Deep Learning for Time Series Anomaly Detection : A Survey

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. 2024.
ACM Comput. Surv. 57, 1, Article 15 (October 2024), 42 pages.

2024년 12월 6일

이규원

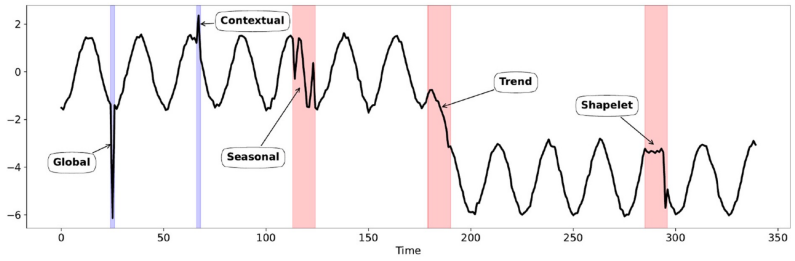
Department of Computer Science and Engineering
Chung-Ang University

Index

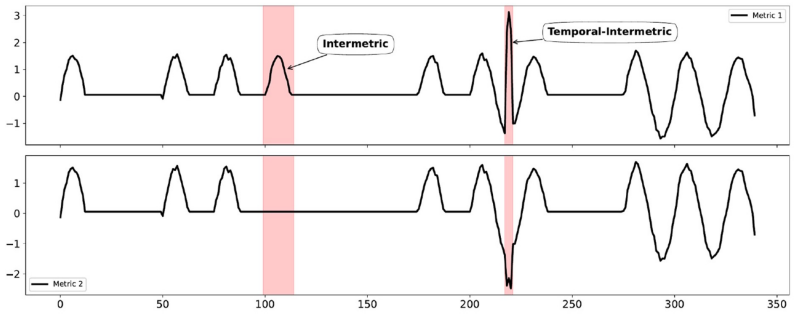
- **Time Series Anomaly Detection(TSAD)**
- **Anomalies in Time Series**
- **TSAD Methods**
  - Main Approaches
  - Aspects of the Models
  - Tables (Multivariate Time Series)
- **Public Dataset**
- **Evaluation Metrics**
- **Challenges**

Time Series Anomaly Detection

- **Time Series**
  - A series of data points indexed sequentially over time
- **Time Series Anomaly Detection**
  - Identifying patterns or events in time-ordered data that deviate significantly from expected behavior or norms



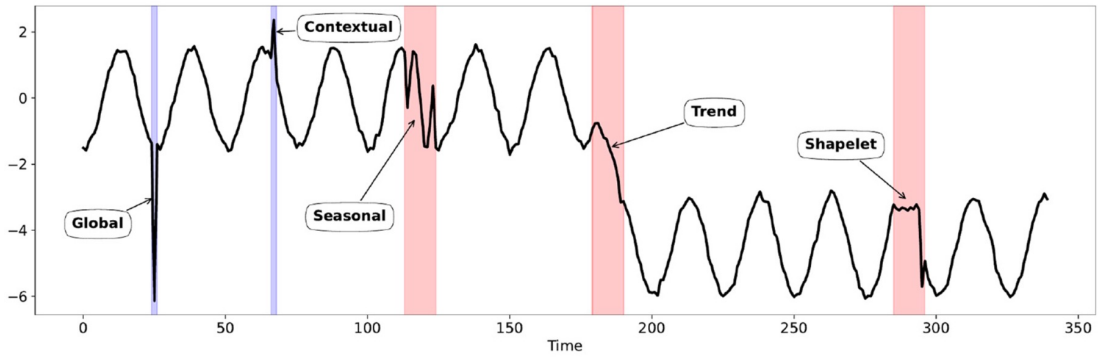
Univariate Time Series



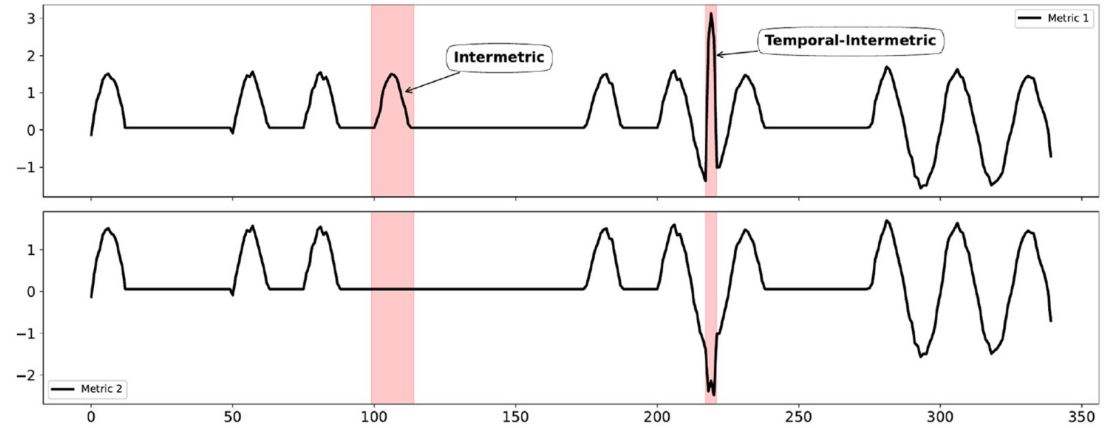
Multivariate Time Series

Anomalies in Time Series

- **Temporal(Time-wise)**

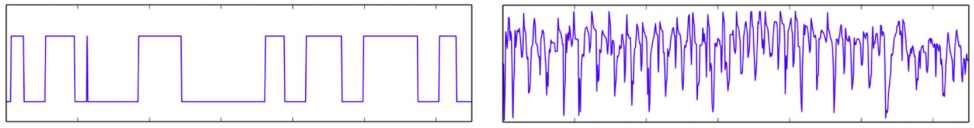


- **Spatial(Intermetric, Input-Tag, Feature-wise)**



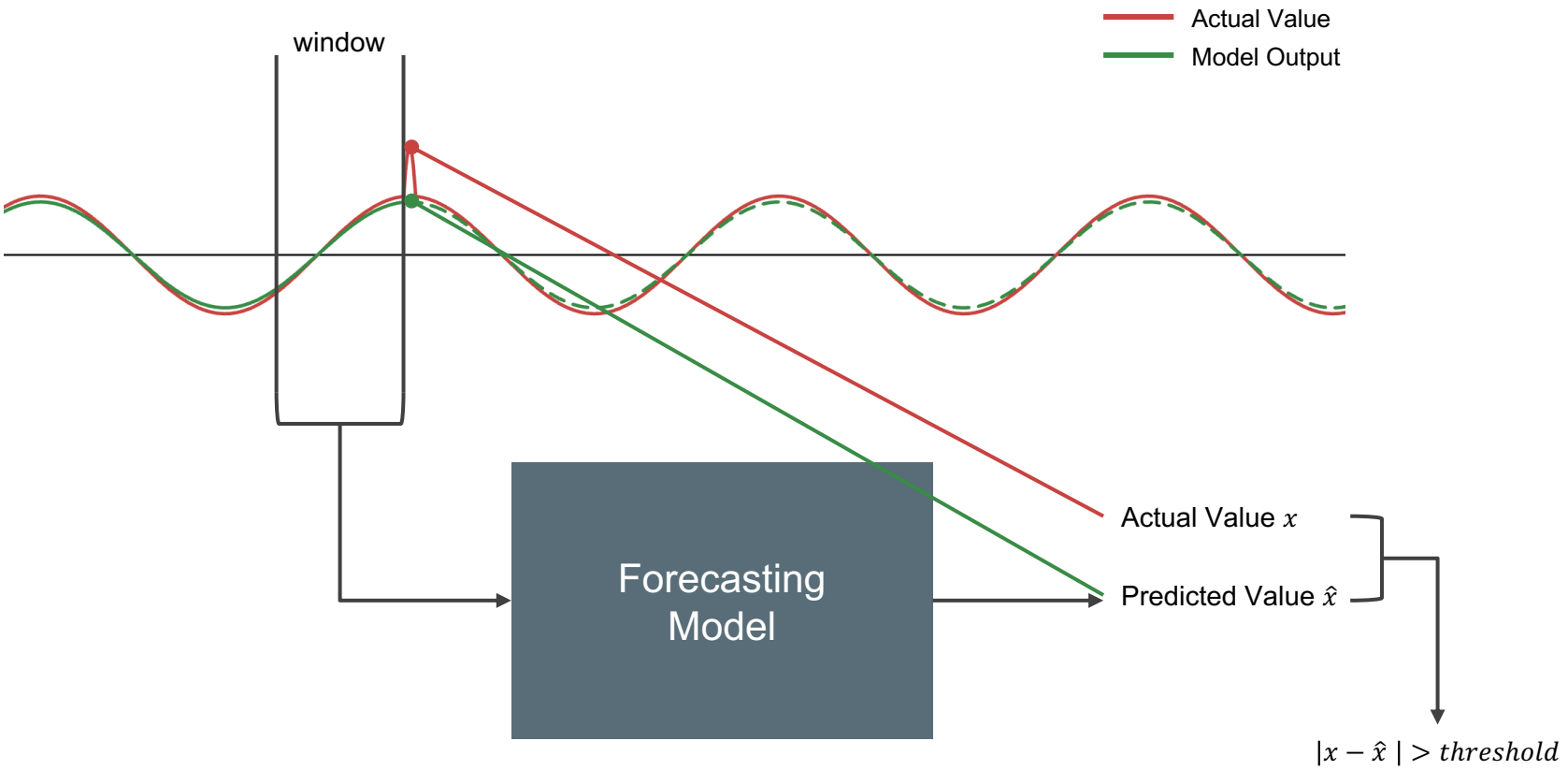
TSAD Methods: Main Approaches

- **Forecasting-based Models**
  - To **predict** a **future point** or subsequence **based on** a point or a **recent window**
- **Reconstruction-based Models**
  - To **reconstruct input** data
- **Representation-based Models**
  - To **learn** rich **representation** of time series data
- **Hybrid Models**
  - Forecasting || Reconstruction || Representation

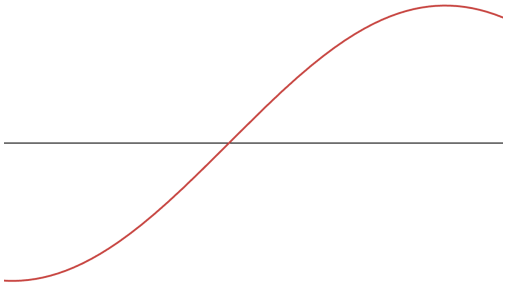


(a) Predictable (b) Unpredictable

TSAD Methods: Main Approaches

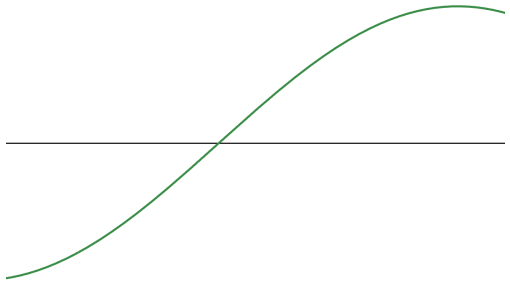


TSAD Methods: Main Approaches

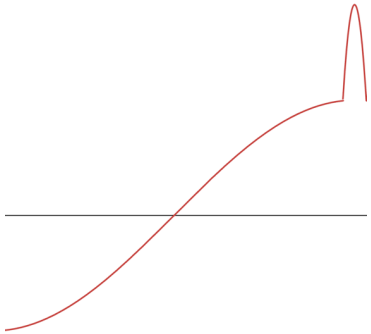


Window

Reconstruction Model



Reconstructed Window



Window

Reconstruction Model

Reconstructed Window

TSAD Methods: Aspects of the Models – 1

- **Temporal(Time) / Spatial(Feature)**
- **Learning Schemes**
  - Supervised: Normal, Anomaly label in train data
  - **Semi-Supervised: Normal label in train data**
  - Unsupervised: No label
  - Self-Supervised: Without explicit label, generate own supervisory signal
- **Input**
  - Point or **Subsequences(windows)**
  - **Representation** or Raw
- **Interpretability**
  - Identifying the dimension or feature that has the greatest impact on an anomaly

TSAD Methods: Aspects of the Models – 2

- **Point / Subsequence Anomaly**
  - Unusual point
  - Subsequence-level abnormality
- **Stochasticity**
  - To handle uncertainties in the inputs
  - ↔ Deterministic
- **Incremental**
  - Updates its knowledge with new data while preserving what it has learned from previous data

TSAD Methods: Forecasting

A[1]	MA[2]		Model		Year	T/S[3]	Su/Un[4]	Input[5]	Int[6]	P/S[7]	Stc[8]	Inc[9]	US[10]
Forecasting	RNN (3.2.1)	474	LSTM-PRED [66]	HASE	2017	T	Un	W	✓	–			
		1449	LSTM-NDT [92]	KDD	2018	T	Un	W	✓	Subseq			
		146	LGMAD [49]		2019	T	Semi	P		Point			✓
		297	THOC [156]	NeurIps	2020	T	Self	W		Subseq			✓
		61	AD-LTI [183]	TKDE	2020	T	Un	P		Point (frame)			
	CNN (3.2.2)	665	DeepAnt [135]	IEEE Access	2018	T	Un	W		Point + Subseq			
		201	TCN-ms [78]	JPCS	2019	T	Semi	W		Subseq			✓
		780	TimesNet [181]	ICLR	2023	T	Un	W		–			✓
	GNN (3.2.3)	922	GDN [45]	AAAI	2021	S	Un	W	✓	–			
		376	GTA* [34]	IEEE IoT Journal	2021	ST	Semi	–		–			
		92	GANF [40]	ICLR	2022	ST	Un	W					
	HTM (3.2.4)	53	RADM [48]	Sensors, MDPI	2018	T	Un	W		–			
	Transformer (3.2.5)	552	SAND [160]	AAAI	2018	T	Semi	W		–			
		376	GTA* [34]	IEEE IoT Journal	2021	ST	Semi	–		–			

TSAD Methods: Reconstruction

A[1]	MA[2]		Model		Year	T/S[3]	Su/Un[4]	Input[5]	Int[6]	P/S[7]	Stc[8]	Inc[9]	US[10]
Reconstruction	AE (3.3.1)	1461	AE/DAE [150]	MLSDA	2014	T	Semi	P		Point			
		2123	DAGMM [203]	ICLR	2018	S	Un	P		Point	✓		
		894	MSCRED [192]	AAAI	2019	ST	Un	W	✓	Subseq			
		837	USAD [10]	KDD	2020	T	Un	W		Point			
		58	APAE [70]	IJCAI	2020	T	Un	W		–			
		213	RANSynCoders [1]	KDD	2021	ST	Un	P	✓	Point		✓	
		59	CAE-Ensemble [22]	VLDB	2021	T	Un	W		Subseq			
		76	AMSL [198]	TKDE	2022	T	Self	W		–			
		11	ContextDA [106]	SIAM SDM	2023	T	Un	W		Point + Subseq			
	VAE (3.3.2)	115	STORN [159]	ICLR	2016	ST	Un	P		Point	✓		
		185	GGM-VAE [74]	PMLR	2018	T	Un	W		Subseq	✓		
		897	LSTM-VAE [143]	IEEE Robotics	2018	T	Semi	P		–	✓		
		1247	OmniAnomaly [163]	KDD	2019	T	Un	W	✓	Point + Subseq	✓		
		49	VELC [191]		2019	T	Un	–		–	✓		
		173	SISVAE [112]	TNNLS	2020	T	Un	W		Point	✓		✓
		166	VAE-GAN [138]	Sensors, MDPI	2020	T	Semi	W		Point	✓		✓
		98	TopoMAD [79]	TNNLS	2020	ST	Un	W		Subseq	✓		
		45	PAD [30]	Neural Computing	2021	T	Un	W		Subseq	✓		✓
		63	InterFusion [117]	KDD	2021	ST	Un	W	✓	Subseq	✓		
		82	MT-RVAE* [177]	Measurement	2022	ST	Un	W		–	✓		
		37	RDSMM [113]	TKDE	2022	T	Un	W		Point + Subseq	✓		✓
	GAN (3.3.3)	1142	MAD-GAN [111]	ICANN	2019	ST	Un	W		Subseq			
		286	BeatGAN [200]	IJCAI	2019	T	Un	W		Subseq			✓
		96	DAEMON [33]	ICDE	2021	T	Un	W	✓	Subseq			
		51	FGANomaly [54]	TKDE	2021	T	Un	W		Point + Subseq			
		84	DCT-GAN* [114]	TKDE	2021	T	Un	W		–			✓
	Transformer (3.3.4)	527	Anomaly Transformer [185]	ICLR	2021	T	Un	W		Subseq			
		84	DCT-GAN* [114]	TKDE	2021	T	Un	W		–			✓
		497	TranAD [171]	VLDB	2022	T	Un	W	✓	Subseq			✓
		82	MT-RVAE* [177]	Measurement	2022	ST	Un	W		–			
		4	Dual-TF [136]	WWW	2024	T	Un	W		Point + Subseq			✓

TSAD Methods: Representation, Hybrid

A[1]	MA[2]		Model		Year	T/S[3]	Su/Un[4]	Input[5]	Int[6]	P/S[7]	Stc[8]	Inc[9]	US[10]
Representation	Transformer (3.4.1)		TS2Vec [190]	549 AAAI	2022	T	Self	P		Point			✓
	CNN (3.4.2)	253	TF-C [196]	Neurlps	2022	T	Self	W		–			✓
		93	DCdetector [187]	KDD	2023	ST	Self	W		Point + Subseq			✓
		6	CARLA [42]	Pattern Recognition	2023	ST	Self	W		Point + Subseq			✓
		2	DACAD [43]		2024	ST	Self	W		Point + Subseq			
Hybrid	AE (3.5.1)	205	CAE-M [197]	TKDE	2021	ST	Un	W		Subseq			
		78	NSIBF* [60]	KDD	2021	T	Un	W		Subseq			
	RNN (3.5.2)	156	TAnoGAN [13]	IEEE SSCI	2020	T	Un	W		Subseq			
		78	NSIBF* [60]	KDD	2021	T	Un	W		Subseq			
	GNN (3.5.3)	567	MTAD-GAT [199]	ICDM	2020	ST	Self	W	✓	Subseq			
		76	FuSAGNet [76]	KDD	2022	ST	Semi	W		Subseq			

Public Dataset: Multivariate

Dataset/Benchmark	Real/Synth	MTS/UTS[1]	# Samples[2]	# Entities[3]	# Dim[4]	Domain
MSL [92]	Real	MTS	132,046	27	55	Aerospace
NAB-realAdExchange [5]	Real	MTS	9,616	3	2	Business
NAB-realAWSCloudwatch [5]	Real	MTS	67,644	1	17	Server machines monitoring
NASA Shuttle Valve Data [62]	Real	MTS	49,097	1	9	Aerospace
OPPORTUNITY [55]	Real	MTS	869,376	24	133	Computer networks
Pooled Server Metrics (PSM) [1]	Real	MTS	132,480	1	24	Server machines monitoring
PUMP [154]	Real	MTS	220,302	1	44	Industrial control systems
SMAP [92]	Real	MTS	562,800	55	25	Environmental management
SMD [115]	Real	MTS	1,416,825	28	38	Server machines monitoring
SWAN-SF [9]	Real	MTS	355,330	5	51	Astronomical studies
SWaT [129]	Real	MTS	946,719	1	51	Industrial control systems
WADI [7]	Real	MTS	957,372	1	127	Industrial control systems

Public Dataset: Univariate

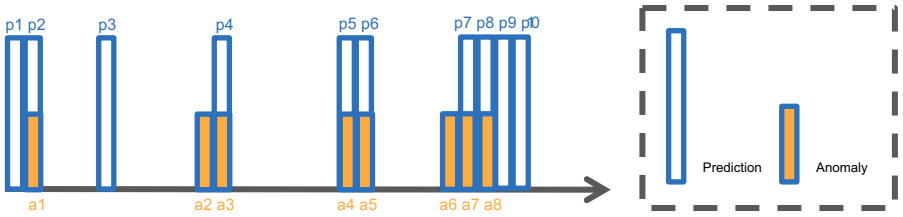
Dataset/Benchmark	Real/Synth	MTS/UTS[1]	# Samples[2]	# Entities[3]	# Dim[4]	Domain
NYC Bike [123]	Real	MTS/UTS	+25M	NA	NA	Urban events management
NYC Taxi [166]	Real	MTS/UTS	+200M	NA	NA	Urban events management
UCR [44]	Real/Synth	MTS/UTS	NA	NA	NA	Multiple domains
Dodgers Loop Sensor Dataset [55]	Real	UTS	50,400	1	1	Urban events management
KPI AIOPS [25]	Real	UTS	5,922,913	58	1	Business
MGAB [170]	Synth	UTS	100,000	10	1	Medical and health
MIT-BIH-LTDB [67]	Real	UTS	67,944,954	7	1	Medical and health
NAB-artificialNoAnomaly [5]	Synth	UTS	20,165	5	1	–
NAB-artificialWithAnomaly [5]	Synth	UTS	24,192	6	1	–
NAB-realKnownCause [5]	Real	UTS	69,568	7	1	Multiple domains
NAB-realTraffic [5]	Real	UTS	15,662	7	1	Urban events management
NAB-realTweets [5]	Real	UTS	158,511	10	1	Business
NeurIPS-TS [107]	Synth	UTS	NA	1	1	–
NormA [18]	Real/Synth	UTS	1,756,524	21	1	Multiple domains
Power Demand Dataset [44]	Real	UTS	35,040	1	1	Industrial control systems
SensoreScope [12]	Real	UTS	621,874	23	1	Internet of things (IoT)
Space Shuttle Dataset [44]	Real	UTS	15,000	15	1	Aerospace
Yahoo [93]	Real/Synth	UTS	572,966	367	1	Multiple domains

Evaluation Metrics

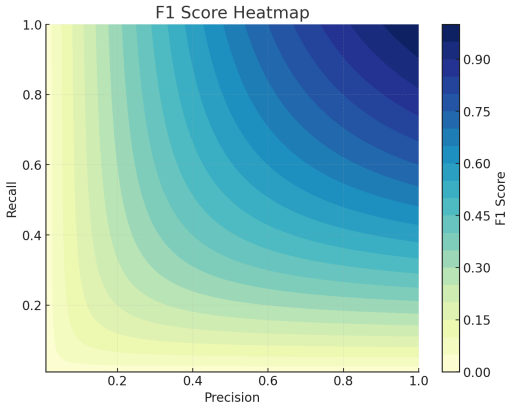
Definitions and Formulas		
Metric	**Definition**	**Formula***
Precision	The proportion of true positive results among all positive results predicted by the model. In time series anomaly detection, it indicates the accuracy of the detected anomalies.	$\text{Precision} = \frac{TP}{TP+FP}$
Recall	The proportion of true positive results among all actual positive cases. It measures the model's ability to detect all actual anomalies.	$\text{Recall} = \frac{TP}{TP+FN}$
F1 Score	The harmonic mean of precision and recall, providing a balance between the two metrics. It is useful when both precision and recall are important.	$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
F1$_{PA}$ Score	**F1$_{PA}$ score** is an F1 score that utilises a segment-based evaluation technique named point adjustment (PA), which at least one point within that segment is detected as abnormal [184]. This method can overestimate the performance of TSAD models (as mentioned in [42]).	$\text{F1}_{\text{PA}} = 2 \cdot \frac{\text{Precision}_{\text{PA}} \cdot \text{Recall}_{\text{PA}}}{\text{Precision}_{\text{PA}}+\text{Recall}_{\text{PA}}}$
PA%K	**F1$_{PA}$** is mitigated by employing a **PA%K** protocol [152] by focusing on segments of data w. A segment is correctly detected as anomalous if at least $K\%$ of its points are TP (TP_w).	$Accuracy_w = \begin{cases} 1 & \text{if } \frac{TP_w}{\lvert w \rvert} \geq \frac{K}{100} \\ 0 & \text{otherwise} \end{cases}$
AU-PR	Area Under Precision-Recall Curve is a performance measurement for classification problems at various thresholds (t). It is particularly useful for imbalanced datasets.	$\text{AU-PR} = \int_0^1 \text{Precision}(t) \frac{d(\text{Recall}(t))}{dt} \, dt$
AU-ROC	Area Under Receiver Operating Characteristic Curve represents the ability of the model to distinguish between classes based on different thresholds (t). A higher AU-ROC indicates better model performance.	$\text{AU-ROC} = \int_0^1 \text{Recall}(t) \frac{d(\text{FPR}(t))}{dt} \, dt$
MTTD	Mean Time to Detect is the average time taken to detect an anomaly at time T_{detect} after it occurs in time T_{true}. This metric evaluates the model's responsiveness.	$\text{MTTD} = \frac{1}{n} \sum_{i=1}^{n} (T_{\text{detect}} - T_{\text{true}})$
Affiliation	The affiliation metric assesses the degree of overlap between the detected anomalies (D) and the actual anomalies (A). It is designed to provide a more nuanced evaluation by considering both the precision and recall of the detected anomalies.	$\text{Affiliation} = \frac{\lvert D \cap A \rvert}{\lvert D \cup A \rvert}$
VUS	The Volume Under the Surface quantifies the volume between the true anomaly signal y and the predicted one, $\hat{y}$, over time. It captures both the temporal and amplitude differences between the two signals, providing a holistic measure of the detection performance.	$\text{VUS} = \int_0^T \lvert y_t - \hat{y}_t \rvert \, dt$

Evaluation Metrics: Conventional

- **Precision**
  - $\frac{True\ Positives}{True\ Positives + False\ Positives} = \frac{Correct\ Predictions}{Total\ Predictions}$
- **Recall**
  - $\frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{Detected\ Anomalies}{Total\ Anomalies}$
- **F1 Score**
  - $2 \times \frac{Precision \times Recall}{Precision + Recall}$
  - Evaluates the balance between precision and recall

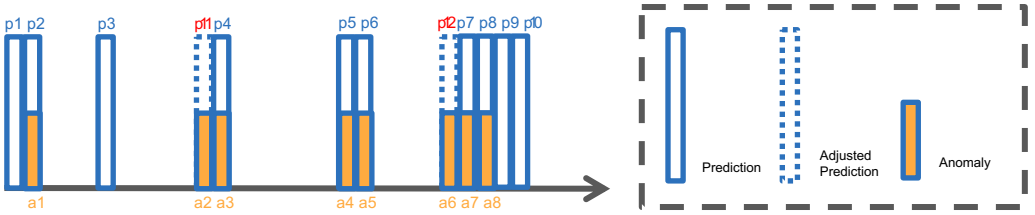


Precision = 0.6, Recall = 0.75, F1 Score = 0.667



Evaluation Metrics: Point Adjust

- **Point Adjust Precision & Recall**
  - Continuous anomalies are considered as a single anomaly segment
  - If even one point within an anomaly segment is detected, the entire segment is considered as detected
- **Point Adjust Precision**
  - $\frac{Correct\ Predictions}{Total\ Predictions}$ $(Predictions = Predictions + Adjusted\ Predictions)$
- **Point Adjust Recall**
  - $\frac{Detected\ Anomalies}{Total\ Anomalies}$

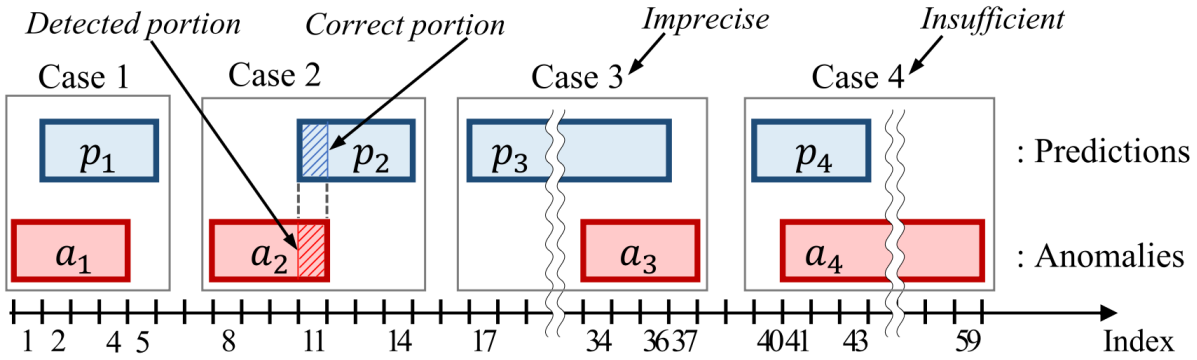


truth	0	0	1	1	1	0	0	1	1	1
score	0.6	0.4	0.3	0.7	0.6	0.5	0.2	0.3	0.4	0.3
point-wise alert	1	0	0	1	1	1	0	0	0	0
adjusted alert	1	0	1	1	1	1	0	0	0	0

PA Precision = 0.75, PA Recall = 1
F1 Score = 0.857

Evaluation Metrics: eTaPR

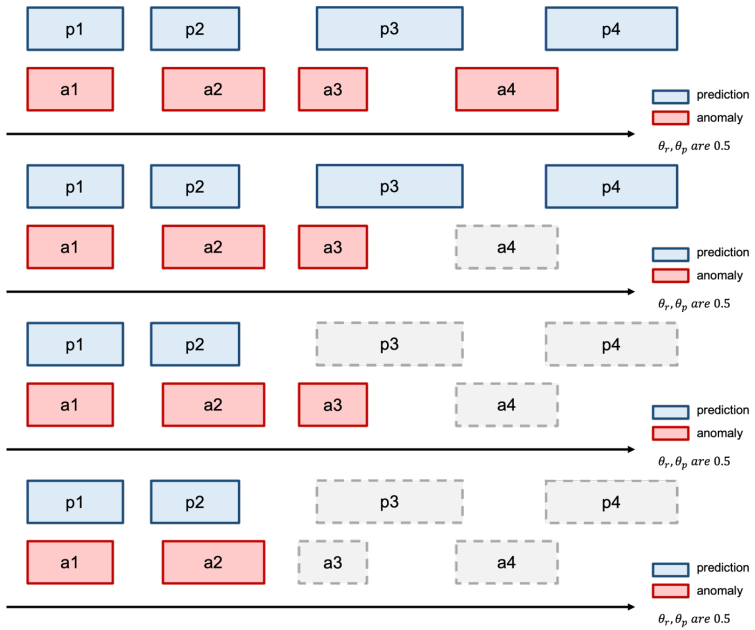
- **Enhanced Time-series Aware Precision & Recall (eTaPR)**
  - Imprecise or insufficient cases are not suitable for expert scenarios
  - Using cross-referencing, imprecise predictions and insufficiently detected anomalies are not scored
  - A weighting scheme is applied to penalize lengthy incorrect predictions



Evaluation Metrics: eTaPR

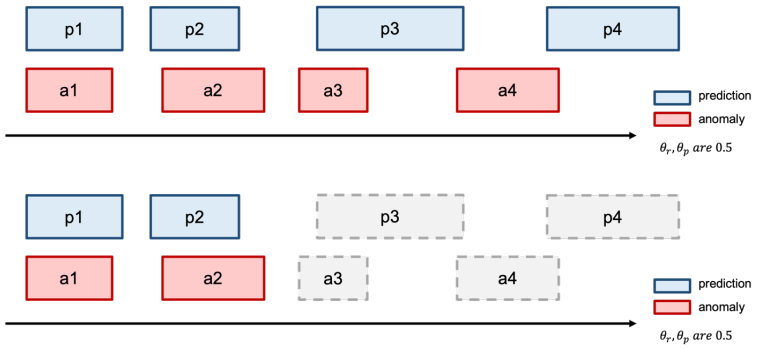
- **Cross referencing**
  - Refine cases of insufficiency or inaccuracy by cross-referencing between A^d and P^c
  - $A^d = \left\{a \mid a \in A \; and \; \frac{\sum_{p \in P^c} |a \cap p|}{|a|} \geq \theta_r\right\}$
  - $P^c = \left\{p \mid p \in P \; and \; \frac{\sum_{a \in A^d} |a \cap p|}{|p|} \geq \theta_p\right\}$
  - P $\rightarrow \boldsymbol{A^d(1)}$, $\boldsymbol{A^d(1) \rightarrow P^c(1)}$
  - $\boldsymbol{P^c(1) \rightarrow A^d(2)}$, $\boldsymbol{A^d(2) \rightarrow P^c(2)}$, $\boldsymbol{P^c(2) \rightarrow A^d(3)}$ …

    until $\boldsymbol{P^c(i) = P^c(i-1)}$ and $\boldsymbol{A^d(i) = A^d(i-1)}$



Evaluation Metrics: eTaPR

- **Enhanced Time-series Aware Precision**
  - $eTaP = \sum_{p \in P} \left(\frac{s^d(p) + s^d(p) \times s^p(p)}{2} \right) \times w_p$
  - $s^d(p) = \begin{cases} 1, & if\ p \in P^c \\ 0, & otherwise \end{cases}$ $s^p(p) = \frac{\sum_{a \in A^d} |a \cap p|}{|p|}$ $w_p = \frac{\sqrt{|p|}}{\sum_{q \in P} |q|}$ weighting scheme
- **Enhanced Time-series Aware Recall**
  - $eTaR = \frac{1}{|A|} \sum_{a \in A} \left(\frac{s^d(a) + s^d(a) \times s^p(a)}{2} \right)$
  - $s^d(a) = \begin{cases} 1, & if\ a \in A^d \\ 0, & otherwise \end{cases}$ $s^p(a) = \frac{\sum_{p \in P^c} |a \cap p|}{|a|}$



Challenges

- **Non-Stationary Data**
  - Incremental learning
- **Multivariate Time Series**
  - High dimensional time series (Many input tags)
- **Unlabeled Data**
  - Unsupervised, Semi-supervised, Self-supervised
  - Minimize false positive rate, Improve recall
- **Noise in the input data**
- **Interpretability**
- **Periodic Outlier Detection**
  - Appear at regular intervals but are rare compared to other patterns in the dataset
  - Not noise

Thank you