This survey groups and summarizes anomaly detection cessing solutions under a process-centric taxonomy in the time series context. In addition to giving an original categorization of anomaly detection methods, we also perform a meta-analysis of the literature and outline general trends in time-series anomaly detection research.

#### ACM Reference Format:

Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparitos. 2024. Dive into Time-Series Anomaly Detection: A Decade Review. In Proceedings of Males use to enter the correct conference tille from your rights notifrmation email (Conference acronym '20, ACM, New York, NY, USA, 51 gaes, https://doi.org/2005/00020200200020

#### 1 Introduction

A wide range of cost-effective sensing, networking, storage, and processing solutions enable the collection of earonmous amounts of measurements over time [109–111, 122, 137, 138, 141, 143, 179, 181, 186]. Recording these measurements results in an ordered sequence of rel-levaled data points commonly referred to a stime series. Nore generic terms, such as data series or data sequences, have also been used to refer to cases where the ordering of data relies on a dimension other than time (e.g., the angle in data from astronomy, the mass in data from spectrometry, or the position in data from biology [176]. Analytical tasks over time series data are necessary virtually in every scientific discipline and their corresponding industries [14.6.4, 6.78, 161, 182, 190–192, 201], including in astronomy [4, 102, 245], biology [11–13, 4.4], economics [26, 74, 148, 155, 213, 221, 240], energy sciences [6, 9, 158], engineering [112, 162, 203, 243, 248], environmental sciences [77, 84, 100, 101, 164, 207, 247], meldrine [52, 199, 206], neuroscience [21, 119], and social sciences [26, 160]. The analysis of time series has become increasingly prevalent for understanding a multitude of natural or human-made processes [157, 188]. Unfortunately, inherent complexities in the data generation of these

P. Boniol et al. Arxiv (2025) [28]

### Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress

#### Renjie Wu and Eamonn J. Keogh

Abstract—Time series aromally detection has been a perionially important topic in data science, with papers daing back to the 1950s. However, in recent years there has been an explosion of interest in this topic, much of it driven by the scuoses of deep learning in other domains and for other time series tasks. Most of these papers test on ore or more of a handlu of popular benchmark datasets, created by "show. Namenta, NASA, etc., to hit is work we make a surprising dain." The majointy of the individual exemptions in these datasets suffer from one or more of four farses. Because of these four farses, we believe that many published companies on a damady detection algorithm may be unreliable, and more importantly, much of the apparent progress in recent years may be illusionary. In addition to demonstrating these dams, with this paper we introduce the UCR Time. Sories Anomaly Archive. We believe that this resource waite performs an similar or use the UCR these: Classification Archive, by providing the community with a benchmark that allows meaningful comparisons between approaches and a meaningful appress of evenil progress.

----- + -------

Index Terms—Anomaly detection, benchmark datasets, deep learning, time series analysis

1 INTRODUCTION					
TME series anomaly detection has been a peremailing back to the dawn of computer science, with papers dating back to the dawn of computer science (11, However, in the last five years there has been an coplosion of interest in this topic, with at least one or two papers on the topic spearing each year in virtually every database, data mining and machine learning conference, including SIGKDD [2] [0] (CMO [4] (CDE SIGMOD / UDE) etc. A large fraction of this increase in interest stems to be siderable success of deep learning in other domains and from other time series tasks such as classification. Most of these papers test on one or more of a handful of popular benchmark datasets, created by Yahoo [5], etc. In this work we make a surprising claim. The majority of the individual exercises tasks such as the rest of the mone or more of four flaws. These flaws are trivialing, unrelistic biss. Because of these four flaws, we believe that much of the appertn progress in necent years may be	neural networks, and a variational auto-encoder (V&E) core- sampling model." This description sounds like it has many "moving parts", and indeed, the dozen or so explicitli- listed parameters include: convolution filter, activation, kernel size, strides, padding, LSTM input size, dense in- put size, softmas loss function, window size, kerning rate and batch size. All of this is to demonstrate "accuracy co- ceeding 0.30 (on a subset of the Vahoo's anomaly detection benchmert, datasets)." However, as we will show, much of with a single line of code and a deve minutes of effort. This "ome-line-of-code" argument is so unusual that is us over hyreviswing it before we tormally demonstrate it in Section 2.2 below. Almost daily, the popular press vanuts a new achievement of deep learning. Picking one at random, in a recent paper [8], we learn that deep learn- ing can be used to classify messatify species. In particu- lar, the proposed algorithm had an accurred of 37% when distinguishing Aders excesses from Culter trainningu- dues. Should we be impressed? One of the current authors Suppose however that someone downloaded the origi-				
R. Wu et al. TKDE (2021) [18]					
[±0]					

Google search for "novel deep learning applications". We hav to doubt the claims of this paper, which we only skimmed.

### If you are interested in anomaly detection in time series...

#### A review on outlier/anomaly detection in time series data

ANE BLÁZQUEZ-GARCÍA and ANGEL CONDE, Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BRTA), Spain

USUE MORI, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain

JOSE A. LOZANO, Intelligent Systems Group (ISG), Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Spain and Basque Center for Applied Mathematics (BCAM), Spain

Recent advances in technology have brought major breakthroughs in data collection, enabling a large amount of data to be gathered over time and thus generating time series. Mining this data has become an important task for researchers and practitioners in the past for years, including the detection of outliers or anomalies that may represent errors or events of interest. This review aims to provide a structured and comprehensive state-of-the-art on outlier detection techniques in the context of time series. To this end, a taxonomy is presented based on the main aspects that characterize on outlier detection technique.

Additional Key Words and Phrases: Outlier detection, anomaly detection, time series, data mining, taxonomy, software

#### 1 INTRODUCTION

Recent advances in technology allow us to collect a large amount of data over time in diverse research areas. Observations that have been recorded in an orderly fashion and which are correlated in time constitute a time series. Time series data mining aims to extract all meaningful knowledge from this data, and several mining tasks (e.g., classification, classtering, forecasting, and outlier detection) have been considered in the literature [Esling and Agon 2012; Fu 2011; Ratamanihastane et al. 2010).

Outline detection has become a field of interest for many researchers and practitioners and is now one of the main tasks of time series data mining. Outlie detection has been studied in a variety of application domains such as credit card fraud detection, intrusion detection in cybersecurity, or fault diagnosis in industy, in particular, the analysis of outliers in time series data examines anomalous behaviors across time [Copta et al. 2014a]. In the first study on this projec, which was conducted by Forc [1972], two types of outliers in univariate time series were defined type I, which affects a single observation; and type II, which affects both a particular observation and the subsequent observations. This work was first extended to four outlier types [Tasy 1988], and then to the case of multivariate time series 41 al. 2000]. Since then, many definitions of the term outlier and numerous detection methods have been proposed in the literature. However, to this day; there is still no consensus on the terms used [Carreño et al. 2019]; for example, outlier observations are often referred to as anomalies, discordant observations, discords, exceptions, aberrations, surprises, precultarities or contaminants.

Authors' addresses: Ane Blärquer-García, ablarquer@ikerlan.es; Angel Condo, aconde@ikerlan.es, Ikerlan Technology Research Centre, Basque Research and Technology Alliance (BKTA), P. J.M. Arizmendiamieta, Z. Arrastate/Mondragón, 2030, Spain; Usue Mori, usue morig@ehues, Intelligent Systems (Toorogu (SK), Department of Computer Science and Artificial Intelligence, University of the Basque County (UPV/FBIN), Mantel de Lardinaba), 1240

### A. Blazquez-Garcia et al. ACM Computing Survey (2021) [24]

#### Dive into Time-Series Anomaly Detection: A Decade Review

PAUL BONIOL, Inria, DI ENS, PSL, CNRS, France QINGHUA LIU, The Ohio State University, USA MINGYI HUANG, The Ohio State University, USA THEMIS PALPANAS, Université Paris Cité; IUF, France

JOHN PAPARRIZOS, The Ohio State University, USA

Recent advances in data collection technology, accompanied by the ever-rising volumes and velocity of atreaming data, undercore the vital need for time series analytics. In this regard, time-series anomaly detection has been an important activity, entailing various applications in fields wolt as velyer security. financial markets, law enforcement, and bealth care. While relational literature on anomaly detection is centered on attaistical measures, the increasing number of machine learning algorithms in nevent years call for a structured, general characterization of the research methods for time-series anomaly detection. This survey groups and summarizes anomaly detection is existing solutions under a process-centric taxosomy in the time series context. In addition for logicing an original categorization of anomaly detection methods, we also perform a meta-analysis of the literature and outline general trends in time-series anomaly detection research.

#### ACM Reference Format

Paul Bonio, Qinghua Liu, Mingri Huang, Themis Palqunas, and John Paparitos. 2024. Dive into Time-Series Anomaly Detection A Decade Review. In Proceedings of Made sure to enter the correct onference till from your rights confirmation emai (Conference acronym '20, ACM, New York, NY, USA, S1, pages. https://doi.org/XXXX00XXXXXXXXX

#### 1 Introduction

A wide range of cost-effective sensing, networking, storage, and processing solutions enable the collection of enromous amounts of measurements over time [109–111, 122, 137, 138, 141, 143, 179, 181, 186]. Recording these measurements results in an ordered sequence of rel-levaled data points commonly referred to a time series. More generic terms, such as *data series* or *data sequences*, have also been used to refer to cases where the ordering of data relies on a dimension other than time (e.g., the angle in data from astronomy, the mass in data from spectrometry, or the position in data from biology [176]. Analytical tasks over time series data are necessary virtually in every scientific discipline and their corresponding industries [14.c.], 6, 78, 161, 132, 190–192, 201], including in astronomy [4, 102, 245], biology [11–13, ed], economics [35, 74, 148, 155, 213, 221, 240], energy sciences [6, 9, 158], engineering [112, 162, 203, 243, 248], environmental sciences [77, 84, 100, 101, 164, 207, 247], medicine [22, 199, 200], neuroscience [11, 116], and social sciences [35, 160]. The analysis of time series has become increasingly prevalent for understanding a multitude of natural or human-made processes [187, 188]. Unfortunately, inherent complexities in the data generation of these

P. Boniol et al. Arxiv (2025) [28]

### Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress

Renjie Wu and Eamonn J. Keogh

Abstract—Time series anomaly detection has been a perennially important topic in data science, with papers dating back to the 1956s. However, in creativy sars there has been an explosion of interest in this lock to much of it driven by the scuces of deep learning in other domains and for other time series tasks. Most of these papers test on ore or more of a handlu of popular benchmark datasets, created by Yahoo, Numenta, NASA, etc. In this work is emake as surprising diam. The unginity of the individual exemptions in these datasets suffer from one or more of four flaws. Because of these four flaws, we believe that many published comparisons of anomaly detection algorithms may be unreliable, and more importantly, much of the apparent progress in recert years may be illusionary. In addition to demonstrating these dams, with this paper we introduce the UCR Time Series Anomaly Achev. We believe that his resource wall perform a saintar or lose as th UCR Time Series Classification Archive, the providing the community with a benchmark that allows meaningful appe of overall progress.

----- + -------

Index Terms—Anomaly detection, benchmark datasets, deep learning, time series analysis

#### 1 INTRODUCTION

TIME series anomaly detection has been a perennially neural networks, and a variational auto-encoder (VAE) over L important topic in data science, with papers dating back to the dawn of computer science [1]. However, in the "moving parts", and indeed, the dozen or so explicitly last five years there has been an explosion of interest in listed parameters include: convolution filter, activation, this topic, with at least one or two papers on the topic kernel size, strides, padding, LSTM input size, dense intins topic, with at least one or two papers on the topic series size, strices, padaing, L51M imput size, derise in-appearing each year in virtually every database, data put size, softmak loss function, window size, L82mming rate mining and machine learning conference, including and batch size. All of this is to demonstrate "accuracy ac-SIGKDD [2], [3], ICDE, SIGMOD, VLDB, etc. Alarge fraction of this increase in interest seems to be benchmark datasets)." However, as we will show, much of largely driven by researchers anxious to transfer the con- the results of this complex approach can be duplicated siderable success of deep learning in other domains and with a single line of code and a few minutes of effort. from other time series tasks such as classification. This "one-line-of-code" argument is so unusual that it Most of these papers test on one or more of a handful is worth previewing it before we formally demonstrate it of popular benchmark datasets, created by Yahoo [5], in Section 2.2 below. Almost daily, the popular press Numenta [6], NASA [2] or Pfei Jah (OMNI) [3], etc. in vanits a new achievement of deep learning. Picking one this work we make a surprising claim. The majority of the at random, in a recent paper [8], we learn that deep learnindividual exemplars in these datasets suffer from one or ing can be used to classify mosquitos' species. In particumore of four flaws. These flaws are triviality, unrealistic lar, the proposed algorithm had an accuracy of 97.8% anomaly density, mislabeled ground truth and run-to-failure when distinguishing Aedes vexans from Culex triaeniorhynbias. Because of these four flaws, we believe that most drus. Should we be impressed? One of the current authors published comparisons of anomaly detection algorithms (Keogh) has significant computational experience workmay be unreliable. More importantly, we believe that ing with mosquitos, and he is impressed much of the apparent progress in recent years may be Suppose however that someone downloaded the origi-



### The Elephant in the Room: Towards A Reliable Time-Series Anomaly Detection Benchmark

Qinghua Liu and John Paparrizos Department of Computer Science and Engineering The Ohio State University {liu.11085,paparrizos.l}@osu.edu

#### Abstract

Time-series anomaly detection is a fundamental task across scientific fields and industries. However, the field has long faced the "elephant in the room:" critical issues including flawed datasets, biased evaluation measures, and inconsistent benchmarking practices that have remained largely ignored and unaddressed. We introduce the TSB-AD to systematically tackle these issues in the following three aspects: (i) Dataset Integrity: with 1070 high-quality time series from a diverse collection of 40 datasets (doubling the size of the largest collection and four times the number of existing curated datasets), we provide the first large-scale, heterogeneous, meticulously curated dataset that combines the effort of human perception and model interpretation; (ii) Measure Reliability: by revealing issues and bi-ases in evaluation measures, we identify the most reliable and accurate measure, namely, VUS-PR for anomaly detection in time series to address concerns from the community; and (iii) Comprehensive Benchmarking: with a broad spectrum of 40 detection algorithms, from statistical methods to the latest foundation models, we perform a comprehensive evaluation that includes a thorough hyperparameter tuning and a unified setup for a fair and reproducible comparison. Our findings challenge the conventional wisdom regarding the superiority of advanced neural network architectures, revealing that simpler architectures and statistical methods often yield better performance. The promising performance of neural networks on multivariate cases and foundation models on point anomalies highlights the need for further advancements in these methods. We open-source the benchmark



38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

KDD 2025 | 03/08/2025 | 267

## Reference

[1] N. Laptev, S. Amizadeh, and Y. Billawala. 2015. S5 - A Labeled Anomaly Detection Dataset, version 1.0(16M).

[2] Markus Thill, Wolfgang Konen, and Thomas Bäck. 2020. MGAB: The Mackey-Glass Anomaly Benchmark.

[3] Pawel Benecki, Szymon Piechaczek, Daniel Kostrzewa, and Jakub Nalepa. 2021. Detecting Anomalies in Spacecraft Telemetry Using Evolutionary Thresholding and LSTMs. In Proceedings of the Genetic and Evolutionary Computation Conference Companion (Lille, France) (GECCO '21)

[4] Scott David Greenwald. 1990. Improved detection and classication of arrhythmias in noise-corrupted electrocardiograms using contextual information. Thesis. Massachusetts Institute of Technology.

[5] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. Proc. VLDB Endow. 15, 9 (May 2022), 1779–1797.

[6] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn J. Keogh. 2016. Matrix Prole I: All Pairs Similarity Joins for Time Series. In ICDM.

[7] Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh. 2016. Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins. In Proceedings of the International Conference on Data Mining (ICDM), 739–748.

[8] Yue Lu, Renjie Wu, Abdullah Mueen, Maria A. Zuluaga, and Eamonn Keogh. 2022. Matrix Profile XXIV: Scaling Time Series Anomaly Detection to Trillions of Datapoints and Ultra-fast Arriving Data Streams. In Proceedings of the 28th ACM SIGKDD.

[9] C. -C. M. Yeh, N. Kavantzas and E. Keogh, Matrix Profile VI: Meaningful Multidimensional Motif Discovery, 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 2017, pp. 565-574, doi: 10.1109/ICDM.2017.66. Data Mining (KDD '22).

[10] Paul Boniol, Michele Linardi, Federico Roncallo, Themis Palpanas, Mohammed Meftah, and Emmanuel Remy. 2021. Unsupervised and scalable subsequence anomaly detection in large data series. The VLDB Journal 30, 6 (Nov 2021), 909–931.

[11] F. T. Liu, K. M. Ting and Z. -H. Zhou, Isolation Forest, 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422

[12] Markus Goldstein and Andreas Dengel. 2012. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: poster and demo track 9 (2012).

[13] Paul Boniol and Themis Palpanas. 2020. Series2Graph: graph-based subsequence anomaly detection for time series. Proc. VLDB Endow. 13, 12 (August 2020), 1821–1834.

[14] Ali Abdul-Aziz, Mark R Woike, Nikunj C Oza, Bryan L Matthews, and John D lekki. 2012. Rotor health monitoring combining spin tests and data-driven anomaly detection methods. Structural Health Monitoring (2012).

[15] Pankaj Malhotra, Lovekesh Vig, Gautam Shro, and Puneet Agarwal. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. (2015).

[16] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed. 2019. DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. IEEE Access 7 (2019), 1991–2005.

[17] Mayu Sakurada and Takehisa Yairi. 2014. Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction. In Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis (Gold Coast, Australia QLD, Australia) (MLSDA'14).

[18] R. Wu and E. Keogh, Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress in IEEE Transactions on Knowledge & Data Engineering, vol. 35, no. 03, pp. 2421-2429, 2023.

[19] John Paparrizos, Yuhao Kang, Paul Boniol, Ruey S. Tsay, Themis Palpanas, and Michael J. Franklin. 2022. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. Proc. VLDB Endow. 15, 8 (April 2022), 1697–1711.

[20] Tom Fawcett. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 8 (2006), 861-874.

## Reference

[21] Jesse Davis and Mark Goadrich. 2006. The Relationship between Precision-Recall and ROC Curves. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06).

[22] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S. Tsay, Aaron Elmore, and Michael J. Franklin. 2022. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. Proc. VLDB Endow. 15, 11 (July 2022), 2774–2787.

[23] Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. 2018. Precision and Recall for Time Series. In Advances in Neural Information Processing Systems, Vol. 31.

[24] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. 2021. A Review on Outlier/Anomaly Detection in Time Series Data. ACM Comput. Surv. 54, 3, Article 56 (April 2022), 33 pages.

[25] Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J. Franklin. 2021. SAND: streaming subsequence anomaly detection. Proc. VLDB Endow. 14, 10 (June 2021), 1717–1729.

[26] Schneider, J., Wenig, P. & Papenbrock, T. Distributed detection of sequential anomalies in univariate time series. The VLDB Journal 30, 579–602 (2021).

[27] Liu, Q. and Paparrizos, J., 2024. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. Advances in Neural Information Processing Systems, 37, pp.108231-108261.

[28] Boniol, P., Liu, Q., Huang, M., Palpanas, T. and Paparrizos, J., 2024. Dive into time-series anomaly detection: A decade review. arXiv preprint arXiv:2412.20512.

[29] Maroua Bahri, Flavia Salutari, Andrian Putina, and Mauro Sozio: AutoML: state of the art with a focus on anomaly detection, challenges, and research directions. International Journal of Data Science and Analytics 14(2): 113-126 (2022).

[30] Mononito Goswami, Cristian Challu, Laurent Callot, Lenon Minorics, Andrey Kan. 2023. Unsupervised Model Selection for Time-series Anomaly Detection. In Proceedings of the International Conference on Learning Representations.

[31] Emmanouil Sylligardos, Paul Boniol, John Paparrizos, Panos Trahanias, Themis Palpanas. 2023. Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. Proceedings of the VLDB Endowment 16(11): 3418-3432.

[32] Lin Xu, Frank Hutter, Holger H Hoos, Kevin Leyton-Brown. 2008. SATzilla: portfolio-based algorithm selection for SAT. Journal of Artificial Intelligence Research 32: 565-606.

[33] Lei Cao, Yizhou Yan, Yu Wang, Samuel Madden, Elke A Rundensteiner. 2023. Autood: Automatic outlier detection. Proceedings of the ACM on Management of Data, 1(1): 1-27. ACM, New York, NY, USA.

[34] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[35] Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023. Large models for time series and spatio-temporal data: A survey and outlook. arXiv preprint arXiv:2310.10196.

[36] Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, Qingsong Wen. 2024. Position: What Can Large Language Models Tell Us about Time Series Analysis. In Proceedings of the Forty-first International Conference on Machine Learning.

[37] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2023. One fits all: Power general time series analysis by pretrained Im. Advances in Neural Information Processing Systems 36: 43322-43355.

[38] Siwon Kim, Kukjin Choi, Hyun-Soo Choi, Byunghan Lee, Sungroh Yoon. 2022. Towards a rigorous evaluation of time-series anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(7): 7194-7201.
[39] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, Artur Dubrawski. 2024. MOMENT: A Family of Open Time-series Foundation Models. In Proceedings of the International Conference on Machine Learning.
[40] Daochen Zha, Kwei-Herng Lai, Mingyang Wan, Xia Hu. 2020. Meta-AAD: Active anomaly detection with deep reinforcement learning. In Proceedings of the 2020 IEEE ICDM, 771-780. IEEE.

[41] Liu, Q., Lee, S. and Paparrizos, J., 2025. TSB-AutoAD: Towards Automated Solutions for Time-Series Anomaly Detection. Proceedings of the VLDB Endowment (VLDB 2025).

[42] Kong, Y., Yang, Y., Hwang, Y., Du, W., Zohren, S., Wang, Z., ... & Wen, Q. (2025). Time-mqa: Time series multi-task question answering with context enhancement. arXiv preprint arXiv:2503.01875.

# Thank you for attending!

## Any Questions?