

SUPPLEMENTARY MATERIAL

This is the supplementary material of Future of Time Series Anomaly Detection: From Historical Review to Unified Outlook. Table S1 lists the common abbreviations used throughout the paper.

TABLE S1: List of Common Abbreviations

Abbr.	Expansion	Abbr.	Expansion
TSAD	Time series anomaly detection	Prot	Prototype-based
SSM	State space model	Diff	Diffusion-based
LLM	Large language model	CC	Channel correlation
RNN	Recurrent neural network	Rec	Reconstruction-based
GNN	Graph neural network	Fcst	Forecasting-based
TCN	Temporal convolutional network	Gen	Generative-based
AE	Autoencoder	Sim	Similarity-based
VAE	Variational autoencoder	Cls	Classification-based
FFN	Feed-forward neural network	Stat	Statistical-based

A. RELATED WORKS

To analyze the research trends of different taxonomies of paradigms and architectures in current TSAD, we used Google Scholar¹ to search for various types of algorithms. We used a different set of search queries for each category, which can be found on our project website: <https://emorzzlg.github.io/FTSAD/>. We recorded the number of papers published each year, covering a ten-year period from 2015 to 2025, with data collected up to June 24, 2025. Fig. 1 (in the main text) shows the number of publications for each algorithm type per year, and the raw data is available in Table S2. The top sub-figure is a statistical breakdown based on the taxonomy of different paradigms, while the bottom sub-figure is a breakdown based on different algorithm architectures. The architecture categories shown in the figure correspond to the detailed taxonomy of paradigms presented in Table III (in the main text).

Based on our analysis of the figure, we summarize the following observations:

- 1) The publication trends for the **Forecasting** and **Statistical** paradigms have been gradually slowing down since 2022.
- 2) The publication trends for the **Generative**, **Similarity-based**, and **Reconstruction** paradigms have remained consistent, indicating that they are still mainstream research paradigms in current TSAD studies.
- 3) Publications related to the **LLM-based** paradigm have been increasing since 2024. As of the data cutoff date, the publication count for 2025 is approximately 2,200, which is already 84.6% of the total for 2024. This represents the fastest growth rate in publication count among all paradigms, suggesting that this paradigm is receiving relatively widespread research attention.
- 4) The publication trends for different algorithm architectures are similar, which indicates that algorithms across these architectures are equally popular.

¹<https://scholar.google.com/>

TABLE S2: Number of papers published each year for various algorithm types.

Paradigm/Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Reconstruction	10300	11100	11500	13000	14700	15200	19500	22700	25500	28200	13300
Forecasting	31000	33500	35100	40600	46300	52900	59700	64300	62300	60800	36100
Generative	8870	9640	9280	11500	13400	15600	19200	23200	24800	32000	16300
Similarity-based	22270	24300	26100	29300	33400	38100	44900	52000	57700	58800	32400
LLM-based	0	0	2	3	3	6	4	44	480	2600	2200
Statistical	28700	30600	32800	37500	43000	49400	57100	57500	52100	46800	30900
Architecture/Year	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Unified & LLM-based Models	98	126	137	193	284	366	539	882	1640	6230	3530
Diffusion & Transformer	436	462	1140	1400	1880	2530	4670	7970	13800	22100	13200
State Space Model	847	921	902	1080	1130	1250	1510	1680	1800	2400	1300
RNN, GNN, TCN & Prototype	3630	4280	5100	7200	9880	13500	17900	22600	29100	31900	17000
Classical Machine Learning	5390	6670	8480	11600	15600	19700	25400	33200	39800	42600	25300

B. ALGORITHM INNOVATIONS

Algorithmic innovations in time series anomaly detection advance along two orthogonal dimensions: (i) *detection paradigms and scoring objectives*, which govern the theoretical definition of abnormal behavior (detailed in SM B-B), and (ii) *representation architectures*, which govern how deep feature extractors capture temporal dynamics and inter-variable dependencies (detailed in SM B-C). In addition, emerging cross-domain unified models integrate multi-dataset pre-training strategies (detailed in SM B-D). Table S4 summarizes these complementary perspectives along with key literature references.

A. Paradigm Classification Criteria

To substantiate the Taxonomy of Paradigms introduced in Sec. IV, this section outlines the three-tier classification framework and provides formal disambiguation criteria for algorithms exhibiting overlapping characteristics across paradigm boundaries.

Algorithm classification follows a hierarchical three-tier decision protocol applied in sequence:

Tier 1: LLM Priority Rule. The emergence of foundational models represents a paradigm shift from learning task-specific distributions from scratch to leveraging pre-trained cross-domain knowledge for zero-shot generalization. The primary criterion therefore evaluates the origin of the representation capacity. Algorithms whose anomaly detection capability fundamentally relies on pre-trained, cross-domain knowledge transferred from large language models or foundational models are assigned to the **LLM-based Paradigm**, irrespective of whether their downstream output format uses forecasting error, reconstruction, or embedding similarity.

Tier 2: Core Detection Assumption. For algorithms trained strictly on domain-specific data, categorization is governed by their fundamental assumption regarding how abnormal patterns deviate from normal temporal behavior:

- *Reconstruction Paradigm:* Assumes normal series can be losslessly compressed and deterministically reconstructed; anomalies yield elevated point-wise residual errors in input space.
- *Forecasting Paradigm:* Posits that normal sequences follow predictable temporal progression; anomalies cause discrepancies between predicted horizons and actual observations.
- *Generative Paradigm:* Models normal data as a probability distribution via stochastic neural objectives (e.g., ELBO, adversarial loss, or score matching); anomalies correspond to low-density samples or degraded reverse diffusion trajectories.

TABLE S3: Full Overview of Time Series Anomaly Detection Surveys.

Citation	Application	Taxonomy of Paradigms	Anomaly Taxonomy	Novel Algorithms		Library	Datasets	Metrics	Discussion		
				Prot/SSM/Diff	LLM				Interpretability	Multi-Modality	Multi-Domain
Landauer et al. [1] (2018)	✓							✓			
Zhao et al. [2] (2019)	✓						✓				
Cook et al. [3] (2020)	✓		✓								
Braei and Wagner [4] (2020)							✓	✓			
Blázquez-García et al. [5] (2020)			✓			✓					
Choi et al. [6] (2021)	✓	✓					✓				
Shaukat et al. [7] (2021)			✓								
Lai et al. [8] (2021)			✓			✓	✓				
Garg et al. [9] (2021)		✓				✓	✓	✓			
Zhang et al. [10] (2021)			✓								
Schmidl et al. [11] (2022)		✓	✓				✓	✓			
Wu and Keogh [12] (2022)							✓				
Alimohammadi et al. [13] (2022)		✓	✓								
Kim et al. [14] (2022)								✓			
Belay et al. [15] (2023)		✓	✓						✓		
Kim et al. [16] (2023)	✓										
Wen et al. [17] (2023)											✓
Li and Jung [18] (2023)	✓		✓								
Yan et al. [19] (2023)	✓	✓	✓								
Jin et al. [20] (2024)	✓	✓				✓			✓		✓
Sun et al. [21] (2024)								✓			
Ye et al. [22] (2024)	✓				✓			✓	✓		✓
Su et al. [23] (2024)	✓				✓		✓	✓	✓		
Zamanzadeh Darban et al. [24] (2024)	✓	✓	✓	✓			✓	✓	✓	✓	
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

TABLE S4: Comprehensive summary of algorithmic innovations and their key references.

Dimension	Category	References
Detection Paradigms (SM B-B)	Reconstruction	EncDec-AD [25], MSCRED [26], UAE [27], [28], USAD [29], RE-ADTS [30], NPSR [31]
	Forecasting	ARIMA, FEDformer, Autoformer, PatchTST, DLinear [32]
	Generative	LSTM-VAE [33], [34], InterFusion [35], MUTANT [36], FCVAE [37], FGANomaly [38], [39], [40], TadGAN [41], AnoFormer [42], BeatGAN [43], UAC-AD [44], D3R [45], ImDiffusion [46], [47], DDMT [48], TimeADMM [49], DDTAD [50]
	Similarity-based	KNN [51], DTW [52], [53], [54], PSOM [55], LOF [56], HBOS [57], ECOAD [58], [59], OC-TCN [60], COCA [61], COUTA [62], THOC [63], PatchAD [64], CARLA [65], SimAD [66], TimeAutoAD [67], USD [68], BroadCAE [69], MEMTO [70], AMSL [71], AdaMemBLS [72], OCSVM [73], EB-OCSVM [74], [75], [76], Isolation Forest [77]
	Statistical	Z-Scores, 3-Sigma, IQR
Representation Architectures (SM B-C)	LLM-based	TS-Bert [78], GPT4TS [79], aLLM4TS [80], Timer [81], AnomalyLLM [82], LLMAD [83], SIGLLM [84], ParInfoGPT [85], Holmes-VAD [86], MERIT [87], [88], [89]
	RNN Backbone	OmniAnomaly [90], CAE-AD [91], MSDMM [92], LSGA [93]
	GNN Backbone	[94], [95], STADN [9], DGINet [96], DPGLAD [97], SGTrans [98]
	TCN Backbone	[99], [100], [101], [102], TCN-AE [103], WaRTEm-AD [104], [105], TFAD [106]
	State Space Model	S-Mamba [107], [108], [109], DMamba [110], C-Mamba [111], Mambaformer [112], SGFM [113], TSAD-C [114]
	Transformer Backbone	Anomaly Transformer [115], TransAnomaly [116], CauseFormer [102], TSAT [117], TransAD [118], AnomalyBERT [119], DCDetector [120], D3TN [121], [122]
	Others	[123], [124], [125]
Unified TSAD Models (SM B-D)	Cross-Domain Models	DADA [126], DACAD [127], ContextTDA [128], FS-ADAPT [129], VANDA [130], UNITS [131]

- *Similarity-based Paradigm*: Formulates normality as geometric proximity or cluster compactness; anomalies are detected via metric distances, neighborhood densities, or boundary isolation.
- *Statistical Paradigm*: Presupposes an analytically specified parametric distribution with parameters estimated in closed form without deep neural optimization; anomalies deviate from this theoretical model.

Tier 3: Dominant Mechanism Rule. For composite architectures combining multi-objective losses, taxonomy assignment is determined by the dominant scoring signal producing the anomaly metric at inference time. Underlying network backbones (such as Graph Neural Networks or Transformer encoders) are treated as representational media rather than taxonomic discriminators. When two scoring signals contribute with comparable significance, the assignment to a single paradigm serves as an analytical abstraction, and the hybrid nature is explicitly documented.

To resolve potential ambiguities in boundary cases, Ta-

ble S5 summarizes the critical mathematical and operational distinctions between paradigm pairs with overlapping surface characteristics. Specifically, Variational Autoencoders and Diffusion models are classified under the Generative Paradigm due to their stochastic, distribution-matching training objectives rather than deterministic reconstruction error. Foundation forecasters (e.g., GPT4TS [79]) are assigned to the LLM-based Paradigm by Tier 1, as their representational power stems from pre-trained weights rather than sequence modeling from scratch. Non-parametric density estimators (e.g., LOF [56]) belong to the Similarity-based Paradigm because they compute neighborhood geometric ratios without closed-form parametric assumptions. Finally, multi-task models combining reconstruction and forecasting (e.g., certain spatial-temporal GNNs [94, 95]) are categorized by their dominant inference scoring mechanism under Tier 3.

TABLE S5: Distinguishing criteria for paradigm pairs with overlapping surface characteristics.

Paradigm Comparison	First Paradigm Characteristics	Second Paradigm Characteristics
Reconstruction vs. Generative	Deterministic optimization (e.g., ℓ_1/ℓ_2 reconstruction loss); latent representation is a fixed point z ; does not support probabilistic sampling	Probabilistic or distribution-matching objective (e.g., ELBO, adversarial loss, score matching); latent space is stochastic $q(z x)$; supports generative synthesis
Generative vs. Statistical	Distribution implicitly parameterized by neural network weights; learned via iterative gradient optimization; high-capacity nonlinear manifold	Distribution specified by an analytical closed-form density (e.g., $\mathcal{N}(\mu, \Sigma)$); parameters estimated via exact MLE or method of moments; requires no deep neural training
Similarity (Density) vs. Statistical	Non-parametric local density estimation; makes no prior assumption on distributional form; score derived from neighborhood spatial geometry (e.g., reachability density in LOF)	Parametric density estimation; presupposes a specific theoretical distribution; anomaly score is the log-likelihood or tail probability under the fitted analytical model
Reconstruction vs. Forecasting	Autoassociative mapping of the current observation window onto itself: $\hat{t}_{(i-m+1):i}$	Autoregressive or sequence-to-sequence mapping from historical context to future horizons: $\hat{t}_{(i-m+1):i} \rightarrow \hat{t}_{(i+1):(i+p)}$
LLM-based vs. Forecasting	Core representation power originates from pre-trained foundational models; supports cross-domain transfer and zero-shot inference; scoring format is an output interface	Representations and temporal dynamics optimized from scratch exclusively on the target time series; anomaly score is the prediction residual; lacks pre-trained foundational backbone
Similarity (Contrastive) vs. Reconstruction	Anomaly score is evaluated as a metric distance, projection radius, or dissimilarity in the latent embedding space	Anomaly score is evaluated directly as residual error between observed and reconstructed signals in the input data space

B. Taxonomy of Paradigms and Methods

This section introduces the taxonomy of TSAD methods categorized by their fundamental algorithmic paradigms. Each paradigm is defined by its core theoretical objective, dictating

how anomalies are conceptualized and distinguished from normal temporal patterns.

1) *Reconstruction Paradigm*: The reconstruction paradigm focuses on learning the intrinsic representations of normal time series data. Anomalies are identified by calculating the reconstruction error, operating under the assumption that abnormal data will deviate from learned normal patterns and thus yield significantly higher errors.

a) *Autoencoder-based Model*: Autoencoders have become a prominent method for detecting anomalies in time series data without labels by learning compressed representations of normal patterns, facilitating dimensionality reduction and feature extraction. Early explorations include **EncDec-AD** [25], which employs a simple LSTM encoder-decoder scheme. Building on this, **MSCRED** [26] enhances anomaly detection by utilizing system feature matrices for global state representation and a convolutional encoder for sensor correlation extraction, along with an attention-based ConvLSTM to capture cross-temporal information across channels.

Further advancements are seen in the univariate fully connected network (**UAE**) proposed by Garg et al. [27], which introduces dynamic scoring functions that outperform static ones in multivariate settings, particularly when modeling channels separately.

The unsupervised **G-LSTM-AE** model, developed by Hu and Xia [28], combines Gaussian confidence intervals with LSTM for rapid detection in industrial signals, meeting real-time requirements. Additionally, **USAD** [29] features a dual-decoder network structure, employing two-stage training to optimize performance.

Novel approaches, such as **RE-ADTS** [30], use autoregressive models for optimal window width determination, while the **NPSR** algorithm [31] introduces a unique modeling technique that combines point-based and sequence-based reconstruction to quantify anomalies, incorporating a Nominality Score to derive an Induced Anomaly Score. Collectively, these methods illustrate the diverse strategies employed in leveraging autoencoders for effective anomaly detection in time series data.

Additionally, usually the variational autoencoder can be regarded as a special case based on the autoencoder. Therefore, this part is introduced in SM B-B3a.

2) *Forecasting Paradigm*: The forecasting paradigm operates by modeling the temporal evolution of normal time series sequences. Under the assumption that normal data follow predictable temporal dynamics while anomalies cause unexpected future deviations, forecasting models detect anomalies by computing the discrepancy between predicted values and actual observations.

Historically, statistical methods such as Autoregressive Integrated Moving Average (ARIMA) predict future points by modeling linear temporal correlations, using the prediction residual as the anomaly score. In modern TSAD, general deep forecasting architectures (e.g., Autoformer, FEDformer, PatchTST, and DLinear) can naturally serve as anomaly detectors out-of-the-box by leveraging their capacity to capture long-range trends, seasonal patterns, and channel dependencies. Because most time series forecasting models can be directly

adapted for anomaly scoring, we refer readers to dedicated forecasting surveys [32] for comprehensive discussions of specific forecasting backbones and architectures.

3) *Generative Paradigm*: The generative paradigm captures the underlying true distribution of normal time series data. By simulating the data generation process, it differentiates between normal and abnormal patterns based on the likelihood of the observed sequences.

a) *Variational Autoencoder Model*: The primary advantage of Variational Autoencoders (VAEs) in TSAD lies in their variational inference mechanism, which aligns generated samples with observed data by optimizing the parameters of the latent space distribution. This framework enhances the model's understanding of data structure, providing a robust foundation for anomaly detection by balancing reconstruction error and latent distribution consistency.

A notable implementation is **LSTM-VAE** [33], which combines LSTM with VAE to dynamically address task execution changes through state-based threshold selection.

In another approach, Pham et al. [34] utilize convolutional kernels of varying sizes to capture both long- and short-scale information, learning hidden space representations through variational inference. **InterFusion** [35] introduces a Hierarchical VAE that explores a low-dimensional two-view embedding strategy for time series data, employing MCMC-based imputation for improved accuracy in embeddings.

MUTANT [36] constructs feature maps for each time window, leveraging a graph neural network to learn variable embeddings, and incorporates a reconstruction model with attention mechanisms, LSTM, and a VAE module to assess variable importance. Lastly, **FCVAE** [37] integrates Global and Local Frequency Information Extraction modules, using frequency features as conditional factors within the CVAE framework and enhancing the capture of short-term trends through an attention mechanism. Together, these approaches illustrate the versatility and effectiveness of VAEs in addressing the complexities of anomaly detection in time series data.

b) *Generative Adversarial Model*: Generative Adversarial Network (GAN) have emerged as a powerful tool in TSAD by training a generator to replicate normal data patterns while a discriminator learns to differentiate between generated data and actual anomalies. This adversarial mechanism enhances the understanding of normal patterns, thereby improving anomaly detection in dynamic time series. Several approaches [43, 38, 39, 40] have been developed within this framework.

FGANomaly [38] utilizes both reconstructed and original time series data as inputs to its Filter module and discriminator, incorporating a progressive Adaptive Weighted Label function to enhance optimization during training.

In a different approach, Niu et al. [39] combined LSTM networks with VAE to learn the deep probability distribution of time series, applying GAN principles to assess whether real-time data fits the normal distribution using both reconstruction and discriminator errors.

The integration of advanced architectures is also evident. Zeng et al. [40] introduced a GAN with Transformer architecture, leveraging the discriminator to capture short-term trends, which amplifies anomaly scores. **TadGAN** [41] employs

LSTM at its core, combining reconstruction loss with three auxiliary functions to refine anomaly discrimination, while Cycle Consistency Loss is included to prevent model collapse.

AnoFormer [42] enhances normal and anomalous data modeling through a two-step masking method, adjusting decision boundaries by calculating uncertainty using entropy. **BeatGAN** [43] merges CNN and RNN structures, optimizing parameters by minimizing reconstruction and discriminator errors, and employs Extreme Value Theory to generate thresholds for anomaly detection without making assumptions about data distribution.

UAC-AD [44] presents a GAN-based method for multimodal data, utilizing the mixup technique to create mismatched samples, thereby improving difficult sample recognition and anomaly detection accuracy through contrastive learning, while optimizing the discriminator via adversarial learning.

c) Diffusion-based Model: Diffusion models simulate a physical process in which data points transition from an ordered state (normal data) to a disordered state (anomalous data). In the context of TSAD, these models are trained to replicate the generation process of normal data. After training, anomalies can be detected by analyzing the model's behavior during the generation phase. This approach offers significant flexibility and controllability over the data generation process, making it particularly effective for managing complex time series datasets.

A notable contribution in this area is **D3R** [45], which tackles the instability of time series data in non-stationary environments, a challenge that often results in a high false positive rate. D3R employs a decomposition and reconstruction method to manage data drift, integrating a noise diffusion model that directly recovers contaminated data. This strategy allows for avoiding the substantial retraining costs typically associated with changes in information bottlenecks. Building on this foundation, **ImDiffusion** [46] presents a TSAD model that merges time series imputation techniques with diffusion models. It utilizes a grating data masking strategy to generate missing data points, enhancing the decision boundary between normal and anomalous data. Research indicates that this imputation approach is more effective than traditional prediction and reconstruction methods in improving model performance.

Additionally, Pintilie et al. [47] propose a novel use of diffusion models specifically for TSAD, introducing DiffusionAE, which is a jointly trained autoencoder and diffusion model. This model conditions the diffusion process on the autoencoder's reconstruction, demonstrating that training with reconstructed data results in a more robust model compared to using the original data.

Another significant advancement is **DDMT** [48], which introduces the Adaptive Dynamic Neighbor Mask (ADNM) mechanism to address the Weak Identity Mapping issue in time series reconstruction. ADNM dynamically adjusts the mask based on the differences between the reconstructed and original time series, thereby enhancing the model's ability to understand and fit normal data patterns.

Finally, **TimeADDM** [49] combines an RNN embedding module with the denoising capabilities of diffusion models for TSAD. The process begins by mapping the original time series to a low-dimensional latent space, followed by

introducing noise through a forward diffusion process and reconstructing features via reverse denoising. Anomaly scores are computed based on reconstruction errors across various diffusion steps. Similarly, **DDTAD** [50] utilizes a TCN for pre-embedding time series data, incorporates a one-dimensional UNet diffusion network for data restoration, and subsequently detects anomalies.

4) Similarity-based Paradigm: The similarity-based paradigm identifies anomalies by evaluating metric distances, density profiles, or similarity functions among time series instances. Observations are flagged as anomalies when they deviate significantly from established normal clusters, prototypes, or dense representation manifolds.

a) Distance-based Model: Distance-based approaches quantify outlierness through direct pairwise geometric proximity. K-Nearest Neighbors (**KNN**) [51] computes the distance to the k -th nearest neighbor in feature space, where isolated samples exhibiting large neighbor distances are identified as outliers. Dynamic Time Warping (**DTW**) [52] performs non-linear temporal alignment across sequence pairs using dynamic programming, measuring morphological similarity invariant to local phase shifts and speed variations.

b) Density-based Model: Density-based algorithms evaluate anomalies based on local or global probability density estimation, positing that normal observations reside in high-density regions while anomalies inhabit sparse low-density areas. Local Outlier Factor (**LOF**) [56] assesses the ratio between a sample's local reachability density and those of its k -nearest neighbors, isolating anomalies in datasets characterized by non-uniform density distributions. Histogram-based Outlier Score (**HBOS**) [57] models univariate feature densities using static or dynamic histogram binning, computing anomaly scores under an assumption of feature independence for high computational throughput. The **ECOD** algorithm [58] estimates tail probabilities via Empirical Cumulative Distribution Functions (ECDF) across dimensions independently and aggregates these probabilities into a unified outlier score, eliminating hyperparameter tuning requirements.

c) Clustering-based Model: Clustering-based methods partition the temporal feature space into cohesive groups, flagging instances that either fall outside major cluster boundaries or belong to sparse micro-clusters. Wang and Qin [53] combined K-means with Particle Swarm Optimization (PSO) to discover optimal initial cluster centroids, mitigating sensitivity to initialization and local optima in time series clustering. Li et al. [54] developed an enhanced Fuzzy C-Means (FCM) formulation utilizing PSO to optimize feature weights and evaluating anomaly severity via reconstruction errors derived from optimal cluster centers and fuzzy membership matrices. The Periodic Self-Organizing Map (**PSOM**) [55] aligns topological neurons with intrinsic temporal periodicities, restricting winning neuron competitions within temporal columns and scoring anomalies by assessing local neighborhood disparities.

d) Classification-based Model: Classification-based approaches (also termed one-class boundary learning) formulate unsupervised anomaly detection as identifying a compact geometric boundary enclosing normal observations in transformed feature spaces. Points residing outside this decision boundary

are classified as anomalies.

In classical machine learning, One-Class Support Vector Machines (**OCSVM**) [73] map input temporal features into high-dimensional reproducing kernel Hilbert spaces, constructing maximal-margin hyperplanes separating normal patterns from the origin. **EB-OCSVM** [74] integrates change-point causal graphs with boundary extensions to attenuate noise sensitivity and facilitate root cause localization. Bevilacqua and Tsafaris [75] coupled dictionary learning (DL) with OCSVM, iteratively optimizing dictionary reconstruction errors and assigning adaptive slack penalties. Wang et al. [76] employed adaptive velocity weighting within PSO to automate OCSVM parameter selection.

In deep learning, one-class architectures replace fixed kernel mappings with end-to-end neural feature projections. Mauceri et al. [59] constructed prototype-derived dissimilarity vectors evaluated via one-class nearest neighbors. **OC-TCN** [60] trains Temporal Convolutional Networks using one-class objectives combined with cross-domain transfer learning. **THOC** [63] leverages multi-scale dilated RNNs with skip connections to encapsulate temporal representations within multi-layer hyperspheres. To prevent hypersphere collapse during representation learning, **COCA** [61] combines data augmentation with negative-free contrastive optimization, while **COUTA** [62] calibrates irregular industrial sensor noise through prior-guided uncertainty regularization.

e) Contrastive Learning Model: Contrastive self-supervised learning constructs representations by maximizing mutual information between augmented positive pairs while separating representations of distinct contexts or synthetic anomalies. **PatchAD** [64] implements a multi-scale patch-based MLP-Mixer with dual projection constraints to suppress representation collapse. **CARLA** [65] employs a two-stage regime, synthesizing realistic anomaly instances during pre-training to guide the separation of nominal and perturbed windows. **SimAD** [66] uses a ContrastFusion module and dissimilarity-driven objectives to isolate anomalous patterns. **TimeAutoAD** [67] establishes automated hyperparameter search over self-supervised contrastive objectives under noisy training regimes. **USD** [68] models representations using Gaussian mixture formulations optimized via soft contrastive losses. **BroadCAE** [69] incorporates contrastive objectives into broad learning systems to generate dynamic scenario-adaptive decision thresholds in low-dimensional projection spaces.

f) Prototype-based Model: Prototype and memory-bank mechanisms address the over-generalization pitfall inherent in highly expressive deep neural networks (where models inadvertently reconstruct unseen anomalies with low error). By restricting the latent representation space to convex combinations or nearest-neighbor queries over a discrete set of explicitly learned normal prototypes, prototype-based models establish robust constraints against anomaly reconstruction. **MEMTO** [70] incorporates a gated memory-augmented Transformer alongside a two-phase optimization schedule, scoring anomalies via Latent Space Deviation. **AMSL** [71] integrates multi-strategy augmentations with hierarchical global-to-local memory networks. **AdaMemBLS** [72] couples broad learning system backbones with prototype memory banks, achieving

efficient streaming inference and competitive anomaly discrimination.

g) Tree-based Model: Tree-based anomaly detection isolates anomalous instances through recursive subspace partitioning. The Isolation Forest algorithm [77] constructs ensembles of randomized isolation trees by iteratively selecting features and split points. Because anomalous points reside in low-density or peripheral regions, they require fewer recursive splits to become isolated, manifesting as noticeably shorter average path lengths across tree ensembles.

5) Statistical Paradigm: The statistical paradigm utilizes mathematical frameworks to model the underlying distribution of time series data. Anomalies are identified as data points that fall outside predefined statistical boundaries, such as confidence intervals or standard deviations.

a) Statistical Method: The Z-Scores algorithm detects anomalies in time series data by calculating each point's deviation from the mean, normalized by the standard deviation. The 3-Sigma algorithm, assuming a normal distribution, identifies anomalies when a data point's deviation exceeds three standard deviations from the mean. The Interquartile Range (IQR) method uses quartiles to detect anomalies, computing the difference between the upper and lower quartiles and typically applying 1.5 times the IQR to identify outliers.

6) LLM-based Paradigm: The LLM-based paradigm leverages the extensive pre-trained knowledge and contextual reasoning capabilities of Large Language Models. By adapting these models to time series data, it enables powerful zero-shot detection and enhances the interpretability of anomalies.

a) Predictive LLM-based Model: Recent advancements in TSAD have been significantly influenced by the integration of pre-trained models and large language models (LLMs). **TS-Bert** [78] is a pioneering pre-trained model specifically designed for TSAD. It leverages a substantial volume of unlabeled time series data to learn temporal behavior features, employing pre-training tasks analogous to those used in BERT, such as Masked Language Modeling and Next Sentence Prediction. Following pre-training, TS-Bert integrates the input and output of the TSAD task, allowing for end-to-end fine-tuning of all parameters based on target data. Notably, to alleviate the dependency on labeled data, the model incorporates unsupervised training by generating labels within the training data using self-reconstruction during the fine-tuning stage.

Building on this foundation, **GPT4TS** [79] introduces a large language model tailored for general time series analysis, based on pre-trained architectures like GPT-2. It begins by applying Instance and Norm Patching to the input time series data, followed by linear probing to redesign and retrain the Input Embedding. This projection minimizes the parameter count while adapting the model for downstream tasks, including TSAD. During training, GPT4TS freezes self-attention layers and feedforward networks containing extensive pre-training knowledge, fine-tuning only the positional embeddings and layer normalization, which balances the use of pre-trained knowledge with task-specific adaptations.

Similarly, **aLLM4TS** [80] exemplifies a class of LLMs focused on general time series analysis tasks. This model adapts large language models through a two-stage self-supervised

multi-patch prediction task, enhancing time series representation learning. By utilizing a unique patch-level decoding layer, aLLM4TS improves its sequential representation capabilities for time patches, demonstrating robust performance across various downstream tasks, including TSAD.

Timer [81] further advances this space by presenting a class of large language models designed for time series analysis, based on a decoder-only Transformer architecture that generates the next token autoregressively. Timer aggregates and cleans public time series datasets to construct a high-quality, large-scale Unified Time Series Dataset, which encompasses heterogeneous domains and hierarchical capabilities. Pre-training on this diverse data, Timer reformats it into a unified single-series sequence format, promoting channel independence. For TSAD tasks, Timer frames the problem as predicting the next token based on prediction errors, achieving robust anomaly detection.

b) Exploring LLM-Powered Model: Meanwhile, several innovative TSAD models leverage knowledge distillation and LLMs to enhance detection capabilities. **AnomalyLLM** [82] employs a knowledge distillation approach to identify temporal anomalies by training a student network to replicate the behavior of a teacher network derived from a pre-trained LLM, specifically GPT-2. This model addresses the challenge of limited access to large-scale time series datasets by capitalizing on the representation differences between the teacher and student networks. By fine-tuning the positional embedding layer and layer normalization of the student network, AnomalyLLM improves its temporal representation. To mitigate overfitting to abnormal features, it incorporates prototype signals and utilizes data augmentation techniques to synthesize anomalies, thereby increasing the representation gap between the two networks.

LLMAD [83] introduces an LLM framework aimed at enhancing the interpretability of TSAD. This model utilizes the In-Context Learning mechanism to retrieve the top-k most similar normal and abnormal samples from a pre-constructed database. By incorporating this information into the prompts sent to the language model, LLMAD establishes clear boundaries between normal and abnormal patterns. Additionally, it enhances both performance and interpretability through Domain Knowledge Injection and Expert Step-Wise Inference via the Anomaly Contextualization technique. LLMAD ultimately generates natural language reports that aid decision-making by providing explanations of detected anomalies, classifying anomaly types, and determining appropriate alarm levels.

Further advancing the field, **SIGLLM** [84] presents a zero-shot large language model framework tailored for TSAD, built on architectures like GPT-3.5 and MISTRAL. This framework preprocesses raw time series data into a compatible format using techniques such as Scaling, Quantization, Rolling windows, and Tokenization. It features two detection branches: **PROMPTER**, which merges input prompts with time series data to assist the model in identifying anomalous indices, and **DETECTOR**, which leverages the model's predictive capabilities to detect anomalies based on prediction errors. Preliminary results indicate that DETECTOR generally outperforms PROMPTER in detection accuracy, while PROMPTER offers greater cost efficiency.

Lastly, **ParInfoGPT** [85] is a two-stage TSAD model that

assesses the reliability of rotating machinery using partial information. In the first stage, the model performs self-supervised reconstruction on a substantial corpus of unlabeled time series data to extract key features. It employs an information masking strategy based on mutual information to identify critical patches for zeroing, thereby capturing essential temporal information. The data undergoes processing through a GPT with embedding layers and a GPT encoder for feature extraction, followed by reconstruction using a feed-forward neural network (FFN). In the second stage, ParInfoGPT conducts Weakly Supervised Classification utilizing a small labeled dataset. By initializing the GPT encoder with parameters from the first stage, it extracts representative features and fine-tunes classification to infer fault types based on the learned feature representations.

C. Deep Representation Architectures

The aforementioned paradigms establish the theoretical objectives and optimization targets for time series anomaly detection. To instantiate these paradigms on complex, multi-dimensional time series data, robust feature extractors are indispensable. This section summarizes the core deep representation architectures that serve as the fundamental backbones across different paradigms.

1) Recurrent Neural Network Model: Recurrent Neural Networks (RNNs) leverage their recurrent connections to capture long-term dependencies in time series data, making them highly effective for learning complex temporal patterns and detecting anomalies. Their ability to maintain temporal memory allows RNNs to handle sequences with significant temporal dynamics, which is crucial for TSAD.

For instance, **OmniAnomaly** [90] employs a stochastic RNN to explicitly explore temporal dependencies among random variables, utilizing Stochastic Gradient Variational Bayes and planar Normalizing Flows to estimate the posterior distribution of non-Gaussian variables. Similarly, **CAE-AD** [91], an LSTM-based algorithm, features a siamese structure that integrates self-attention for positional information extraction, alongside Discrete Fourier Transform-based data augmentation and Instance and Contextual Contrasting losses to capture local invariant features.

MSDMM [92] utilizes a GRU network to extract time series features, enhancing normal window reconstruction by retrieving global normal features through a memory module and constructing local memory from neighboring windows. **LPGA** [93] focuses on learning a sparse adjacency matrix between sensors via node embedding similarity and Gumbel-Softmax sampling, minimizing sensor relationship density while employing CNNs for short-term local pattern capture and Skip-GRU for long-term trend analysis.

To mitigate the issue of excessive temporal distribution overlap in RNN models across different time stages, Sun et al. [132] proposed the Ranking-Based Cross-Entropy Loss. This loss function jointly learns class features and the priority of temporal positions, thereby assisting the classifier in generating probability scores with clearer boundaries between distinct phases.

2) *Temporal Convolution Network Model*: Temporal Convolution Networks (TCNs) are specialized deep learning architectures designed for efficient feature extraction in time series data, utilizing convolution operations to capture local patterns and temporal dependencies.

The foundational work by Bai et al. [99] proposed the basic TCN, enhancing traditional CNN networks through one-dimensional convolutions to better learn temporal information. Building upon this, Park et al. [100] successfully applied TCN to various applications, including ECG anomaly detection and gesture recognition.

In a different approach, Ren et al. [101] introduced the Spectral Residual (SR) combined with a CNN, termed SR-CNN, which extracts saliency features within time series data. This model employs tail interpolation to actively identify anomalies, facilitating integration with online monitoring systems for rapid diagnostics.

DKN [133] constructs a dual-tower distillation network architecture, based on convolutional and attention networks. This design aims to extract both local and global features while enhancing information flow across different layers. It achieves dense mutual supervision between low-level and high-level semantic information, enabling DKN to uncover rich regularizations and relationships hidden within the data.

Thill et al. [103] introduced the **TCN-AE**, which incorporates multiple skip connections and dilated convolutions, improving the model's capacity to process long-distance dependencies. Their experimental validation demonstrated the advantages of skip connections and dimensionality reduction of feature maps in enhancing anomaly detection performance.

Moreover, Abilasha et al. [104] proposed **WaRTem-AD**, which employs time warping operations for data augmentation in TSAD, including replication and interpolation warping techniques. They also introduced a twin TCN autoencoder structure aimed at learning warping-robust features. In a novel approach, Huang et al. [105] presented a multi-resolution self-supervised discriminative network for anomaly detection. This architecture utilizes multi-scale down-sampling and self-supervised modules, diverging from traditional reconstruction or prediction-based scoring methodologies. Their TCN block integrates Conv1d and max-pooling layers to enhance feature extraction.

Additionally, Zhang et al. [134] developed a multi-scale convolutional attention algorithm for anomaly detection. By employing one-dimensional convolution layers with varying receptive fields and an amplified attention mechanism, their model emphasizes anomalies for improved identification. Lastly, Zhang et al. [106] introduced the **TFAD** architecture, which leverages time-frequency analysis to identify various component anomalies by utilizing decomposition details across multiple frequency domains.

3) *Graph Neural Network Model*: Graph Neural Networks (GNNs) have emerged as powerful deep learning architectures designed specifically for graph-structured data, capturing intricate relationships among nodes. In the realm of TSAD, GNNs are increasingly applied to graph-based time series data, encompassing dynamic social networks, traffic flow, and bioinformatics.

A notable approach is presented by Zhao et al. [94] with **MTAD-GAT**, which treats each univariate time series as an independent feature. This model leverages graph attention mechanisms to elucidate feature dependencies and temporal relationships, thereby enhancing interpretability. Additionally, it incorporates both prediction and reconstruction methodologies to improve the effectiveness of anomaly detection.

Similarly, Ding et al. [95] introduce **MST-GAT**, which employs multi-head attention alongside inter-modal and intra-modal learning to capture spatial features, coupled with convolutional networks for temporal feature extraction. This architecture integrates prediction and reconstruction techniques to deliver a joint anomaly scoring system.

In the work of [9], **STADN** constructs a sensor correlation matrix utilizing graph attention to extract spatiotemporal characteristics from time series data. Notably, it features an Error Refinement Module aimed at assisting users in establishing appropriate thresholds for anomaly detection.

Further extending the capabilities of GNNs, **DGINet** [96] combines a dynamic graph attention network with Informer to capture correlations among sensor features. Meanwhile, **DPGLAD** [97] employs unidirectional graph structure learning to model causal relationships among sensors within a dynamic framework, utilizing Diffusion Convolution RNN with times-tamp masking to capture both temporal and spatial features. **MagNet** [135] leverages wavelet transformations to categorize time series into multi-resolution time-frequency components. It incorporates an orthogonal kernel to restrict signals at varying frequencies, which helps in mitigating information redundancy. Furthermore, the model uses a frequency-layered GNN to capture dynamic spatial relationships, with the ultimate features being weighted and aggregated through average pooling.

Lastly, **SGTrans** [98] introduces a joint optimization approach involving a reconstruction module based on sparse autoencoders and a prediction module utilizing Graph Transformer, aimed at extracting spatiotemporal correlations across different time sequences.

4) *Transformer-based Model*: Transformer-based architectures have significantly advanced time series anomaly detection by leveraging self-attention mechanisms, which enable the processing of sequential data in parallel while being independent of fixed positional relationships among elements. This architecture excels in capturing long-range dependencies, thereby enhancing the modeling of complex temporal patterns and improving anomaly detection accuracy.

A variety of innovative approaches have emerged, building upon the foundational capabilities of Transformers. **Anomaly Transformer** [115] introduced an anomaly attention module accompanied by a correlation discrepancy mechanism, integrating prior and sequence correlations to refine normal time series modeling. Their findings suggest that while normal time series exhibit greater correlation with global sequences, abnormal instances tend to correlate more with local sequences.

Expanding the capabilities of traditional models, Zhang et al. [116] proposed the **TransAnomaly** model, which incorporates variational learning concepts into a Transformer framework by enhancing the LSTM-VAE and OmniAnomaly models. They also introduced a gated transition function, significantly

improving the model’s ability to capture dependencies among hidden variables.

To make anomalies more interpretable, CauseFormer [102] introduces a rule-based tree-structured attention mechanism that simultaneously harvests high-order features and quantifies each metric’s contribution, delivering both sample-level and metric-level explanations.

In a bid to reduce computational complexity, Doshi et al. [117] developed **TiSAT**, which features a probability attention mechanism tailored for anomaly detection. Meanwhile, TranAD [118] introduced a purely adversarial generative network based on Transformer architecture, where two decoders compete to enhance anomalies for detection purposes.

AnomalyBERT [119] further advanced the field by utilizing 1D relative positional bias embeddings along with cross-entropy for prediction loss. They implemented four data augmentation techniques—Soft replacement, Uniform replacement, Length adjustment, and Peak noise—demonstrating how self-supervised classification can enhance model adaptability across various datasets.

Moreover, **DCdetector** [120] proposed a patch-based dual self-attention representation learning mechanism that captures normal sequence feature representations through both global and local patches. Their multi-scale representation mechanism mitigates information loss during downsampling.

Lastly, the **D3TN** model [121] showcases innovative feature and temporal modules, utilizing graph convolutional networks and self-attention for extracting both global and local features. Additionally, it incorporates the POT-MoM algorithm for determining the final anomaly threshold. In a different approach, Kim et al. [122] introduced a Transformer-CNN hybrid anomaly detection network, employing attention mechanisms for multi-layered feature extraction complemented by CNNs for anomaly prediction.

5) *State Space Model*: State Space Models represent a significant class of state-based sequential models, with notable implementations such as Mamba [136]. This survey primarily focuses on deep SSM methods for TSAD, specifically Mamba (or selected state space model), which can be broadly categorized into two types: (i) **forecasting Mamba** [107, 108, 109, 112], and (ii) **TSAD-oriented Mamba** [113, 114]. Mamba is characterized as an S4 model that employs a selection mechanism computed through a scan operation. By refining the state selection equation, Mamba achieves linear complexity, thus facilitating the learning of complex long-range dependencies while minimizing computational load.

Forecasting Mamba: Building on the Mamba architecture, **S-Mamba** [107] is a simplified state space model based on Mamba, designed for time series prediction tasks. Empirical comparisons with state of the art Transformer models show that S-Mamba requires less training time and GPU memory, making it especially suitable for TSAD tasks with strict real time requirements. Several works [108, 109] have built upon the Mamba architecture to propose models better suited for long time series modeling. For example, **Mambaformer** [112] integrates the SSM with the Transformer architecture to create a hybrid Mamba-Transformer framework, showcasing scalability for large-scale long-short range time series forecasting tasks.

TSAD-oriented Mamba: **DMamba** [110] combines Mamba with a multi-stage detrending mechanism. This integration utilizes SSM’s capacity to capture long-range dependencies while the detrending process stabilizes inputs for the sequence model, thereby enhancing generalization capabilities for non-stationary data.

SGFM [113] comprises a feature extraction module and a conditional flow matching module for TSAD. The feature extraction module leverages state space models to capture long-term temporal dependencies and graph convolutional networks to model spatial relationships within the time series data. The extracted features are then used as conditions to guide the flow matching process, which models the normal data distribution by transforming a simple prior distribution to the target distribution through learned vector fields.

TSAD-C [114] introduces a Decontaminator that combines masking strategies with an SSM-based conditional diffusion model. By leveraging an SSM to capture intra-variable dynamics and a GNN to model inter-variable relationships, it excels at detecting anomalies that span long temporal horizons.

ASSM [137] is a streaming TSAD model that suppresses redundant noise by dynamically adjusting hidden features, thereby accelerating the response to abrupt anomalies. It is tailored for resource-constrained sensor environments.

Lastly, **MemMambaAD** [136] first decomposes the time series into trend and seasonal components, then employs the proposed Mamba-TCN module to extract features from each component separately. A dynamic memory bank is introduced to continuously store the decomposed representations of normal samples, enhancing TSAD performance.

6) *Others*: In TSAD, some algorithms utilize flow-based neural network architectures, like GANF, to improve data distribution estimation. Ensemble-based algorithms aggregate multiple base models in various ways to address TSAD challenges with a more extensive model set. Furthermore, researchers propose integrating federated learning with TSAD to meet high privacy and security demands.

Dai and Chen [123] presents a novel framework called Graph Augmented Normalizing Flows (**GANF**) for TSAD. This framework identifies distribution drift by incorporating Bayesian networks to model conditional dependencies among variables while jointly estimating flow parameters.

A prior study [124] significantly advances anomaly detection by introducing an ensemble of deep learning models for multivariate time series data. By combining RNN, LSTM, CNN, Transformers, and GNN models, the detection of anomalies is greatly improved. This ensemble approach harnesses the strengths of various algorithms to enhance accuracy and resilience, especially in real-time processing and decision-making in diverse industrial contexts.

Xu et al. [125] introduces **PeFAD**, a Parameter-Efficient Federated Anomaly Detection framework tailored for time series data. Aimed at providing privacy-preserving and scalable anomaly detection in decentralized environments, PeFAD employs a LLM to tackle data scarcity issues and enhance the model’s ability to generalize from limited data.

Moreover, [138, 139] fuse fuzzy theory with ensemble learning to devise lightweight, online TSAD algorithms tailored

for IoT systems.

D. Unified Time Series Anomaly Detection Model

The Unified Time Series Anomaly Detection Models adopt a unique strategy, enabling a single model with shared parameters to manage multiple distinct domains of time series data independently. Several recent models have advanced the field of TSAD through innovative mechanisms and cross-domain adaptability. **DADA** [126] represents a universal TSAD model capable of zero-shot anomaly detection without requiring fine-tuning on specific datasets. This capability is achieved through adaptive bottleneck and dual adversarial decoder mechanisms, which enhance its generalization across diverse downstream scenarios. DADA’s adaptive bottleneck module creates pools of varying latent space sizes tailored to the information bottleneck requirements of different datasets, while its dual adversarial decoder module facilitates the learning of normal and abnormal time series patterns through adversarial training, thereby establishing clear decision boundaries. Complementing DADA’s approach, **DACAD** [127] introduces a cross-domain TSAD model that integrates domain adaptation with contrastive learning. By leveraging labeled data from one domain, DACAD enhances unsupervised TSAD tasks. It generates triplets of target data, source data, and labels by injecting anomalies into both domains, utilizing TCN for representation. The model employs supervised average margin contrast loss for the source domain and self-supervised contrast triplet loss for the target domain, distinguishing between normal and anomalous data through adversarial training, and utilizes an extended Deep SVDD structure as the anomaly detector.

In the realm of unsupervised TSAD, **ContextTDA** [128] introduces a context-aware domain adaptation model. It conceptualizes the time series context sampling problem as a Markov decision process and employs deep reinforcement learning algorithms with customized reward functions to facilitate knowledge transfer, thereby enhancing anomaly detection capabilities. **FS-ADAPT** [129] specifically targets few-shot learning and unsupervised domain adaptation. Its dueling triplet network processes rearranged and augmented quadruplet data while engaging in adversarial training of feature extractors. FS-ADAPT extracts class-specific and domain-specific features to learn domain-invariant representations, addressing concept drift in anomalies through incremental adaptation. By maintaining exemplar sets and dynamically updating the feature extractor, it enhances online detection performance amidst challenges posed by limited labeled data in source domains and unlabeled data in target domains.

VANDA [130] exemplifies a domain-adaptive unsupervised model for TSAD that utilizes trainable embedding vectors to represent features of each time series variable. By constructing a graph structure based on these embeddings and engaging in self-supervised graph training, VANDA enhances embedding learning. It leverages attention mechanisms to aggregate neighboring representations and employs techniques like sliding similarity and optimal transport distance to facilitate the transfer of inter-domain variate associations.

UNITS [131] further advances cross-domain time series analysis through a modified transformer architecture. By

utilizing three token types—Sample tokens for embedding heterogeneous multivariate time series, Prompt tokens for task-specific context, and Task tokens for TSAD. UNITS efficiently detects anomalies via reconstruction errors. The model’s two-way time and variable self-attention mechanism, along with a dynamic linear operator in the feedforward layer, optimizes computation and captures complex relationships in time series data.

C. EVALUATION METHODS

Several measures for TSAD have been proposed to quantify the quality of AD methods. We categorize these measures into threshold-based and threshold-independent measures. In this section, we first introduce these two major categories, and then we will list key features of a fair metric for TSAD. Comprehensive benchmarking in TSAD is often hindered by inconsistent and flawed evaluation metrics. For a rigorous theoretical foundation, we refer readers to the pioneering work by Wagner et al. [140], which establishes a formal framework comprising verifiable properties to characterize metric behaviors, demonstrating that most existing metrics lack solid axiomatic foundations.

A. Threshold-based Metrics

In time series anomaly detection, models output an anomaly score S_i for each point p_i in the time series T . This score indicates the degree of abnormality, with higher values signifying more abnormal points. To evaluate the performance of these models, threshold-based metrics are employed, which involve setting a threshold to classify points as normal or anomalous.

Thresholding functions. Various methods have been proposed to determine optimal thresholds in anomaly detection, aiming to maximize the F1-score and enhance detection accuracy. $F1_{\text{best}}$ [141] calculates the threshold yielding the best F1-score by enumerating all possible thresholds. The top-k threshold [142] predicts k points as abnormal, matching the actual number of anomalies in the test set. Hundman et al. [143] introduces unsupervised dynamic error thresholds that identify the value causing the greatest percent decrease in the mean and standard deviation after removing samples above it. The tail probability threshold [144] assesses anomaly likelihood based on the complement of the Q-function applied to the mean and standard deviation of the recent short-term average of prediction errors. In addition, thresholding methods based on Extreme Value Theory (EVT), such as Peaks Over Threshold (POT) [145] and its online variant SPOT [146], offer statistically grounded and dynamically adjustable thresholds by modeling the tail behavior of anomaly scores. In summary, these methods provide diverse approaches to threshold determination in anomaly detection, each utilizing different statistical properties and optimization techniques to improve prediction accuracy and reliability.

1) *Point-Based Metrics: Precision, Recall, and F1 Score* are key metrics in TSAD. In point-based anomaly detection, a predicted point within an anomaly sequence is considered a True Positive. Precision measures the ratio of true positives to total actual positives, while Recall assesses the ratio of

true positives to total identified positives. The F1 score is the harmonic mean of Precision and Recall [24].

The point-based metric views each data point as an independent sample, ignoring the temporal aspects of anomaly segments, which limits its effectiveness for sequence anomaly detection. To address this, range-based metrics treat the entire range of anomalies as a single event. Additionally, statistical analysis methods evaluate model performance using foundational metrics [147, 9, 148, 149, 150].

2) *Range-based Metrics: Point Adjustment Metrics* (or event-based metrics) introduced by Xu et al. [141] provide an intuitive approach by treating each contiguous anomaly sequence as a single event. This method recognizes that detecting just one anomaly within a range is often enough to alert experts. Thus, every timestamp in an anomaly segment receives the highest anomaly score from that segment. The segment is deemed detected if this score exceeds the threshold. The point-adjusted F-score is then calculated using a selected anomaly threshold.

The point adjustment metric has a fundamental flaw: it assigns high scores to random guess models. This occurs because errors are penalized only once, while the entire anomaly sequence is rewarded for detection [117, 14]. To address this, Garg et al. [9] propose the composite F1 score, which combines point-based precision and event-based recall, promoting the detection of at least one point in every anomaly sequence while discouraging incorrect predictions. Additionally, Si et al. [151] introduce a Reduced-length PA that balances point-wise and event-wise detection, utilizing a severity coefficient $\ln(k)$ as the reduced sequence length, where k is the anomaly sequence length. Similarly, the PA%K protocol [14] considers a segment detected only if the ratio of correctly detected points in the sequence exceeds a specified threshold.

Position Relativity between anomaly sequences and detection results is a crucial measurement for TSAD that point adjustment metrics overlook. Xu et al. [141] employs average alert delay, defined as the time difference between the first point and the first detected point in the anomaly segment, to measure position relativity. Range-based Precision and Recall metrics (R-based) [152] effectively calculate $Recall_T$ and $Precision_T$ based on the existence and overlap in anomaly sequences. These metrics penalize fragmented predictions for a single anomaly while rewarding detections that are close to a defined relative position to the anomaly. The Time Series Aware Precision (TaP), Time Series Aware Recall (TaR), and Time Series Aware F1 (TaF1) metrics, collectively known as TaPR [153], and their enhanced version, eTaPR [154], evaluate both the detection and the portion scoring of anomalies. These metrics assess whether any part of an anomaly is detected and reward methods that capture a larger portion of the anomalous sequence. Improved F1 scores can be derived from these variations of recall and precision.

From a formal properties perspective, Wagner et al. [140] theoretically proved that none of these ranges-based or point-adjusted metrics satisfy all crucial requirements of TSAD metrics simultaneously. For instance, point-wise and point-adjusted F1 scores inherently fail to distinguish varied prediction patterns inside anomaly windows, severely violating

basic prerequisites such as detection timing bias and alarm redundancy penalization.

Moreover, Affiliation metrics [155] utilize the duration between ground truth and predictions, providing an intuitive interpretation. These metrics normalize precision and recall against random sampling, quantifying improvements over random predictions. However, Affiliation lacks sensitivity for accurate predictions and does not effectively highlight performance differences among models, demonstrating weaknesses in robustness against random predictions. To address these issues, Zhong et al. [66] propose Normalized Affiliation and Unbiased Affiliation (N&U Affs). The former is a special case of the latter but operates independently of the dataset, enhancing versatility. This metric reweights precision to address some limitations of the original Affiliation metric. Besides, the Proximity-Aware Time series Anomaly Evaluation (PATE) [156] incorporates proximity-based weighting, accounting for buffer zones around anomaly intervals.

B. Threshold-Independent Metrics

Threshold-based metrics require threshold selection, compelling models to learn domain-specific thresholds from training data. In practical applications, the ratio of anomaly labels often differs from that in the training data, and factors like noise and delays can significantly reduce performance on these thresholds [157, 158]. Additionally, comparing models across different datasets poses challenges. Therefore, statistical metrics that reflect overall model performance across all possible threshold settings, known as threshold-independent metrics, are necessary.

1) *AUC-ROC and AUC-PR*: Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Area Under the Precision-Recall Curve (AUC-PR) are used to assess the performance of an anomaly detection model across all threshold settings. The ROC curve visually represents the diagnostic ability of a binary classifier as the discrimination threshold varies, plotting the true positive rate (TPR) against the false positive rate (FPR) at different thresholds. In contrast, the Precision-Recall curve is particularly beneficial for imbalanced datasets, plotting precision (positive predictive value) against recall (sensitivity or true positive rate) across various threshold values. Furthermore, VUS-ROC and VUS-PR [158] enhance traditional AUC measures by evaluating the volume under the surface formed by the ROC or PR curves as thresholds and buffer lengths change. This approach recognizes the continuous nature of time series data, offering a more robust evaluation, particularly in scenarios involving noise, misalignments, and varying anomaly cardinality ratios.

2) *Statistical Analysis*: For comparisons of multiple classifiers across various datasets, statistical analysis [147] is employed in TSAD [9, 157]. The Wilcoxon test [148] is frequently used for pairwise comparisons, allowing for the evaluation of significant differences between two methods across multiple datasets. For a more thorough assessment that considers multiple methods at once, the Friedman test [149] is utilized, followed by the post-hoc Nemenyi test [150], which ranks methods and identifies statistically significant performance differences.

The relationship between threshold independent metrics and threshold based metrics lies in their foundational components. Threshold-based metrics redefine which points are detected as anomalies and can be directly converted into statistical metrics. For instance, R-based F1 [152] is calculated as AUC-PR using range-based precision and recall. Thus, threshold-independent metrics can be viewed as an extension or complement to threshold-based metrics, providing a broader evaluation framework.

C. Essential Features of Qualified Metrics

Designing effective metrics for evaluating TSAD models requires careful consideration of several key aspects. Unlike static data, time series data presents unique challenges due to its sequential nature, the importance of timely detection, and the presence of noise. Here are the essential features that a qualified metric for TSAD should encompass:

- **Temporal Adjacency Awareness (TAA):** Adapt the metric to the sequential nature of time series data, recognizing the importance of temporal relationships by valuing the proximity between predicted and actual anomalies, treating close predictions as true positives.
- **Early Detection (ED):** Prioritize the early detection of anomalies by assessing how closely detection aligns with the onset of the event.
- **Balance between Existence and Duration (BED):** Ensure the longer and potentially more severe anomalies are emphasized while not dominating the evaluation, thus recognizing both short and long anomalies appropriately.
- **Robust to Noise, Lag, and Imbalance Category Ratio (RNLI):** Develop a metric that is resilient to noise in anomaly scores, can manage delays in detection, and is sensitive to the ratio of normal to abnormal labels, providing a balanced evaluation.
- **Exactly-Once Detection (EOD):** Encourage each anomaly instance to be reported exactly once to avoid duplicate detections, which can lead to repeated alerts for a single event and confusion about whether alerts refer to the same anomaly.
- **Threshold-Free Evaluation (TFE):** Create a metric that does not rely on predefined threshold values for classifying anomalies, ensuring fair evaluation across various methods and datasets.

Inspired by the work of Zamanzadeh Darban et al. [24], widely used and recommended metrics have been integrated for comparison. Meanwhile, equally important metrics required for different scenarios have been appended. Table S6 presents a comparison of the advantages of different metrics and explains their overall characteristics, aiming to provide guidance for metric evaluation. Although metrics for TSAD have been extensively developed, we believe that no single metric is suitable for all scenarios. Therefore, we recommend using multiple different metrics to evaluate models simultaneously.

D. Evaluation Metrics' Formulas

In Table S7, we provide a compilation of the formulas for the metrics introduced in the main text, along with

necessary explanations and clarifications in the *Comment* column. Moreover, the R-based F1, TaPR-F1, and eTaPR F1 metrics share a common foundational logic, computed using the formula: $2 \cdot \frac{\text{Precision}_T \cdot \text{Recall}_T}{\text{Precision}_T + \text{Recall}_T}$. Specifically, all three metrics define Precision_T and Recall_T to weight interval predictions, integrating factors such as prediction accuracy, overlap, and positional fidelity. The final score adheres to a calculation paradigm analogous to the classical F1 metric, with the added nuance of incorporating anomaly persistence. Given the complexity inherent in their underlying principles and computational frameworks, we recommend consulting the original literature to avoid oversimplification and to maintain technical precision. Similarly, the definition of the more intricate VUS metric has not been introduced in this section. To bridge the gap between intuitive requirements and counter-intuitive metric evaluations, Wagner et al. [140] proposed LARM. As detailed in Table S7, LARM establishes a tiered scoring mechanism by explicitly decomposing the evaluation into true positive patterns, false alarm frequencies, and point-wise false positives. We recommend incorporating LARM/ALARM alongside conventional F_1 variants to deliver a more trustworthy and consistent model ranking.

E. Flaws of Existing Metrics: A Case Analysis

To evaluate the reliability of conventional evaluation protocols, we analyze standard metrics including Precision, Recall, point-wise F1, F1-PA [141], Aff-F1 [155], and UAff-F1 [66]. Fig. S1 illustrates a synthetic scenario for time-series anomaly detection, where the top row displays the ground-truth binary labels. The sequences *Random1*, *Random2*, *Pred1*, and *Pred2* denote anomaly scores generated by two trivial baselines and two candidate detectors, respectively. Labels with the PA suffix designate point adjustment, wherein an entire anomaly segment is marked as detected if any single time point within it is identified.

As shown, *Random1* flags all time points as anomalous, while *Random2* randomly and uniformly samples anomalies. Despite having zero practical utility, both baselines achieve inflated F1-PA scores, confirming that point adjustment can introduce severe evaluation artifacts.

Point-wise F1 is sensitive to anomaly duration and provides a more faithful reference for localized events (such as the segment at the end of the series), serving as a baseline anchor for non-trivial predictions.

Among the evaluated models, *Pred1* demonstrates the highest detection fidelity. However, its Aff-F1 score is paradoxically lower than that of the all-anomaly baseline *Random1*, because the original Aff-F1 formulation does not adequately penalize false positives. In contrast, UAff-F1 remedies this deficiency by imposing an explicit false-positive penalty, correctly ranking *Pred1* (18.18) above *Pred2* (0.57) and suppressing trivial random baselines. Consequently, conventional F1-PA is vulnerable to artificial inflation, whereas metrics incorporating false-alarm penalties provide a substantially more reliable assessment.

F. Experimental Analysis of Metrics

Dataset Configuration. To thoroughly evaluate metric behaviors, we employ both synthetic and real-

TABLE S6: Comparison of metric features and interpretability.

Metric	TAA	ED	BED	RNLI	EOD	TFE	Overview
Precision							Enforces low false alarms, emphasizing high detection confidence at the point level.
Recall							Prioritizes capturing all anomalous points, tolerating higher false alarm rates.
F1-Score							Harmonic mean of point-wise Precision and Recall, providing balanced baseline evaluation.
F1-PA [141]	✓						Applies point adjustment where identifying any single point marks the entire segment detected; highly vulnerable to score inflation.
Reduced F1 [151]	✓		✓				Logarithmically compresses segment lengths post-adjustment to mitigate length bias.
PA%K [14]	✓		✓				Requires at least $K\%$ of points within an anomaly segment to be identified before triggering adjustment.
R-based F1 [152]	✓	✓	✓		✓		Evaluates overlap cardinality and temporal continuity while penalizing fragmented predictions.
TaPR [153]	✓		✓				Balances detection across segments and points, prioritizing complete anomaly interval coverage.
eTaPR [154]	✓	✓	✓				Extends TaPR with explicit emphasis on early detection and critical long anomaly durations.
Affiliation [155]	✓	✓	✓		✓		Measures distance between predicted and ground-truth intervals, providing continuous proximity-based scoring.
N&U Affs [66]	✓	✓	✓		✓		Normalizes and removes precision bias in Affiliation F1 under high anomaly density regimes.
PATE [156]	✓	✓	✓	✓		✓	Parameter-free range-based AUC metric weighting early detection, duration, and delay.
AUC-ROC				✓		✓	Measures threshold-free ranking capability across varying operating conditions; sensitive to label imbalance.
AUC-PR						✓	Evaluates precision-recall trade-off across all thresholds, well-suited for heavily imbalanced anomaly distributions.
VUS-ROC [158]	✓			✓		✓	Computes volume under the ROC surface across variable range-buffer sizes for threshold-free range evaluation.
VUS-PR [158]	✓					✓	Measures volume under the PR surface, combining threshold-free range evaluation with imbalance resilience.
LARM [140]	✓	✓	✓	✓	✓		Provably satisfies all 9 formal metric properties, decomposing scores into TP pattern alignment, false alarm count, and FP penalties.
ALARM [140]	✓	✓	✓	✓	✓		Extends LARM with alarm tolerance (t) and explicit differentiation among early, late, and true false alarms.

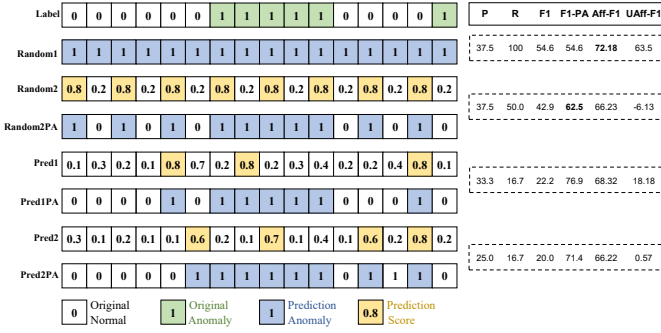


Fig. S1: Synthetic case study demonstrating the vulnerabilities and scoring biases of conventional evaluation metrics.

world datasets (summarized in Table S8). We construct 30 synthetic time series following the schema $\langle \text{Length} \rangle - \langle \text{Segments} \rangle - \langle \text{MeanLen} \rangle \langle \text{Variance} \rangle$. For instance, $100k-20seg-50L$ designates a sequence of length $L = 100k$ with $N_{seg} = 20$ anomaly segments, where the mean segment length is 50 with low variance (L, lengths in $[40, 60]$). Conversely, the suffix H denotes high variance (H, lengths in $[1, 99]$). The columns *Max/Min Seg Length* in Table S8 denote the extreme boundaries of the generated anomaly intervals. We generate 15 synthetic configurations each for sequence lengths of 100k and 10k, supplemented by 6 diverse real-world datasets covering varied anomaly lengths and frequencies.

Algorithm Configuration. To ensure an objective evaluation of metric properties independent of specific detector biases, we simulate synthetic anomaly detectors across controlled accuracy levels. Specifically, let Model- q denote a detector with nominal prediction accuracy $q\%$ across both normal and anomalous intervals (e.g., Model-60 or Model-70). For anomalous intervals, correct predictions yield scores drawn from $0.1 \cdot \mathcal{U}(0, 1) + 0.9$, while erroneous predictions produce $0.1 \cdot \mathcal{U}(0, 1) + 0.05$, where $\mathcal{U}(0, 1)$ is a standard uniform distribution. For normal intervals, these scoring distributions are reversed. This stochastic noise injection ensures algorithmic

diversity and enables evaluating metric robustness against score perturbations.

We evaluate computational latency and scoring sensitivity across four key dimensions:

- 1) Detector accuracy levels Model- q (Fig. S2);
- 2) Overall sequence length L (Fig. S3);
- 3) Number of anomaly segments N_{seg} (Fig. S4);
- 4) Mean anomaly segment duration (Fig. S5).

Each factor is evaluated on both synthetic and real-world benchmarks. Complete raw experimental data are available in the supplementary repository.

Performance of Metrics Against Different Anomaly Scoring Models. As shown in Fig. S2(a–b), detector accuracy has negligible impact on the computational latency of most metrics, with a minor exception in PATE due to its dependence on the number of generated interval candidates. When a model approaches random guessing, it generates the highest number of candidate events, resulting in peak latency.

However, metric discrimination varies considerably across accuracy tiers (Fig. S2(c–d)):

- 1) Point-wise F1 yields expected values around 0.5–0.6 even for an ideal model with theoretical 100% accuracy (Model-100). This phenomenon occurs because point-wise F1 is sensitive to continuous score noise across decision thresholds, preventing the empirical score from reaching 1.0.
- 2) F1-PA exhibits extreme score variance under identical models, leading to rank instability across decision thresholds. For instance, Model-100 may yield an F1-PA score of only ≈ 0.3 (comparable to Model-20–Model-30), whereas Model-50 can reach 0.8, demonstrating severe unreliability for model comparison.
- 3) Reduced-F1, R-based F1, TaPR, eTaPR, PATE, and AUC-PR display limited discriminability across moderate accuracy regimes (10%–90%), exhibiting a sharp non-linear jump only at Model-100. Moreover, R-based F1 and TaPR fail to attain the theoretical maximum of 1.0.

TABLE S7: Overview of the formulas for evaluation metrics along with accompanying comments.

Metric	Formula	Comment
Precision	$\text{Precision} = \frac{\sum_{i=1}^N y_i \cdot \hat{y}_i}{\sum_{i=1}^N \hat{y}_i}$	Classical precision.
Recall	$\text{Recall} = \frac{\sum_{i=1}^N y_i \cdot \hat{y}_i}{\sum_{i=1}^N y_i}$	Classical recall.
F1-Score	$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	Classical F1 score.
F1-PA [141]	$\hat{y}_i = \begin{cases} 1, & \text{if } \mathcal{F}_t > \delta \\ & \text{or } t \in S \text{ and } \exists t' \in S, \mathcal{F}_{t'} > \delta \\ 0, & \text{otherwise.} \end{cases}$	The F1-PA metric follows the same calculation framework as the classical F1-Score. However, F1-PA applies a post-processing adjustment to the predicted results $\hat{y}_i$. Here, S denotes an arbitrary anomaly segment, and $\mathcal{F}$ represents the anomaly score function.
Reduced F1 [151]	$ S_i = \ln(S_i) + 1,$ $\text{Pre} = \frac{\sum_{i=1}^{ S } S_i \cdot \hat{S}_i}{\sum_{i=1}^{ S } \hat{S}_i}, \text{Rec} = \frac{\sum_{i=1}^{ S } S_i \cdot \hat{S}_i}{\sum_{i=1}^{ S } S_i},$ $\text{ReducedF1} = 2 \cdot \frac{\text{Pre} \cdot \text{Rec}}{\text{Pre} + \text{Rec}}$	Reduced F1 also employs a point adjustment strategy, but its improvement lies in adjusting the length of anomaly segments. For any anomaly segment S_i , after point adjustment, the anomaly length is post-processed as shown in the first formula. The F1 score is then calculated on an event-by-event basis to mitigate the bias introduced by point adjustments.
PA%K [14]	$\hat{y}_i = \begin{cases} 1, & \text{if } \mathcal{F}_t > \delta \text{ or} \\ & t \in S \text{ and } \frac{ \{t' t' \in S, \mathcal{F}_{t'} > \delta\} }{ S } \geq K \\ 0, & \text{otherwise} \end{cases}$	PA%K is an improvement over F1-PA, modifying the original point adjustment strategy. In the traditional approach, an entire anomaly segment S was considered successfully predicted if any single point within the segment was correctly identified. By contrast, PA%K requires that the proportion (e.i. $\frac{ \{t' t' \in S, \mathcal{F}_{t'} > \delta\} }{ S }$) of correctly predicted anomaly points within a segment exceeds $K\%$ (where K is a pre-defined value ranging from 0 to 100) to trigger post-processing and mark the segment as successfully predicted. When K varies across 0 to 100, the area under the PA%K curve can be computed as a new metric, denoted by $F1_{\text{PA}\%K}$.
Affiliation F1 [155]	$S = \{j \in [1, n]; \text{pred} \cap I_j \neq \emptyset\},$ $\text{Aff-Pre} = \frac{1}{ S } \sum_{j \in S} P_{\text{precision}_j},$ $\text{Aff-Rec} = \frac{1}{n} \sum_{j=1}^n P_{\text{recall}_j} \cdot$ $\text{Aff-F1} = 2 \cdot \frac{\text{Aff-Pre} \cdot \text{Aff-Rec}}{\text{Aff-Pre} + \text{Aff-Rec}}.$	In the Aff-F1 metric, the predicted results are partitioned into n intervals, where I_j denotes the j -th zone of affiliation of the ground truth event. Aff-Pre and Aff-Rec denote Affiliation Precision and Affiliation Recall, respectively.
N&U Affs [66]	$\text{UAff-Pre} = \frac{\text{Aff-Pre} - \text{Aff-Pre}_{\text{bias}}}{1 - \text{Aff-Pre}_{\text{bias}}},$ $\text{UAff-F1} = \frac{2 \cdot \text{UAff-Pre} \cdot \text{Aff-Rec}}{ \text{UAff-Pre} + \text{Aff-Rec}} \cdot (-1)^{ \text{UAff-Pre} < 0 },$ $\text{NAff-Pre} = 2 \cdot (\text{Aff-Pre} - 0.5),$ $\text{NAff-F1} = \frac{2 \cdot \text{NAff-Pre} \cdot \text{Aff-Rec}}{ \text{NAff-Pre} + \text{Aff-Rec}} \cdot (-1)^{ \text{NAff-Pre} < 0 }.$	NAff-F1 and UAff-F1 denote Normalized Affiliation F1 and Unbiased Affiliation F1, respectively. Here, Aff-Pre _{bias} represents the dataset-related precision bias estimated by the algorithm.
PATE [156]	$\text{Precision}_{e,d}(\theta) = \frac{\sum_{t=1}^N w^{\text{TP}}(t)}{\sum_{t=1}^T w^{\text{TP}}(t) + \sum_{t=1}^N w^{\text{FP}}(t)},$ $\text{Recall}_{e,d}(\theta) = \frac{\sum_{t=1}^N w^{\text{TP}}(t)}{\sum_{t=1}^N w^{\text{TP}}(t) + \sum_{t=1}^N w^{\text{FN}}(t)},$ $\text{PATE} = \frac{1}{ E \times D } \sum_{e \in E, d \in D} \text{AUC} - \text{PR}_{e,d}$	In PATE, the weights for true positive rate, false positive rate, and false negative rate are defined as $w^{\text{TP}}(t)$, $w^{\text{FP}}(t)$, and $w^{\text{FN}}(t)$, respectively. The overall PATE score is determined by averaging the computed AUC-PRs across all combinations of e and d , where $ D $ and $ E $ denote the number of distinct values for d and e within their respective sets.
LARM [140]	$\text{LARM}(g, p) = \text{Score}_{\text{TP}} - \text{Pen}_{\text{FA}} - \text{Pen}_{\text{FP}}$ $\text{Score}_{\text{TP}} = \frac{1}{ I_1(g) } \sum_{A \in I_1(g)} \frac{\alpha(p_A) + 1}{2^{ I_1(p_A) }},$ $\text{Pen}_{\text{FA}} = 2 \sum_{A \in I_0(g)} I_1(p_A) $ $\text{Pen}_{\text{FP}} = \sum_{A \in I_0(g)} \beta(p_A^{-1}(1))$	A tiered metric provably satisfying all 9 formal properties. Score_{TP} measures pattern alignment inside true anomaly windows $I_1(g)$, while Pen_{FA} and Pen_{FP} explicitly penalize the number of false alarms and point-wise false positives in normal regions $I_0(g)$, respectively.
AUC-ROC	$\int_0^1 \text{TPR}(\text{FPR}) d\text{FPR}$	It is obtained by calculating the area under the curve of the True Positive Rate (TPR) and False Positive Rate (FPR).
AUC-PR	$\int_0^1 \text{Precision}(\text{Recall}) d\text{Recall}$	Here, Precision and Recall respectively denote the precision curve and the recall curve.

4) Although VUS is threshold-independent, its event buffer introduces an operational bias. While buffer size preserves relative model rankings, consistent configurations must be maintained to ensure comparable absolute scores.

Impact of Overall Time Series Length. As shown in Fig. S3(a–b), sequence length directly scales metric evaluation latency. PATE is computationally the most demanding; for series lengths exceeding 100k, its latency is two or more orders of magnitude higher than other interval metrics. Among interval-based formulations, VUS-ROC and VUS-PR exhibit moderate latency, whereas R-based F1 remains the fastest.

Fig. S3(c–d) depicts metric sensitivity with respect to sequence length. For a balanced baseline with 50% expected accuracy, theoretical expectations should align with 0 or 0.5

depending on metric formulation. However, Aff-F1, VUS-ROC, and VUS-PR exhibit systematic baseline shifts (expected scores of 0.67 and 0.8, respectively). Furthermore, F1-PA displays volatile baseline fluctuations across real-world datasets (ranging from 0 to 0.8), making it unsuitable as an isolated evaluation criterion in practical deployments.

Impact of Anomaly Segment Count. As shown in Fig. S4(a–b), the latency of most interval-based metrics increases with segment density (with the exception of F1-PA, Reduced-F1, and PA%K). PATE is especially sensitive to segment count, experiencing a tenfold latency surge when segment count exceeds 50. As depicted in Fig. S4(c), Reduced-F1, TaPR, PATE, VUS-ROC, and VUS-PR show increasing bias with higher segment counts in synthetic data, though this effect

is attenuated in real-world benchmarks containing sparser anomalies.

Impact of Average Segment Length. Segment length exerts an inverse effect on interval metric latency (Fig. S5(a)): latency decreases as average segment duration increases for a fixed sequence length, whereas point-wise metrics exhibit constant execution times. In terms of scoring behavior (Fig. S5(c)), F1-PA scores monotonically increase with segment length, whereas TaPR, PATE, VUS-ROC, and VUS-PR continuously decrease. Under high segment variance (e.g., 20L vs. 20H, or 100L vs. 100H), bias increases for point-oriented metrics (F1, F1-PA, PA%K, Reduced-F1, TaPR, PATE, AUC-PR), while NAff-F1, UAff-F1, VUS-ROC, and VUS-PR remain more stable, indicating their suitability for evaluating extended anomaly durations.

G. Practical Recommendations for Metric Selection

TABLE S8: Time series characteristics of synthetic and real data.

Dataset Name	TS Length	Segments	Max Seq Length	Min Seq Length
100k-20seg-50L	100000	20	60	40
100k-200seg-50L	100000	200	60	40
100k-20seg-50H	100000	20	99	1
100k-200seg-50H	100000	200	99	1
100k-50seg-20L	100000	50	30	10
100k-500seg-20L	100000	500	30	10
100k-50seg-20H	100000	50	39	1
100k-500seg-20H	100000	500	39	1
100k-10seg-100L	100000	10	110	90
100k-100seg-100L	100000	100	110	110
100k-10seg-100H	100000	10	199	1
100k-100seg-100H	100000	100	199	1
100k-2seg-500L	100000	2	550	450
100k-20seg-500L	100000	20	550	450
100k-2seg-500H	100000	2	999	1
100k-20seg-500H	100000	20	999	1
10k-2seg-50L	10000	2	60	40
10k-20seg-50L	10000	20	60	40
10k-2seg-50H	10000	2	99	1
10k-20seg-50H	10000	20	99	1
10k-5seg-20L	10000	5	30	10
10k-50seg-20L	10000	50	30	10
10k-5seg-20H	10000	5	39	1
10k-50seg-20H	10000	50	39	1
10k-1seg-100L	10000	1	110	90
10k-10seg-100L	10000	10	110	110
10k-1seg-100H	10000	1	199	1
10k-10seg-100H	10000	10	199	1
10k-2seg-500L	10000	2	550	450
10k-2seg-500H	10000	2	999	1
MSL	73729	36	1141	11
Creditcard	284807	465	5	1
SWAT	449919	35	35900	101
SMD-1-1	28479	8	721	2
SMD-2-1	23694	13	452	8
SMD-3-1	28700	4	131	21

The selection of an appropriate evaluation metric is contingent upon the specific objectives of the analysis and the inherent characteristics of the time series data and anomaly detection task. Drawing from our experimental analysis, the following practical strategies and recommendations are provided:

- 1) **Consider Latency Requirements:** When evaluation speed is a primary concern, particularly for very long time series ($> 100k$) or datasets containing numerous anomaly segments (> 50), PATE should be avoided due to excessive computational overhead. In such scenarios, point-based metrics, or interval-based metrics less sensitive to segment count (such as F1-PA, Reduced-F1, and PA%K), are preferable for achieving lower latency.

2) Understand Anomaly Characteristics:

- For scenarios involving short anomaly segments or low variance in segment length, metrics such as F1, F1-PA, PA%K, Reduced-F1, TaPR, PATE, and AUC-PR are more appropriate, given their reduced bias under these conditions.
- Conversely, for long anomaly segments or high variance in segment length, NAff-F1, UAff-F1, VUS-ROC, and VUS-PR generally demonstrate superior stability with lower inherent bias.

3) Evaluate Discriminability and Robustness:

- To effectively discriminate between models across moderate accuracy regimes (e.g., 10%–90%), caution is advised when using Reduced-F1, PATE, and AUC-PR, as these metrics demonstrate poor discriminability in non-saturated regimes.
- F1 and F1-PA exhibit sensitivity to noise and variations in threshold. Consequently, F1-PA should not be employed in isolation for real-world evaluations due to its variable bias baseline and high score variance; it should be complemented by more stable range-based metrics.
- The inability of R-based F1 and TaPR to achieve theoretical maximums may result in a consistent underestimation of performance, which should be accounted for when absolute scores are critical.

- 4) **Threshold Dependency:** For threshold-independent evaluation, VUS metrics (VUS-ROC, VUS-PR) represent robust options. Nevertheless, it is essential to standardize the event buffer size across all comparative analyses to prevent artificial bias.

- 5) **Interpretability of Metric:** In practical applications requiring an intuitive understanding of model performance, metrics relying on unrestricted point adjustment (such as F1-PA) are generally not recommended. Standard point-wise metrics (F1, AUC-ROC, AUC-PR) provide direct point-level interpretability, while range-based metrics (eTaPR, Aff-F1, NAff-F1, UAff-F1, PATE, VUS-ROC, VUS-PR) offer direct interval-level assessment. A structured comparison of all metrics is provided in Table S6.

- 6) **Utilize Multiple Metrics for a Holistic View:** Given the inherent trade-offs and sensitivities of individual metrics, it is recommended practice to report results using a complementary combination of metrics. This approach facilitates a more comprehensive and robust evaluation by mitigating individual metric limitations (e.g., pairing a threshold-dependent metric with a threshold-independent one, or combining point-wise and interval-based formulations).

By carefully considering these factors, researchers can make an informed decision regarding the most suitable metrics for their specific time series anomaly detection evaluation requirements.

D. GENERAL TECHNOLOGIES IN TSAD

This section reviews foundational general-purpose methodologies supporting modern TSAD pipelines, focusing on data

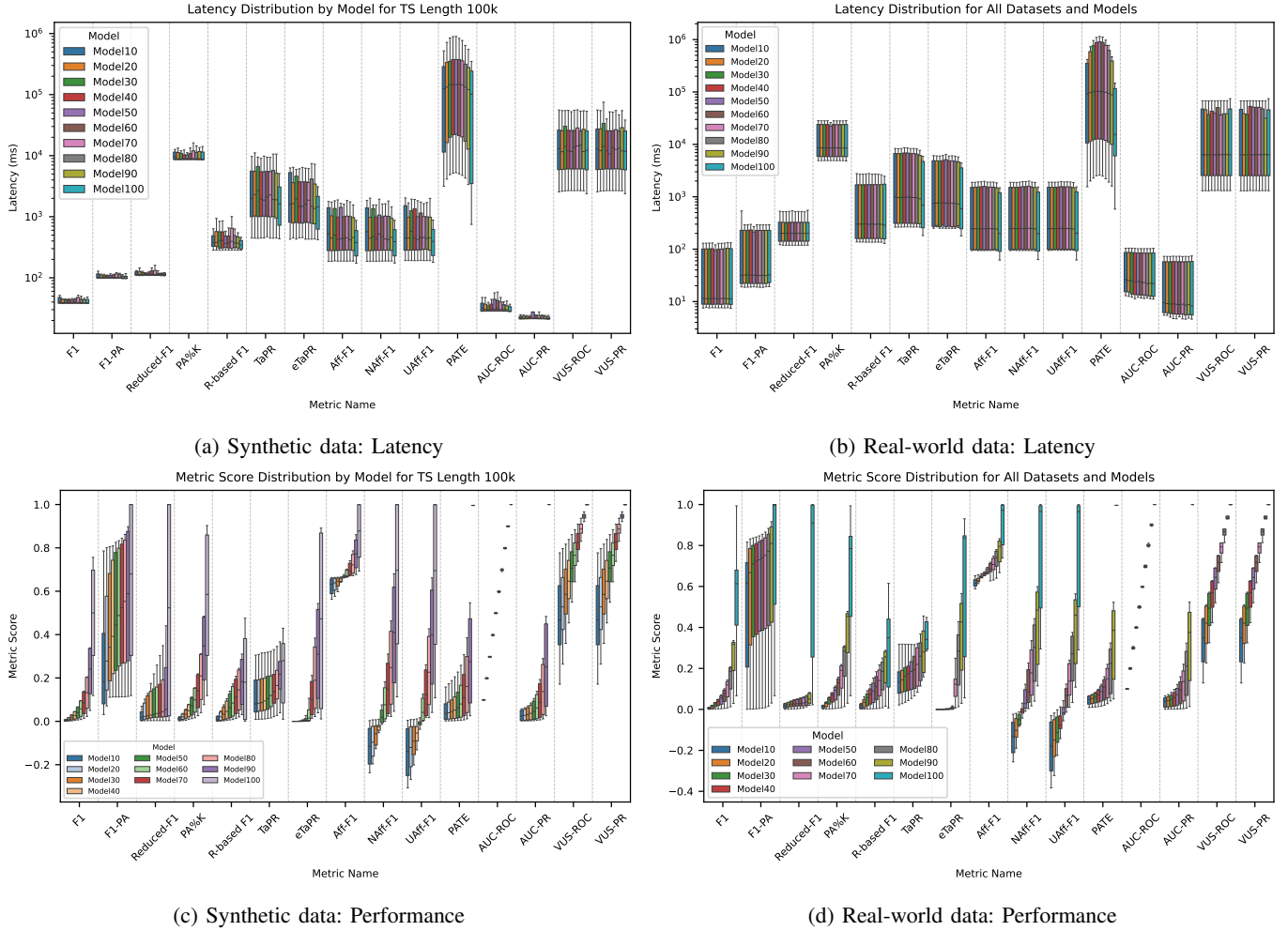


Fig. S2: Performance variation in latency and metrics across different models for synthetic and real-world data.

augmentation frameworks and the architectural paradigm of channel independence.

A. Time Series Data Augmentation

Data augmentation enhances model generalizability and representation robustness by expanding the effective training distribution [62, 72, 126]. Mainstream time series augmentation techniques can be categorized into four methodological families:

1) Basic Signal Transformations: Elementary operations perturb temporal sequences without altering fundamental dynamics. *Jittering and Gaussian noise injection* [66] perturb signal values to force models to learn noise-invariant representations. *Scaling* modifies signal amplitudes by a scalar factor ($t' = \alpha t$), encouraging scale invariance. *Temporal inversion* reverses the sequence index order ($t' = [t_N, \dots, t_1]$) to assess directional symmetry. *Smoothing* applies moving-average or low-pass filtering to attenuate high-frequency fluctuations.

2) Cross-Channel and Representation Mixing: *Mixup* [111] generates convex combinations of distinct channels to mitigate distribution drift and prevent over-smoothing. **Multi-PSCA** [159] leverages Dynamic Time Warping to synthesize

temporal perturbations, applying PCA decomposition to extract path signatures as complementary augmented features.

3) Self-Supervised Masking and Imputation: Masking strategies replace contiguous or random intervals with missing indicators, training models to reconstruct corrupted inputs. Empirical findings demonstrate that masked imputation tasks (e.g., ImDiffusion [46] and DDMT [48]) yield tighter prediction bounds and superior representation fidelity compared to unconstrained autoregressive forecasting.

4) Model-Agnostic Anomaly Injection: To expose models to synthetic outliers during self-supervised or weakly supervised training, diverse anomaly injection protocols are employed:

- *Peak Injection:* Overwriting isolated timestamps with global extrema.
- *Contextual Injection:* Perturbing timestamps toward local mean values while maintaining global amplitude bounds.
- *Collective/Pattern Injection:* Replacing contiguous segments with constant values, random walks, or mismatched historical subsequences.
- *Seasonal/Trend Perturbation:* Rescaling periodic components by period shifts τ or applying non-linear scaling α to decomposed trend vectors s^{trend} . Generative frameworks such as GenIAS [160] further employ TCN-VAE

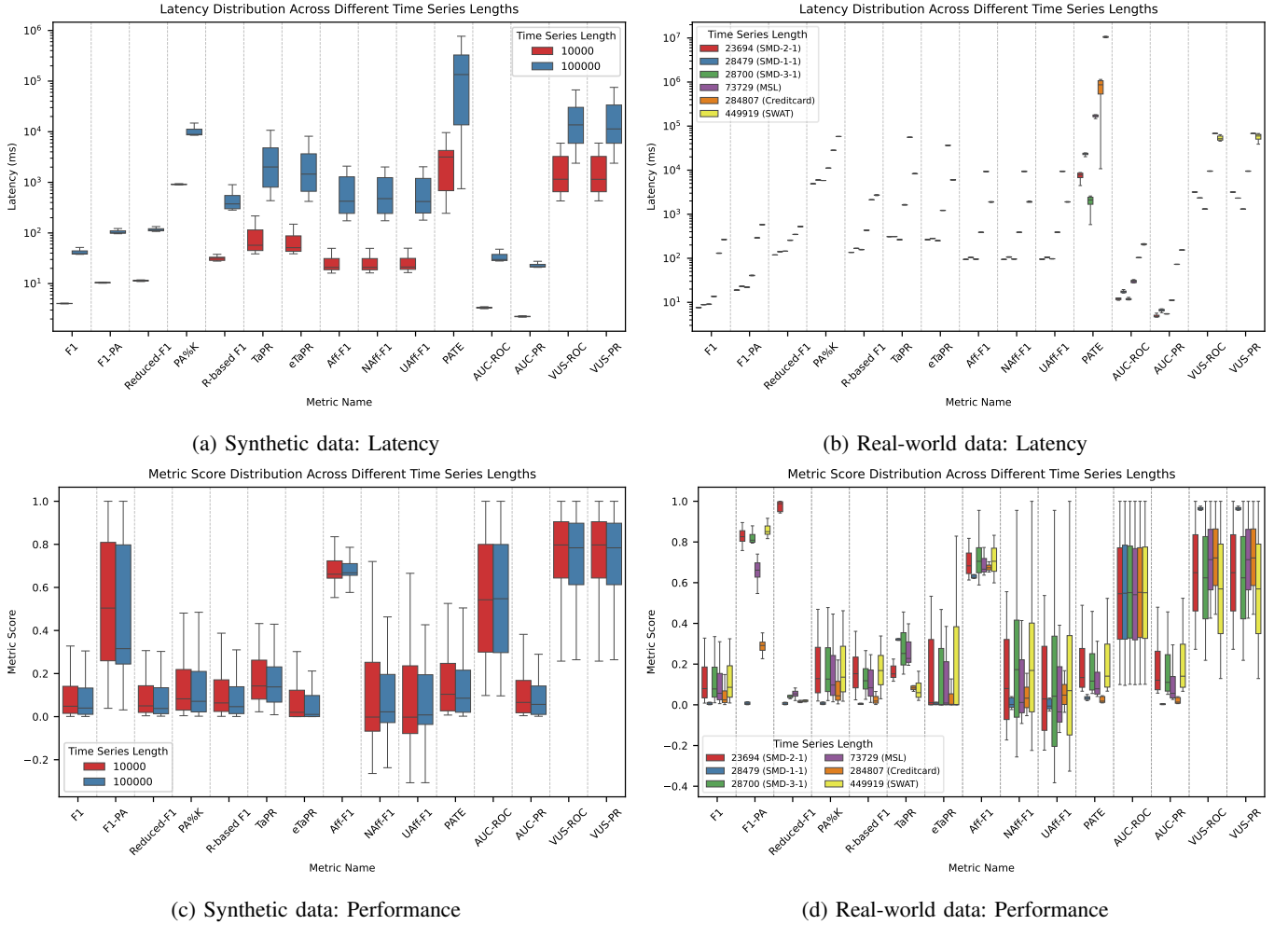


Fig. S3: Performance variation in latency and metrics across different time series lengths for synthetic and real-world data.

backbones to synthesize realistic anomalous trajectories.

Methodological Challenges and Frontiers: Despite their widespread utility, existing augmentation methods exhibit notable theoretical limitations: (i) Linear mixup interpolations can disrupt temporal causality and violate dynamical system constraints (e.g., Granger causality [161]); (ii) Unstructured Gaussian perturbations lack semantic selectivity and are rapidly filtered out by deep architectures; adversarial noise formulations (such as TimeGAN-based perturbations [162]) offer more effective hard boundary samples; (iii) Independent univariate augmentations destroy inter-variable correlations, compromising detection for channel-correlation anomalies. Future frameworks should prioritize structure-preserving spatial augmentations, such as graph-attention modeling via STG-NF [163] or causal discovery-guided synthesis [161].

B. Channel Independence

Channel independence processes each multivariate dimension through shared univariate model parameters, treating a C -channel series as C independent 1D sequences during feature extraction [164]. This design confers two decisive advantages: (i) it mitigates over-parameterization and suppresses overfitting to spurious cross-channel correlations in noisy regimes; and (ii)

it provides structural agnosticism for unified cross-domain models, accommodating datasets with arbitrary channel dimensions without architectural modification.

However, strict channel independence is inherently limited when evaluating channel-correlation (CC) anomalies, as individual marginal channels may exhibit normal behavior while joint distributions are violated. Consequently, effective unified architectures should incorporate hierarchical processing: maintaining channel-independent feature extraction in early layers while deploying adaptive cross-attention, graph message passing, or dynamic channel gating in deeper layers to capture inter-variable dependencies.

E. APPLICATIONS

This section reviews key application domains for TSAD algorithms, followed by a compilation of standardized benchmark datasets and open-source software libraries.

A. Application Domains

Different operational environments impose distinct architectural priorities on TSAD algorithms:

Business KPI Monitoring: Enterprise monitoring frameworks (such as Yahoo EGADS [165]) process millions of

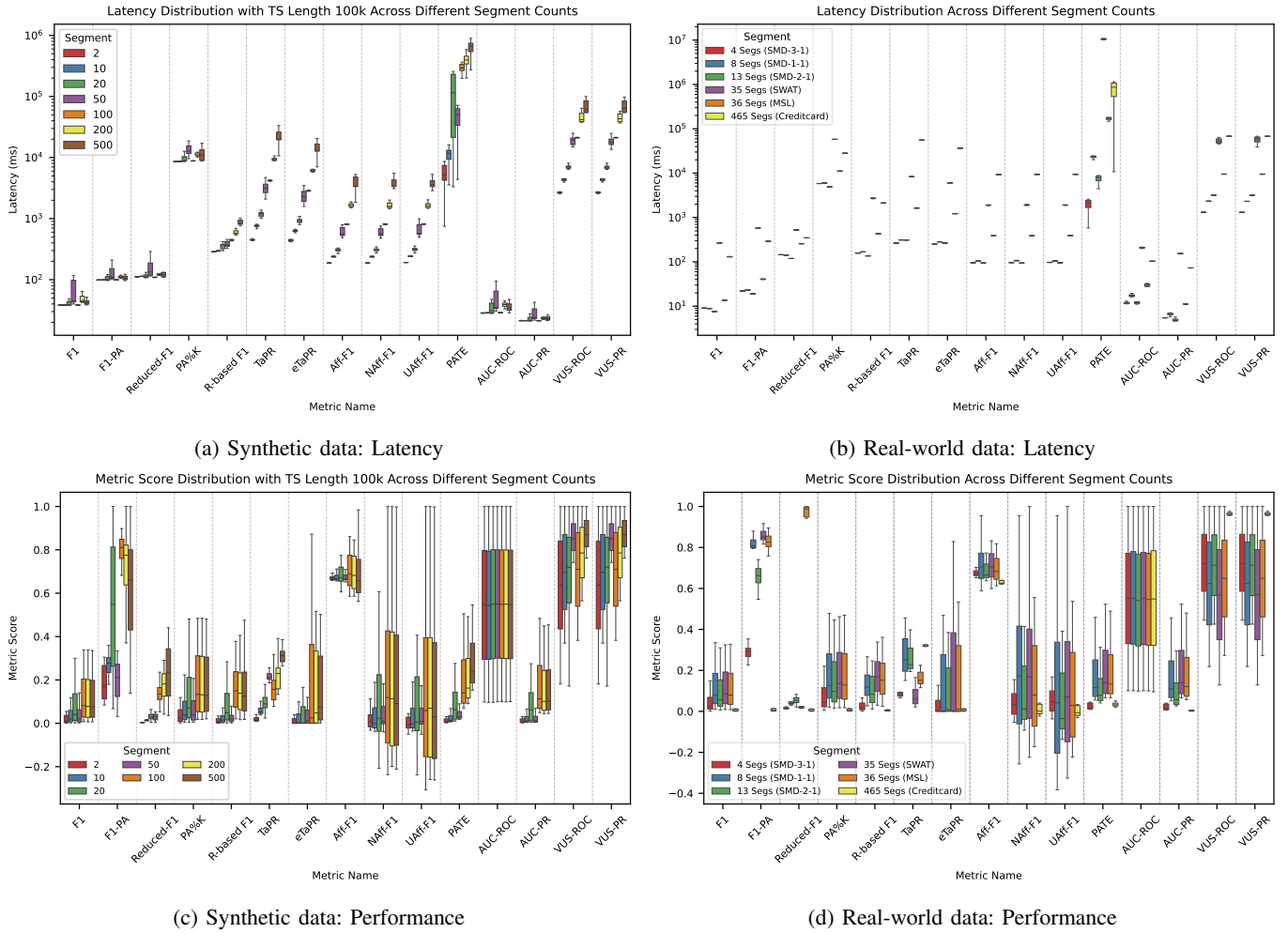


Fig. S4: Performance variation in latency and metrics across the number of anomaly segments for synthetic and real-world data.

time series concurrently to identify service outages. Key requirements include high throughput, linear computational scalability, and robustness against non-stationary trend shifts.

Power Grid and Converter Systems: Monitoring modern electrical infrastructures [166] requires sub-second streaming inference and rigorous post-hoc interpretability, enabling operators to identify physical failure causes under tight real-time constraints.

Intelligent Transportation and Vehicle Trajectories: Trajectory anomaly detection (exemplified by DiffTAD [167]) models continuous multi-agent spatial geometries and temporal evolution simultaneously, requiring models to capture complex spatio-temporal dependencies.

Clinical Physiological Signal Processing: Healthcare monitoring (e.g., multi-lead ECG analysis via ECGGAN [168]) requires high diagnostic precision and cross-channel consistency across coupled physiological leads, with strict demands on model transparency to support clinical verification.

Industrial Predictive Maintenance: Smart manufacturing controllers [169] operate under strict edge-hardware constraints, requiring lightweight online stream detectors capable of continuous self-adaptation with minimal memory footprints.

B. Datasets

Table S9 compiles the most commonly used datasets in state-of-the-art methods, along with advanced synthetic datasets such as GutenTAG and TSB-UAD. We have summarized the links for fully open access and download. Additionally, in this table, we have documented the domain/application of each dataset, as well as its characteristics. For instance, #Subset indicates the number of subsets within a dataset, which often vary in time series length, collection location, or individual subject. Avg TS Len represents the average length of time series in each subset, Dim denotes the dimensionality of the time series data (*i.e.*, the number of features), and Anomaly Ratio indicates the average proportion of anomalies within the dataset. All these datasets are labeled. Furthermore, GutenTAG is a synthetic data generation toolkit that can produce diverse data subsets with customizable properties such as the number of channels and anomaly ratio. Therefore, its intrinsic dataset characteristics are not documented. These datasets cover common scenarios in TSAD, including server operation monitoring, human health tracking, and specific anomaly detection in industrial and commercial settings.

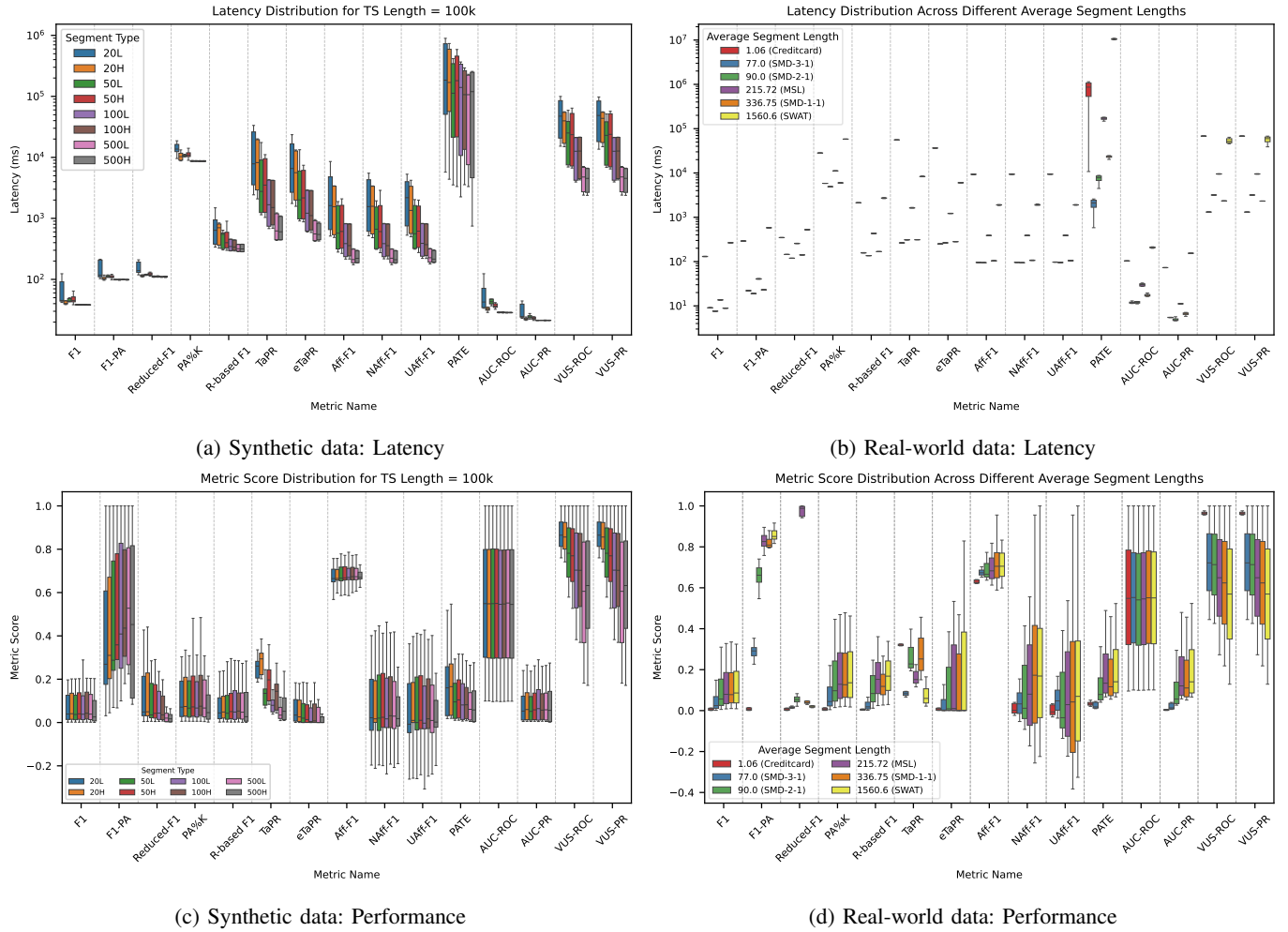


Fig. S5: Performance variation in latency and metrics across different segment lengths for synthetic and real-world data.

TABLE S9: Time series datasets from different domains.

Name	Type	Paper	Public Access Link	Domain/Application	#Subset	Avg TS Len	Dim	Anomaly Ratio (%)
IOPS	real	[101]	https://github.com/iopsai/iops	Business KPI	58	72792.3	1	1.3
Yahoo	real	-	https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70	Business KPI	367	1560.2	1	2.5
UCR	real	[170]	https://wu.renjie.im/research/anomaly-benchmarks-are-flawed/#ucr-timeseries-anomaly-archive	Health tracking	250	67818.7	1	0.6
ASD	real	[35]	https://github.com/zhilee/InterFusion	Application server	12	6423.79	19	4.61
SMD	real	[90]	https://github.com/NetManAI/Ops/OmniAnomaly	Server machine	28	25466.4	38	3.8
WaQ	real	-	https://www.spotseven.de/gecco/gecco-challenge	Water quality	1	230172	9	2.02
MSL	real	[143]	https://github.com/khundman/telemanom	Sensor	27	3119.4	55	5.1
SMAP	real	[143]	https://github.com/khundman/telemanom	Sensor	54	7855.9	25	2.9
WADI	real	[171]	https://itrust.sutd.edu.sg/itrust-labs_datasets/	Water distribution	241921	34561	123	5.74
SWaT	real	[172]	https://itrust.sutd.edu.sg/itrust-labs_datasets/	Secure water treatment	4	207457.5	59	12.7
SKAB	real	-	https://www.kaggle.com/dsv/1693952	Water circulation system	34	1200	8	35.48
GAIA	real	-	https://github.com/CloudWise-OpenSource/GAIA-DataSet	Log detection	279	10156.41	1	0.87
Exathlon	real	[173]	https://github.com/exathlonbenchmark/exathlon	Spark application monitoring	32	44075.8	1	11.0
Creditcard	real	-	https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud	Financial fraud detection	1	284807.0	29	0.2
CoMuT	synthetic	[174]	https://hpi.de/naumanns/comut	Synthetic data	180	10000	9	0.93
NAB	real/synthetic	[144]	https://www.kaggle.com/datasets/boltzmannbrain/nab	Streaming media	58	5099.7	1	10.6
GutenTAG	synthetic	[175]	https://github.com/TimeEval/gutentag	Synthetic data tool	-	-	-	-
TSB-UAD	real/synthetic	[157]	https://github.com/TheDatumOrg/TSB-UAD	Dataset collection	1125	91694.87	1	5.95

C. Algorithm Library

Table S10 lists open-source implementations of representative TSAD algorithms. The **Paradigm** column indicates the core detection paradigm: Reconstruction (Rec), Forecasting (Fcst), Generative (Gen), Similarity-based (Sim), or LLM-based (LLM).

REFERENCES

- [1] M. Landauer, M. Wurzenberger, F. Skopik, G. Settanni, and P. Filzmoser, “Time Series Analysis: Unsupervised
- [2] Z. Zhao, Y. Zhang, X. Zhu, and J. Zuo, “Research on Time Series Anomaly Detection Algorithm and Application,” *Proceedings of 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2019*, no. Iaeac, pp. 16–20, 2019.
- [3] A. A. Cook, G. Misirli, and Z. Fan, “Anomaly Detection for IoT Time-Series Data: A Survey,” *IEEE Internet of*

Anomaly Detection Beyond Outlier Detection,” *Information Security Practice and Experience*, vol. 11125, pp. 19–36, 2018.

TABLE S10: Open-source implementations of representative TSAD algorithms.

Algorithm	Paradigm	Public Access Link
AMSL	Sim	https://github.com/zhangyuxin621/AMSL
MEMTO	Sim	https://github.com/gunny97/MEMTO
PatchAD	Sim	https://github.com/EmorZz1G/PatchAD
SimAD	Sim	https://github.com/EmorZz1G/SimAD
DCdetector	Rec	https://github.com/DAMO-DI-ML/KDD2023-DCdetector
USD	Sim	https://github.com/zangzelin/code_USD
TS-TCC	Sim	https://github.com/emadeldeen24/TS-TCC
COUTA	Sim	https://github.com/xuhongzuo/couta
COCA	Sim	https://github.com/ruiking04/COCA
OCAN	Sim	https://github.com/PanpanZheng/OCAN
FGANomaly	Gen	https://github.com/sxxmason/FGANomaly
BeatGAN	Gen	https://github.com/hi-bingo/BeatGAN
UAC-AD	Gen	https://github.com/lhysgithub/UAC-AD
MSCRED	Rec	https://github.com/7fantasysz/MSCRED
UAE	Rec	https://github.com/astha-chem/mvts-ano-eval
USAD	Rec	https://github.com/manigalati/usad
NPSR	Rec	https://github.com/andrewlai61616/NPSR
InterFusion	Gen	https://github.com/zhlee/InterFusion
MUTANT	Gen	https://github.com/Coac-syf/MUTANT
FCVAE	Gen	https://github.com/CSTCloudOps/FCVAE
OmniAnomaly	Gen	https://github.com/NetManAI/Ops/OmniAnomaly
MTAD-GAT	Rec	https://github.com/ML4ITS/mtad-gat-pytorch
TFAD	Rec	https://github.com/DAMO-DI-ML/CIKM22-TFAD
AnomalyTrans	Rec	https://github.com/thuml/Anomaly-Transformer
AnomalyBERT	Rec	https://github.com/Jhryu30/AnomalyBERT
CauseFormer	Rec	https://github.com/zhongguoxiang/CauseFormer
S-Mamba	Fcst	https://github.com/wzhwzhwh0921/S-D-Mamba
Mambaformer	Fcst	https://github.com/XiongxiaoXu/Mambaformer-in-Time-Series
UniTS	Rec	https://github.com/mims-harvard/UniTS
D3R	Gen	https://github.com/ForestsKing/D3R
ImDiffusion	Gen	https://github.com/17000cyh/IMDiffusion
DiffusionAE	Gen	https://github.com/tbrad/DiffusionAE
DDMT	Gen	https://github.com/yangchaocheng/DMDDT
TimeADDM	Gen	https://github.com/Hurongyao/TIMEADDM
GANF	Gen	https://github.com/EnyanDai/GANF
Timer	LLM	https://github.com/thuml/Large-Time-Series-Model
GPT4TS	LLM	https://github.com/DAMO-DI-ML/One_Fits_All
SIGLLM	LLM	https://github.com/sintel-dev/sigllm
aLLM4TS	LLM	https://github.com/yxbian23/aLLM4TS
AnomalyLLM	LLM	https://github.com/fly-orange/AnomalyLLM

Things Journal, vol. 7, no. 7, pp. 6481–6494, 2020.

- [4] M. Braei and S. Wagner, “Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art,” 2020.
- [5] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, “A review on outlier/anomaly detection in time series data,” 2020.
- [6] K. Choi, J. Yi, C. Park, and S. Yoon, “Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines,” *IEEE Access*, vol. 9, pp. 120 043–120 065, 2021.
- [7] K. Shaukat, T. M. Alam, S. Luo, S. Shabbir, I. A. Hameed, J. Li, S. K. Abbas, and U. Javed, “A Review of Time-Series Anomaly Detection Techniques: A Step to Future Perspectives,” in *Advances in Information and Communication*, K. Arai, Ed. Cham: Springer International Publishing, 2021, vol. 1363, pp. 865–877.
- [8] K.-H. Lai, D. Zha, J. Xu, Y. Zhao, G. Wang, and X. Hu, “Revisiting time series outlier detection: Definitions and benchmarks,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021.
- [9] A. Garg, W. Zhang, J. Samaran, R. Savitha, and C.-S. Foo, “An evaluation of anomaly detection and diagnosis in multivariate time series,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2508–2517, 2021.
- [10] J. E. Zhang, D. Wu, and B. Boulet, “Time Series Anomaly Detection for Smart Grids: A Survey,” 2021.
- [11] S. Schmidl, P. Wenig, and T. Papenbrock, “Anomaly detection in time series: A comprehensive evaluation,” *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, 2022.
- [12] R. Wu and E. J. Keogh, “Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2421–2429, 2023.
- [13] H. Alimohammadi and S. Nancy Chen, “Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis,” *Expert Systems with Applications*, vol. 191, p. 116371, 2022.
- [14] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon, “Towards a rigorous evaluation of time-series anomaly detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7194–7201.
- [15] M. A. Belay, S. S. Blakseth, A. Rasheed, and P. Salvo Rossi, “Unsupervised Anomaly Detection for IoT-Based Multivariate Time Series: Existing Solutions, Performance Analysis and Future Directions,” *Sensors*, vol. 23, no. 5, p. 2844, 2023.
- [16] B. Kim, M. A. Alawami, E. Kim, S. Oh, J. Park, and

- H. Kim, "A Comparative Study of Time Series Anomaly Detection Models for Industrial Control Systems," *Sensors*, vol. 23, no. 3, p. 1310, 2023.
- [17] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: A survey," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 2023, pp. 6778–6786.
- [18] G. Li and J. J. Jung, "Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges," *Information Fusion*, vol. 91, pp. 93–102, 2023.
- [19] P. Yan, A. Abdulkadir, P.-P. Luley, M. Rosenthal, G. A. Schatte, B. F. Grewe, and T. Stadelmann, "A Comprehensive Survey of Deep Transfer Learning for Anomaly Detection in Industrial Time Series: Methods, Applications, and Directions," *IEEE Access*, vol. 12, pp. 3768–3789, 2024.
- [20] M. Jin, H. Y. Koh, Q. Wen, D. Zambon, C. Alippi, G. I. Webb, I. King, and S. Pan, "A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, p. 10466–10485, Dec. 2024.
- [21] W. Sun, H. Li, Q. Liang, X. Zou, M. Chen, and Y. Wang, "On data efficiency of univariate time series anomaly detection models," *Journal of Big Data*, vol. 11, no. 1, p. 83, 2024.
- [22] J. Ye, W. Zhang, K. Yi, Y. Yu, Z. Li, J. Li, and F. Tsung, "A Survey of Time Series Foundation Models: Generalizing Time Series Representation with Large Language Model," 2024.
- [23] J. Su, C. Jiang, X. Jin, Y. Qiao, T. Xiao, H. Ma, R. Wei, Z. Jing, J. Xu, and J. Lin, "Large language models for forecasting and anomaly detection: A systematic literature review," *arXiv preprint arXiv:2402.10350*, 2024.
- [24] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," *ACM Comput. Surv.*, vol. 57, no. 1, Oct. 2024.
- [25] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, and G. Shroff, "LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection," *ArXiv preprint*, vol. abs/1607.00148, 2016.
- [26] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, and N. V. Chawla, "A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.
- [27] A. Garg, W. Zhang, J. Samaran, S. Ramasamy, and C.-S. Foo, "An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2508–2517, 2022.
- [28] M. Hu and P. Xia, "Industrial Time-Series Signal Anomaly Detection Based on G-LSTM-AE Model," in *Artificial Intelligence in China*, Q. Liang, W. Wang, J. Mu, X. Liu, and Z. Na, Eds. Singapore: Springer Nature Singapore, 2023, vol. 871, pp. 383–391.
- [29] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised Anomaly Detection on Multivariate Time Series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event CA USA: ACM, 2020, pp. 3395–3404.
- [30] T. Amarbayasgalan, V. H. Pham, N. Theera-Umpon, and K. H. Ryu, "Unsupervised Anomaly Detection Approach for Time-Series in Multi-Domains Using Deep Reconstruction Error," *Symmetry*, vol. 12, no. 8, p. 1251, 2020.
- [31] C.-Y. A. Lai, F.-K. Sun, Z. Gao, J. H. Lang, and D. Boning, "Nominality score conditioned time series anomaly detection by point/sequential reconstruction," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 637–76 655, Dec. 2023.
- [32] X. Kong, Z. Chen, W. Liu, K. Ning, L. Zhang, S. Muhammad Marier, Y. Liu, Y. Chen, and F. Xia, "Deep learning for time series forecasting: a survey," *International Journal of Machine Learning and Cybernetics*, vol. 16, no. 7, pp. 5079–5112, Aug. 2025. [Online]. Available: <https://doi.org/10.1007/s13042-025-02560-w>
- [33] D. Park, Y. Hoshi, and C. C. Kemp, "A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-Based Variational Autoencoder," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.
- [34] T.-A. Pham, J.-H. Lee, and C.-S. Park, "MST-VAE: Multi-Scale Temporal Variational Autoencoder for Anomaly Detection in Multivariate Time Series," *Applied Sciences*, vol. 12, no. 19, p. 10078, 2022.
- [35] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. Virtual Event Singapore: ACM, 2021, pp. 3220–3230.
- [36] Y. Shi, B. Wang, Y. Yu, X. Tang, C. Huang, and J. Dong, "Robust anomaly detection for multivariate time series through temporal GCNs and attention-based VAE," *Knowledge-Based Systems*, vol. 275, p. 110725, 2023.
- [37] Z. Wang, C. Pei, M. Ma, X. Wang, Z. Li, D. Pei, S. Rajmohan, D. Zhang, Q. Lin, H. Zhang, J. Li, and G. Xie, "Revisiting vae for unsupervised time series anomaly detection: A frequency perspective," in *Proceedings of the ACM Web Conference 2024*, ser. WWW '24. New York, NY, USA: Association for

- Computing Machinery, 2024, p. 3096–3105.
- [38] B. Du, X. Sun, J. Ye, K. Cheng, J. Wang, and L. Sun, “Gan-based anomaly detection for multivariate time series using polluted training set,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12 208–12 219, 2023.
 - [39] Z. Niu, K. Yu, and X. Wu, “LSTM-Based VAE-GAN for Time-Series Anomaly Detection,” *Sensors*, vol. 20, no. 13, p. 3738, 2020.
 - [40] F. Zeng, M. Chen, C. Qian, Y. Wang, Y. Zhou, and W. Tang, “Multivariate time series anomaly detection with adversarial transformer architecture in the Internet of Things,” *Future Generation Computer Systems*, vol. 144, pp. 244–255, 2023.
 - [41] A. Geiger, D. Liu, S. Alnegheimish, A. Cuesta-Infante, and K. Veeramachaneni, “Tadgan: Time series anomaly detection using generative adversarial networks,” *2020 IEEE International Conference on Big Data (Big Data)*, pp. 33–43, 2020.
 - [42] A.-H. Shin, S. T. Kim, and G.-M. Park, “Time Series Anomaly Detection Using Transformer-Based GAN With Two-Step Masking,” *IEEE Access*, vol. 11, pp. 74 035–74 047, 2023.
 - [43] S. Liu, B. Zhou, Q. Ding, B. Hooi, Z. Zhang, H. Shen, and X. Cheng, “Time Series Anomaly Detection With Adversarial Reconstruction Networks,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 4293–4306, 2023.
 - [44] H. Liu, X. Huang, M. Jia, T. Jia, J. Han, Y. Li, and Z. Wu, “Uac-ad: Unsupervised adversarial contrastive learning for anomaly detection on multi-modal data in microservice systems,” *IEEE Transactions on Services Computing*, pp. 1–14, 2024.
 - [45] C. Wang, Z. Zhuang, Q. Qi, J. Wang, X. Wang, H. Sun, and J. Liao, “Drift doesnt matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 10 758–10 774.
 - [46] Y. Chen, C. Zhang, M. Ma, Y. Liu, R. Ding, B. Li, S. He, S. Rajmohan, Q. Lin, and D. Zhang, “Imdiffusion: Imputed diffusion models for multivariate time series anomaly detection,” pp. 359–372, 2023.
 - [47] I. Pintilie, A. Manolache, and F. Brad, “Time Series Anomaly Detection using Diffusion-based Models,” in *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2023, pp. 570–578.
 - [48] C. Yang, T. Wang, and X. Yan, “DDMT: Denoising Diffusion Mask Transformer Models for Multivariate Time Series Anomaly Detection,” 2023.
 - [49] R. Hu, X. Yuan, Y. Qiao, B. Zhang, and P. Zhao, “Unsupervised Anomaly Detection for Multivariate Time Series Using Diffusion Model,” in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 9606–9610.
 - [50] J. Sui, J. Yu, Y. Song, and J. Zhang, “Anomaly Detection for Telemetry Time Series Using a Denoising Diffusion Probabilistic Model,” *IEEE Sensors Journal*, vol. 24, no. 10, pp. 16 429–16 439, 2024.
 - [51] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 427–438.
 - [52] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.
 - [53] K.-W. Wang and S.-J. Qin, “A hybrid approach for anomaly detection using k-means and pso,” in *2nd International Conference on Electronics, Network and Computer Engineering (ICENCE 2016)*. Atlantis Press, 2016, pp. 821–826.
 - [54] J. Li, H. Izakian, W. Pedrycz, and I. Jamal, “Clustering-based anomaly detection in multivariate time series data,” *Applied Soft Computing*, vol. 100, p. 106919, 2021.
 - [55] S. Zhang, C. Fung, S. Huang, Z. Luan, and D. Qian, “Psom: periodic self-organizing maps for unsupervised anomaly detection in periodic time series,” in *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*. IEEE, 2017, pp. 1–6.
 - [56] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: identifying density-based local outliers,” in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.
 - [57] M. Goldstein and A. Dengel, “Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm,” *KI-2012: poster and demo track*, vol. 1, pp. 59–63, 2012.
 - [58] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, and G. H. Chen, “Ecod: Unsupervised outlier detection using empirical cumulative distribution functions,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12 181–12 193, 2022.
 - [59] S. Mauceri, J. Sweeney, and J. McDermott, “Dissimilarity-based representations for one-class classification on time series,” *Pattern Recognition*, vol. 100, p. 107122, 2020.
 - [60] I. T. Nicholas, J.-S. Lee, and D.-K. Kang, “One-Class Convolutional Neural Networks for Water-Level Anomaly Detection,” *Sensors*, vol. 22, no. 22, p. 8764, 2022.
 - [61] R. Wang, C. Liu, X. Mou, K. Gao, X. Guo, P. Liu, T. Wo, and X. Liu, “Deep contrastive one-class time series anomaly detection,” *SIAM*, pp. 694–702, 2023.
 - [62] H. Xu, Y. Wang, S. Jian, Q. Liao, Y. Wang, and G. Pang, “Calibrated one-class classification for unsupervised time series anomaly detection,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2024.
 - [63] L. Shen, Z. Li, and J. T. Kwok, “Timeseries anomaly detection using temporal hierarchical one-class network,” in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell,

- M. Balcan, and H. Lin, Eds., 2020.
- [64] Z. Zhong, Z. Yu, Y. Yang, W. Wang, K. Yang, and C. L. P. Chen, "Patchad: A lightweight patch-based mlp-mixer for time series anomaly detection," *IEEE Transactions on Big Data*, pp. 1–15, 2025.
 - [65] Z. Z. Darban, G. I. Webb, S. Pan, C. C. Aggarwal, and M. Salehi, "Carla: Self-supervised contrastive representation learning for time series anomaly detection," p. 110874, 2025.
 - [66] Z. Zhong, Z. Yu, X. Xi, Y. Xu, W. Cao, Y. Yang, K. Yang, and J. You, "Simad: A simple dissimilarity-based approach for time-series anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2025.
 - [67] Y. Jiao, K. Yang, D. Song, and D. Tao, "Timeautoad: Autonomous anomaly detection with self-supervised contrastive loss for multivariate time series," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1604–1619, May 2022.
 - [68] H. Liu, X. Qiu, Y. Shi, and Z. Zang, "Usd: Unsupervised soft contrastive learning for fault detection in multivariate time series," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
 - [69] G. Zhong, F. Liu, J. Jiang, B. Wang, X. Yao, and C. P. Chen, "Detecting cloud anomaly via broad network-based contrastive autoencoder," *IEEE Transactions on Network and Service Management*, vol. 21, no. 3, pp. 3249–3263, 2024.
 - [70] J. Song, K. Kim, J. Oh, and S. Cho, "Memto: Memory-guided transformer for multivariate time series anomaly detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 57 947–57 963, Dec. 2023.
 - [71] Y. Zhang, J. Wang, Y. Chen, H. Yu, and T. Qin, "Adaptive Memory Networks with Self-supervised Learning for Unsupervised Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4347, no. c, pp. 1–13, 2022.
 - [72] Z. Zhong, Z. Yu, Z. Fan, C. L. P. Chen, and K. Yang, "Adaptive memory broad learning system for unsupervised time series anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024.
 - [73] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.
 - [74] R. Zhang, L. Xia, and Y. Lu, "Anomaly detection of ics based on eb-ocsvm," in *Journal of Physics: Conference Series*, vol. 1267, no. 1. IOP Publishing, 2019, p. 012054.
 - [75] M. Bevilacqua and S. Tsafaris, "Dictionary-decomposition-based one-class svm for unsupervised detection of anomalous time series," in *Proceedings of 23rd European signal processing conference (EUSIPCO)*, 2015, pp. 1776–1780.
 - [76] Z. Wang, Y. Fu, C. Song, P. Zeng, and L. Qiao, "Power system anomaly detection based on ocsvm optimized by improved particle swarm optimization," *IEEE Access*, vol. 7, pp. 181 580–181 588, 2019.
 - [77] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth IEEE international conference on data mining*. IEEE, 2008, pp. 413–422.
 - [78] W. Dang, B. Zhou, L. Wei, W. Zhang, Z. Yang, and S. Hu, "Ts-bert: Time series anomaly detection via pre-training model bert," in *Computational Science-ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part II 21*. Springer, 2021, pp. 209–223.
 - [79] T. Zhou, P. Niu, L. Sun, R. Jin *et al.*, "One fits all: Power general time series analysis by pretrained lm," *Advances in neural information processing systems*, vol. 36, pp. 43 322–43 355, 2023.
 - [80] Y. Bian, X. Ju, J. Li, Z. Xu, D. Cheng, and Q. Xu, "Multi-patch prediction: Adapting llms for time series representation learning," *International Conference on Machine Learning (ICML)*, pp. 1–14, 2024.
 - [81] Y. Liu, H. Zhang, C. Li, X. Huang, J. Wang, and M. Long, "Timer: Generative pre-trained transformers are large time series models," in *Forty-first International Conference on Machine Learning*, 2024.
 - [82] C. Liu, S. He, Q. Zhou, S. Li, and W. Meng, "Large language model guided knowledge distillation for time series anomaly detection," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI '24, 2024.
 - [83] J. Liu, C. Zhang, J. Qian, M. Ma, S. Qin, C. Bansal, Q. Lin, S. Rajmohan, and D. Zhang, "Large language models can deliver accurate and interpretable time series anomaly detection," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, ser. KDD '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 4623–4634.
 - [84] S. Alnegheimish, L. Nguyen, L. Berti-Equille, and K. Veeramachaneni, "Large language models can be zero-shot anomaly detectors for time series?" *arXiv preprint arXiv:2405.14755*, 2024.
 - [85] Z. Pang, Y. Luan, J. Chen, and T. Li, "Parinfopt: An llm-based two-stage framework for reliability assessment of rotating machine under partial information," *Reliability Engineering & System Safety*, p. 110312, 2024.
 - [86] H. Zhang, X. Xu, X. Wang, J. Zuo, C. Han, X. Huang, C. Gao, Y. Wang, and N. Sang, "Holmes-vad: Towards unbiased and explainable video anomaly detection via multi-modal llm," *arXiv preprint arXiv:2406.12235*, 2024.
 - [87] S. Zhou, Y. Xuan, Y. Ao, X. Wang, T. Fan, and H. Wang, "MERIT: Multi-Agent Collaboration for Unsupervised Time Series Representation Learning," in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 24 011–24 028.
 - [88] Z. He, S. Alnegheimish, and M. Reimherr, "Harnessing vision-language models for time series anomaly

- detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 26, 2026, pp. 21 690–21 698.
- [89] S. Mokhtar, A. Mousakhan, S. Galesso, J. Tayyub, and T. Brox, “Detect, classify, act: Categorizing industrial anomalies with multi-modal large language models,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 4058–4067.
- [90] Y. Su, R. Liu, Y. Zhao, W. Sun, C. Niu, and D. Pei, “Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network,” *Kdd*, vol. 1485, pp. 2828–2837, 2019.
- [91] H. Zhou, K. Yu, X. Zhang, G. Wu, and A. Yazidi, “Contrastive autoencoder for anomaly detection in multivariate time series,” *Information Sciences*, vol. 610, pp. 266–280, 2022.
- [92] B. Xue, X. Gao, B. Li, F. Zhai, J. Lu, J. Yu, S. Fu, and C. Xiao, “A robust multi-scale feature extraction framework with dual memory module for multivariate time series anomaly detection,” *Neural Networks*, vol. 177, p. 106395, Sep. 2024.
- [93] Q. Miao, D. Wang, C. Xu, J. Zhan, and C. Wu, “An unsupervised long- and short-term sparse graph neural network for multi-sensor anomaly detection,” *IEEE Sensors Journal*, pp. 1–1, 2024.
- [94] H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, B. Xu, J. Bai, J. Tong, and Q. Zhang, “Multivariate time-series anomaly detection via graph attention network,” *IEEE*, pp. 841–850, 2020.
- [95] C. Ding, S. Sun, and J. Zhao, “MST-GAT: A multimodal spatial-temporal graph attention network for time series anomaly detection,” *Information Fusion*, vol. 89, pp. 527–536, 2023.
- [96] X. Huang, N. Chen, Z. Deng, and S. Huang, “Multivariate time series anomaly detection via dynamic graph attention network and informer,” *Applied Intelligence*, Jun. 2024.
- [97] S. He, G. Li, J. Wang, K. Xie, and P. K. Sharma, “Unidirectional graph structure learning-based multivariate time series anomaly detection with dynamic prior knowledge,” *International Journal of Machine Learning and Cybernetics*, May 2024.
- [98] Q. Yang, J. Zhang, J. Zhang, C. Sun, S. Xie, S. Liu, and Y. Ji, “Graph transformer network incorporating sparse representation for multivariate time series anomaly detection,” *Electronics*, vol. 13, no. 11, p. 2032, Jan. 2024.
- [99] S. Bai, J. Z. Kolter, and V. Koltun, “An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling,” 2018.
- [100] J. Park, Y. Park, and C.-I. Kim, “Tcae: Temporal convolutional autoencoders for time series anomaly detection,” in *2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, Jul. 2022, pp. 421–426.
- [101] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, and Q. Zhang, “Time-series anomaly detection service at microsoft,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 3009–3017.
- [102] G. Zhong, F. Liu, J. Jiang, and C. L. P. Chen, “Cause-former: Interpretable anomaly detection with stepwise attention for cloud service,” *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 637–652, 2024.
- [103] M. Thill, W. Konen, H. Wang, and T. Bäck, “Temporal convolutional autoencoder for unsupervised anomaly detection in time series,” *Applied Soft Computing*, vol. 112, p. 107751, 2021.
- [104] S. Abilasha, S. Bhadra, P. Deepak, and A. Mathew, “Warping resilient scalable anomaly detection in time series,” *Neurocomputing*, vol. 511, pp. 22–33, 2022.
- [105] D. Huang, L. Shen, Z. Yu, Z. Zheng, M. Huang, and Q. Ma, “Efficient time series anomaly detection by multiresolution self-supervised discriminative network,” *Neurocomputing*, vol. 491, pp. 261–272, 2022.
- [106] C. Zhang, T. Zhou, Q. Wen, and L. Sun, “Tfad: A decomposition time series anomaly detection architecture with time-frequency analysis,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, ser. CIKM ’22. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 2497–2507.
- [107] Z. Wang, F. Kong, S. Feng, M. Wang, X. Yang, H. Zhao, D. Wang, and Y. Zhang, “Is mamba effective for time series forecasting?” *Neurocomputing*, vol. 619, p. 129178, 2025.
- [108] A. Liang, X. Jiang, Y. Sun, X. Shi, and K. Li, “Bi-Mamba+: Bidirectional Mamba for Time Series Forecasting,” 2024.
- [109] X. Cai, Y. Zhu, X. Wang, and Y. Yao, “MambaTS: Improved Selective State Space Models for Long-term Time Series Forecasting,” 2024.
- [110] J. Chen, X. Tan, S. Rahardja, J. Yang, and S. Rahardja, “Joint selective state space model and detrending for robust time series anomaly detection,” *IEEE Signal Processing Letters*, vol. 31, pp. 2050–2054, 2024.
- [111] C. Zeng, Z. Liu, G. Zheng, and L. Kong, “C-Mamba: Channel Correlation Enhanced State Space Models for Multivariate Time Series Forecasting,” 2024.
- [112] X. Xu, Y. Liang, B. Huang, Z. Lan, and K. Shu, “Integrating Mamba and Transformer for Long-Short Range Time Series Forecasting,” 2024.
- [113] Y. He, T. Yan, Y. Zhan, Z. Feng, and Y. Xia, “Sgfm: Conditional flow matching for time series anomaly detection with state space models,” *IEEE Internet of Things Journal*, vol. 11, no. 22, pp. 36 979–36 990, 2024.
- [114] T. K. K. Ho and N. Armanfard, “Multivariate Time-Series Anomaly Detection with Contaminated Data,” 2024.
- [115] J. Xu, H. Wu, J. Wang, and M. Long, “Anomaly transformer: Time series anomaly detection with association discrepancy,” in *International Conference on Learning Representations*, 2021, pp. 1–14.

- [116] H. Zhang, Y. Xia, T. Yan, and G. Liu, "Unsupervised Anomaly Detection in Multivariate Time Series through Transformer-based Variational Autoencoder," in *2021 33rd Chinese Control and Decision Conference (CCDC)*. Kunming, China: IEEE, 2021, pp. 281–286.
- [117] K. Doshi, S. Abudalou, and Y. Yilmaz, "Reward once, penalize once: Rectifying time series anomaly detection," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.
- [118] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: deep transformer networks for anomaly detection in multivariate time series data," *Proc. VLDB Endow.*, vol. 15, no. 6, p. 1201–1214, Feb. 2022.
- [119] Y. Jeong, E. Yang, J. H. Ryu, I. Park, and M. Kang, "AnomalyBERT: Self-Supervised Transformer for Time Series Anomaly Detection using Data Degradation Scheme," pp. 1–8, 2023.
- [120] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3033–3045.
- [121] C. Wang, S. Xing, R. Gao, L. Yan, N. Xiong, and R. Wang, "Disentangled Dynamic Deviation Transformer Networks for Multivariate Time Series Anomaly Detection," *Sensors*, vol. 23, no. 3, p. 1104, 2023.
- [122] J. Kim, H. Kang, and P. Kang, "Time-series anomaly detection with stacked Transformer representations and 1D convolutional network," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105964, 2023.
- [123] E. Dai and J. Chen, "Graph-augmented normalizing flows for anomaly detection of multiple time series," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net, 2022.
- [124] A. Iqbal, R. Amin, F. S. Alsubaei, and A. Alzahrani, "Anomaly detection in multivariate time series data using deep ensemble models," *PLOS ONE*, vol. 19, no. 6, p. e0303890, Jun. 2024.
- [125] R. Xu, H. Miao, S. Wang, P. S. Yu, and J. Wang, "Pefad: A parameter-efficient federated framework for time series anomaly detection," in *KDD*, 2024, pp. 3621–3632.
- [126] Q. Shentu, B. Li, K. Zhao, Y. Shu, Z. Rao, L. Pan, B. Yang, and C. Guo, "Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders," 2025.
- [127] Z. Z. Darban, Y. Yang, G. I. Webb, C. C. Aggarwal, Q. Wen, S. Pan, and M. Salehi, "Dacac: Domain adaptation contrastive learning for anomaly detection in multivariate time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 8, pp. 4485–4496, 2025.
- [128] K.-H. Lai, L. Wang, H. Chen, K. Zhou, F. Wang, H. Yang, and X. Hu, "Context-aware domain adaptation for time series anomaly detection," in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 676–684.
- [129] H. Li, W. Zheng, F. Tang, Y. Zhu, and J. Huang, "Few-shot time-series anomaly detection with unsupervised domain adaptation," *Information Sciences*, vol. 649, p. 119610, 2023.
- [130] Y. He, Y. Bian, X. Ding, B. Wu, J. Guan, J. Zhang, and S. Zhou, "Variate associated domain adaptation for unsupervised multivariate time series anomaly detection," *ACM Transactions on Knowledge Discovery from Data*, 2024.
- [131] S. Gao, T. Koker, O. Queen, T. Hartvigsen, T. Tsiligkaridis, and M. Zitnik, "Units: A unified multi-task time series model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 140 589–140 631, 2025.
- [132] C. Sun, H. Li, M. Song, and S. Hong, "A Ranking-Based Cross-Entropy Loss for Early Classification of Time Series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 11 194–11 203, Aug. 2024.
- [133] Z. Xiao, H. Xing, R. Qu, L. Feng, S. Luo, P. Dai, B. Zhao, and Y. Dai, "Densely Knowledge-Aware Network for Multivariate Time Series Classification," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 54, no. 4, pp. 2192–2204, Apr. 2024.
- [134] J. Zhang, C. Wang, X. Zhang, and Z. Li, "Multi-Attention Integrated Convolutional Network for Anomaly Detection of Time Series," in *2022 14th International Conference on Computer and Automation Engineering (ICCAE)*. Brisbane, Australia: IEEE, 2022, pp. 91–96.
- [135] X. Hong, J. Hu, T. Xu, X. Ren, F. Wu, X. Ma, and W. Li, "MagNet: Multilevel Dynamic Wavelet Graph Neural Network for Multivariate Time Series Classification," *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 1, pp. 1–22, Jan. 2025, publisher: Association for Computing Machinery (ACM).
- [136] G. Li, M. Ge, J. Wan, D. Han, M. Li, and M. Zhou, "Memmbaada: Memory-augmented state space model for multivariate time series anomaly detection," *Engineering Applications of Artificial Intelligence*, vol. 158, p. 111308, 2025.
- [137] A. Zhang and C. Li, "Adaptive state-space mamba for real-time sensor data anomaly detection," 2025. [Online]. Available: <https://arxiv.org/abs/2503.22743>
- [138] J. Jiang, F. Liu, Y. Liu, Q. Tang, B. Wang, G. Zhong, and W. Wang, "A dynamic ensemble algorithm for anomaly detection in iot imbalanced data streams," *Computer Communications*, vol. 194, pp. 250–257, 2022.
- [139] J. Jiang, B. Wang, Q. Tang, G. Zhong, X. Tang, and J. J. Rodrigues, "Incremental semi-supervised learning for data streams classification in internet of things," *IEEE Transactions on Network and Service Management*, 2025.
- [140] D. Wagner, A. Nair, B. J. Franks, J. Arweiler, A. Muraleedharan, I. Jungjohann, F. Hartung, M. C. Ahuja, A. Balinsky, S. Varshneya *et al.*, "Formally exploring time-series anomaly detection evaluation metrics," *arXiv preprint arXiv:2510.17562*, 2025.
- [141] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng *et al.*, "Unsupervised anomaly

- detection via variational auto-encoder for seasonal kpis in web applications,” in *Proceedings of the 2018 world wide web conference*, 2018, pp. 187–196.
- [142] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, “Deep autoencoding gaussian mixture model for unsupervised anomaly detection,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
 - [143] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, “Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding,” in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 387–395.
 - [144] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, “Unsupervised real-time anomaly detection for streaming data,” *Neurocomputing*, vol. 262, pp. 134–147, 2017.
 - [145] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*. Springer, 2001, vol. 208.
 - [146] A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet, “Anomaly detection in streams with extreme value theory,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1067–1075.
 - [147] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.
 - [148] F. Wilcoxon, “Individual comparisons by ranking methods,” in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.
 - [149] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” *Journal of the american statistical association*, vol. 32, no. 200, pp. 675–701, 1937.
 - [150] P. B. Nemenyi, *Distribution-free multiple comparisons*. Princeton University, 1963.
 - [151] H. Si, J. Li, C. Pei, H. Cui, J. Yang, Y. Sun, S. Zhang, J. Li, H. Zhang, J. Han *et al.*, “Timeseriesbench: An industrial-grade benchmark for time series anomaly detection models,” *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)*, pp. 61–72, 2024.
 - [152] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, “Precision and recall for time series,” in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., 2018, pp. 1924–1934.
 - [153] W.-S. Hwang, J.-H. Yun, J. Kim, and H. C. Kim, “Time-series aware precision and recall for anomaly detection: Considering variety of detection result and addressing ambiguous labeling,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2241–2244.
 - [154] W.-S. Hwang, J.-H. Yun, J. Kim, and B. G. Min, ““do you know existing accuracy metrics overrate time-series anomaly detections?”,” in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 403–412.
 - [155] A. Huet, J. M. Navarro, and D. Rossi, “Local evaluation of time series anomaly detection algorithms,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 635–645.
 - [156] R. Ghorbani, M. J. Reinders, and D. M. Tax, “Pate: Proximity-aware time series anomaly evaluation,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 872–883.
 - [157] J. Paparrizos, Y. Kang, P. Boniol, R. S. Tsay, T. Palpanas, and M. J. Franklin, “Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection,” *Proceedings of the VLDB Endowment*, vol. 15, no. 8, pp. 1697–1711, 2022.
 - [158] 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,” *Proceedings of the VLDB Endowment*, vol. 15, no. 11, pp. 2774–2787, 2022.
 - [159] C. Wang, L. Luo, and U. Aickelin, “Time Series Classification with Elasticity Using Augmented Path Signatures,” *ACM Transactions on Knowledge Discovery from Data*, vol. 19, no. 3, pp. 1–32, Apr. 2025, publisher: Association for Computing Machinery (ACM).
 - [160] Z. Z. Darban, Q. Wang, G. I. Webb, S. Pan, C. C. Aggarwal, and M. Salehi, “GenIAS: Generator for Instantiating Anomalies in time Series,” Feb. 2025, arXiv:2502.08262 [cs].
 - [161] C. Gong, C. Zhang, D. Yao, J. Bi, W. Li, and Y. Xu, “Causal Discovery from Temporal Data: An Overview and New Perspectives,” *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–38, Apr. 2025.
 - [162] J. Yoon, D. Jarrett, and M. Van der Schaar, “Time-series generative adversarial networks,” *Advances in neural information processing systems*, vol. 32, 2019.
 - [163] O. Hirschorn and S. Avidan, “Normalizing flows for human pose anomaly detection,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 13 499–13 508.
 - [164] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, “A time series is worth 64 words: Long-term forecasting with transformers,” in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
 - [165] N. Laptev, S. Amizadeh, and I. Flint, “Generic and Scalable Framework for Automated Time-series Anomaly Detection,” in *Proceedings of the 21th ACM SIGKDD*

- International Conference on Knowledge Discovery and Data Mining*. Sydney NSW Australia: ACM, 2015, pp. 1939–1947.
- [166] A. Beattie, P. Mulink, S. Sahoo, I. T. Christou, C. Kalalas, D. Gutierrez-Rojas, and P. H. J. Nardelli, “A robust and explainable data-driven anomaly detection approach for power electronics,” in *2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*. IEEE, Oct. 2022, pp. 296–301.
 - [167] C. Li, G. Feng, Y. Li, R. Liu, Q. Miao, and L. Chang, “DiffTAD: Denoising diffusion probabilistic models for vehicle trajectory anomaly detection,” *Knowledge-Based Systems*, vol. 286, p. 111387, 2024.
 - [168] H. Wang, Z. Luo, J. W. Yip, C. Ye, and M. Zhang, “Ecggan: A framework for effective and interpretable electrocardiogram anomaly detection,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5071–5081.
 - [169] R.-J. Hsieh, J. Chou, and C.-H. Ho, “Unsupervised online anomaly detection on multivariate sensing time series data for smart manufacturing,” in *2019 IEEE 12th conference on service-oriented computing and applications (SOCA)*. IEEE, 2019, pp. 90–97.
 - [170] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, “The ucr time series archive,” *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.
 - [171] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, “Wadi: a water distribution testbed for research in the design of secure cyber physical systems,” in *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks*, ser. CySWATER ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 25–28.
 - [172] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, “A dataset to support research in the design of secure water treatment systems,” in *Critical Information Infrastructures Security*, G. Havarneanu, R. Setola, H. Nas-sopoulos, and S. Wolthusen, Eds. Cham: Springer International Publishing, 2017, pp. 88–99.
 - [173] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao, and N. Tatbul, “Exathlon: a benchmark for explainable anomaly detection over time series,” *Proceedings of the VLDB Endowment*, vol. 14, no. 11, pp. 2613–2626, 2021.
 - [174] P. Wenig, S. Schmidl, and T. Papenbrock, “Anomaly detectors for multivariate time series: The proof of the pudding is in the eating,” in *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, 2024, pp. 96–101.
 - [175] —, “Timeeval: A benchmarking toolkit for time series anomaly detection algorithms,” vol. 15, no. 12, pp. 3678–3681, 2022.