DRIFT: A Dual-Branch Time-Frequency Transformer for Time Series Anomaly Detection

Jiaxin Wu, Jiaxuan Xu, Lei Duan*
School of Computer Science, Sichuan University, Chengdu, China
{wujiaxin2, xujx}@stu.scu.edu.cn, leiduan@scu.edu.cn

***Abstract*—Although existing reconstruction-based methods have significantly improved state-of-the-art results for time series anomaly detection, they still struggle with distribution shifts in time series exhibiting complex patterns. To address the issue, this paper proposes a novel transformer-based model that is robust to distribution shifts, called Dual-bRanch tIme-Frequency Transformer (DRIFT). DRIFT proposes a dual-branch attention layer that focuses on the inherent periodic patterns in time series. The core principle is that even when distribution shifts occur, the periodic structure remains stable. Specifically, one branch employs Fourier transforms to extract global periodic patterns, while the other branch uses wavelet transforms to capture local periodic patterns. Finally, the contrastive structure learns pattern consistency across both branches, reinforcing the consistency of normal data while highlighting the distinctiveness of anomalies. Extensive experiments on real-world time series datasets demonstrate the effectiveness of DRIFT.**

***Index Terms*—time series, anomaly detection, time-frequency analysis**

I. Introduction

In critical realms like industrial monitoring [1], finance [2], [3], and healthcare [4]–[6], the widespread application of Internet of Things (IoT) and sensor technology has generated massive time series data. Meanwhile, in materials science, the massive time-series data that requires manual inspection is not only time-consuming but also prone to errors from human oversight. Time series anomaly detection [7]–[9] enables timely anomaly discovery and root cause analysis, helping to avoid unnecessary losses and ensuring the sustainability.

With deep learning advancing across domains, diverse frameworks have been applied to unsupervised time series analysis. The mainstream methods fall into two categories: prediction-based [10]–[12] methods and reconstruction-based [13]–[16] methods. For example, GDN [12] uses a graph attention-based prediction module to forecast future sensor behavior based on their graph-based neighbor information for anomaly detection in multivariate time series. TimesNet [14] converts 1D time series into 2D tensors, uncovering complex temporal patterns for precise reconstruction of time series. Both share a similar core principle—learning normal data patterns and detecting anomalies by comparing the “expected output of normal patterns” with the “actual observed values”.

*Corresponding author
This work was supported in part by the Advanced Materials-National Science and Technology Major Project (Grant No.2024ZD0607700) and the National Natural Science Foundation of China (Grant No.62472294).

However, two mainstream methods both face two key challenges. **Time Series Distribution Shift:** Patterns learned from training data perform poorly or even fail entirely on test data. Prediction errors or reconstruction errors, calculated point-by-point, fail to capture temporal context and even overfit, degrading detection performance. **Complex Time Series Patterns:** Real-world time series often exhibit complex patterns, incorporating periodic patterns while simultaneously being influenced by noise and abrupt changes. Consequently, real-world time series cannot be described by a single, uniform stationary paradigm. Extracting fine-grained information from these complex structures is a significant challenge.

To address these challenges, this paper proposes **DRIFT**. For distribution shifts, DRIFT designs a transformer-based dual-branch attention layer to capture both global and local periodic patterns. Then, a contrastive structure is introduced to enable mutual learning of these representations, reinforcing the consistency of normal patterns. Finally, as the global-local periodicity discrepancy in anomalies exceeds that of normal data, periodicity pattern discrepancy replaces traditional reconstruction error as the primary detection metric. Concurrently, periodicity patterns constitute inherent temporal attributes unaffected by distribution shifts, alleviating the impact on detection. For fine-grained temporal information, DRIFT employs Fast Fourier Transform (FFT) and Continuous Wavelet Transform (CWT) to focus on global and local periodicity respectively, covering periodic patterns of varying granularities. Meanwhile, it combines the base attention with a periodicity distance matrix, enabling the extracted information to align with periodic patterns while preserving random fluctuations, avoiding overfitting in purely data-driven approaches. The main contributions of this paper are as follows:

- We propose a novel method, DRIFT, which innovatively uses the discrepancy between global and local periodicity features as the metric for anomaly detection, effectively reducing the impact of distribution shifts.
- For complex time series patterns, DRIFT employs Fourier and Wavelet transforms to capture periodic patterns at varying granularities. It combines these with base attention to preserve the random fluctuations in time series.
- Compared to 11 existing anomaly detection methods, extensive experiments conducted on real-world time series datasets demonstrate the effectiveness of DRIFT.

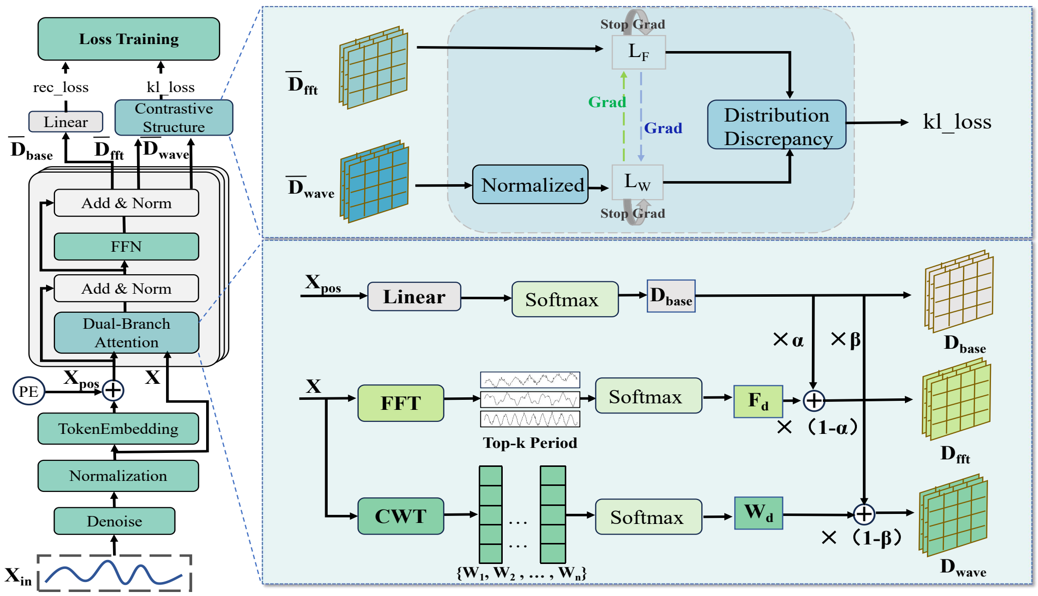


Fig. 1. Overall architecture of DRIFT (left). Details of dual-branch attention and contrastive structure (right).

II. Related Work

Time series anomaly detection has long been a subject of academic research, yielding a variety of approaches. Among these, deep learning-based methods have gained the most prominence, encompassing both prediction-based and reconstruction-based methods. The majority of time series anomaly detection methods employ some form of deep neural network. MTAD-GAT [10] explicitly models inter-variable correlations and temporal dependencies through a Graph Attention Network (GAT), propagating information via a Gated Recurrent Unit (GRU) network. DCdetector [17] is a multi-scale contrastive learning model employing an asymmetric dual-attention module to mitigate information loss during patch segmentation. DConAD [18] employs a Transformer-based architecture that generates differential data to provide supplementary information, thereby enhancing the modeling of normal patterns within time series data. MSCRED [19] detects anomalies across multiple scales using a multi-scale feature matrix, while integrating an LSTM network to capture long-term temporal dependencies in multi-scale features. To capture comprehensive information, this paper employs a Transformer-based model.

Recent research increasingly employs frequency-domain information to complement time-domain insights, with the Fourier transform being the most prevalent time-frequency conversion method. FreTS [20] is a time-series forecasting approach based on MLPs, centered on domain transformation and frequency-domain learning. It converts time-domain multivariate sequences into complex-form frequency-domain representations via the Fourier transform, learning specific dependencies of frequency components through a reconstructive MLP. MtsCID [21] employs a dual-network architecture to learn both intra-variable temporal dependencies and inter-variable relationships. Both components utilize the Fourier transform to achieve feature transformation and optimization through time-frequency domain conversion. However, these approaches neglect the wavelet transform's ability to capture frequency-domain information.

III. Methodology

Consider the multivariate time series data containing T timestamps as $X = (x_1, x_2, \dots, x_T)$, where each observation $x_t \in \mathbb{R}^d$ is a d-dimensional vector at time step t. The univariate scenario is a special case of the multivariate scenario, where $d = 1$. We need to predict its corresponding anomaly label sequence $Y = (y_1, y_2, \dots, y_T)$. Here, $y_t \in \{0, 1\}$, where $y_t = 1$ indicates that the data point is an anomaly, and $y_t = 0$ indicates that the data point is normal.

A. Overall Architecture

Fig. 1 illustrates the overall architecture of DRIFT. DRIFT adopts a "Fourier-Wavelet Dual-Branch Fusion" framework, primarily comprising three main stages: data processing, dual-branch attention, contrastive structure.

1. **Data Processing**: To eliminate noise caused by distribution shifts and ensure the cleanliness of inputs, input X_{in} undergoes denoising and RevIN normalization. Then the data is encoded into high-dimensional vectors via an embedding layer and integrated with positional encoding to learn temporal positional information. Subsequently, it is fed into the encoder alongside the output from the normalization layer to capture fine-grained periodic patterns and original dependencies.

2. **Dual-Branch Attention**: This is the core of the encoder. It simultaneously employs Fourier and wavelet transforms to focus on periodic patterns from shallow to deep levels. The Fourier branch captures global periodic features, while the wavelet branch focuses on local periodic features. The combination of both enables the model to be robust to periodic patterns at varying granularities. Simultaneously, the attention

layer learns base attention by processing inputs that fuse positional information, thereby capturing inherent temporal dependencies and random fluctuations. By separately merging periodic features with this base attention through learnable weights, the model extracts comprehensive information from complex temporal patterns.

3. **Contrastive Structure**: It calculates the Kullback–Leibler (KL) divergence between global periodicity and local periodicity to determine the discrepancy between the two distributions. During training, the model employs two loss functions: one aligns the distributions to reinforce consistency in normal data, while the other prevents overfitting of discrepancy. The global and local periodicity of normal data are inherently consistent, but anomalies disrupt this coherence. Consequently, periodicity distribution discrepancy can be leveraged to detect anomalies. Furthermore, periodicity constitutes an intrinsic property of time series, unaffected by distributional shifts. This ensures the detection criterion remains robust against distribution drifts.

B. Data Processing

In this stage, the model first performs denoising on the input time-series data X_{in}. It filters out minor frequency components in the frequency domain while preserving significant frequencies, ensuring the removal of noise unrelated to the core periodicity. Concurrently, local smoothing is applied to eliminate high-frequency random noise, accurately retaining valid local features. Then normalization is performed to stabilize the temporal distribution without discarding original temporal characteristics. Subsequently, the data is fed into the embedding layer to fuse positional encodings into high-dimensional vectors, enabling the model to perceive temporal sequence and prevent the loss of original relative positional information. Finally, these high-dimensional vectors X_{pos} are input into the encoder alongside the output X from the normalization layer. After processing through these layers, the original data is stripped of redundancy and noise while retaining core information and periodic characteristics, effectively enhancing the model's detection efficacy.

C. Dual-Branch Attention

Traditional transformer attention learns dependencies in time-series data but struggles to distinguish core periodic patterns from noise and random fluctuations. Furthermore, lacking periodic guidance and being purely data-driven, attention scores decay with distance, making long-range correlations easily obscured by short-range dependencies. Hence, dual-branch attention introduces two periodic pattern capturing branches, focusing respectively on global coarse-grained periodicity and local fine-grained periodicity, covering both long-range patterns and short-range details. However, relying solely on periodic patterns risks model homogenization, compromising adaptability to real-world time series. Thus, both the Fourier branch (FB) and wavelet branch (WB) are divided into two stages: "feature extraction" and "feature fusion". After periodic feature extraction, features are fused with the base attention to complement each other. This approach preserves core original features while ensuring extraction of multi-granularity periodic patterns.

Regarding FB, during feature extraction, each feature of every sample without additional position information undergoes a Fourier transform along the time dimension, converting the time-domain signal into a frequency-domain signal:

$$X_{\text{f}}[b, f, d] = \sum_{t=0}^{T-1} X_{\text{in}}[b, t, d] \cdot e^{-i \cdot \frac{2\pi f t}{T}}. \tag{1}$$

The output $X_{\text{f}} \in \mathbb{R}^{B \times T \times D}$ represents the frequency domain signal, where f is the frequency index $(0 \leq f < T)$ denoting different frequency components. B, T, and D correspond to sample size, time step, and feature dimension, respectively. Since more pronounced fluctuations exhibit greater amplitude, once the amplitude for each frequency component is obtained, the strongest k frequencies can be identified. This yields the top-k principal periods of the time-series signal. Based on the extracted top-k global periods, the branch measures the shortest distance between any two positions within the periods, yielding the global period distance matrix F_d, which indicates coarse-grained periodic correlations. At this stage, the distance matrix F_d lacks the original true features and requires additional raw dependency information. During feature fusion, the temporal features incorporating positional information capture temporal dependencies through a linear projection layer, reducing the computational complexity of dot product operations to prevent gradient explosion. Then the projected information undergoes dot product computations to derive an attention distribution matrix, which is then scaled to prevent vanishing of gradients. Following normalization of the distribution matrix, the base attention distribution is obtained. The FB combines the base attention distribution D_{base} with the distribution F_d through learnable weight α:

$$D_{\text{fft}} = D_{\text{base}} \times \alpha + F_d \times (1 - \alpha), \tag{2}$$

where D_{fft} denotes global periodic dependencies.

Regarding WB, during feature extraction, for each feature of every sample without additional position information, the original signal is convolved with wavelet functions of varying scales to obtain wavelet coefficients:

$$W = \int_{-\infty}^{\infty} X_{\text{in}}[b, t, d] \cdot \psi^* \left(\frac{t - \tau}{s} \right) dt, \tag{3}$$

where s is the scale parameter, τ is the shift parameter, and ψ^* is the complex conjugate of the mother wavelet, ensuring real-valued amplitude results. For the obtained wavelet coefficients W, extract their amplitudes. As the amplitude of wavelet coefficient W reflects local feature intensity, local fine-grained periodic values $cycle$ can be confirmed within different time-frequency windows based on amplitude, with robustness processing applied:

$$\text{cycle} = \max(\text{cycle}, 1). \tag{4}$$

Based on $cycle$ and the relative positions of any two locations, the branch derives the local period distance matrix W_d to

TABLE I

ANOMALY DETECTION RESULTS ON MSL, PSM, NAB, AND UCR DATASETS. ALL RESULTS ARE IN %, THE BEST ONES ARE IN BOLD.

Methods		MSL			PSM			NAB			UCR		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
LSTM-NDT	[KDD-18]	62.88	99.99	77.21	80.21	78.44	79.32	64.00	66.67	65.31	52.31	82.94	52.31
DAGMM	[ICLR-18]	73.63	**100.00**	84.82	96.79	90.71	93.65	76.22	72.92	74.53	53.37	97.18	53.37
MSCRED	[AAAI-19]	81.75	92.16	86.64	**99.98**	43.75	71.87	84.21	66.67	74.42	54.41	99.99	70.48
CAE-M	[AAAI-19]	77.51	99.98	87.33	71.09	43.75	71.87	79.18	80.19	79.68	69.81	99.99	69.81
MERLIN	[ICDM-20]	26.13	46.65	33.45	39.27	32.16	35.36	80.13	72.62	76.19	75.42	80.18	75.42
MTAD-GAT	[ICDM-20]	79.17	98.06	87.68	74.04	92.28	82.13	84.21	72.72	78.04	77.56	99.72	69.81
GDN	[ICDM-20]	89.08	99.98	94.22	70.38	83.49	76.37	81.29	78.72	79.98	68.94	99.88	68.94
AnoTran	[ICLR-22]	91.95	96.50	94.17	96.74	97.73	97.23	94.18	99.99	97.00	60.41	99.99	73.08
DCdetector	[KDD-23]	92.08	94.44	93.24	97.40	98.16	97.78	93.65	97.68	96.03	61.62	99.99	74.05
MtsCID	[WWW-25]	93.37	96.97	95.13	97.57	98.51	98.54	85.98	87.78	86.12	60.08	96.52	71.85
DConAD	[IJCNN-25]	91.68	99.30	95.34	96.00	98.17	97.07	89.50	90.08	91.64	63.79	94.05	70.31
DRIFT (Ours)		**94.55**	98.17	**96.33**	97.82	**99.53**	**98.67**	**95.57**	**100.00**	**97.74**	**77.79**	**100.00**	**87.51**

TABLE II

AVERAGE RANKING OF ANOMALY DETECTION METHODS (LOWER VALUE INDICATES BETTER PERFORMANCE). THE BEST RESULT IS IN BOLD.

DRIFT (Ours)	DCdetector	MtsCID	DConAD	AnoTran	GDN	MTAD-GAT	CAE-M	MERLIN	DAGMM	MSCRED	LSTM-NDT
1.0000	3.7500	3.7500	4.2500	5.0000	6.2500	7.5000	8.5000	8.5000	9.2500	9.2500	11.0000

measure the local spatio-temporal detail correlation between any two positions.

$$W_d[i,j] = \min\big(|\text{pos}_i - \text{pos}_j|, \text{cycle} - |\text{pos}_i - \text{pos}_j|\big), \tag{5}$$

where i and j represent any two positions, and pos_i and pos_j denote their relative positions. During feature fusion, the WB similarly uses the base attention distribution D_{base} with the distribution W_d, controlling importance through learnable weight β:

$$D_{\text{wave}} = D_{\text{base}} \times \beta + W_d \times (1-\beta), \tag{6}$$

where D_{wave} emphasizes local periodic dependencies. Following subsequent processing by the encoder, the D_{fft} and D_{wave} are output as $\overline{D}_{\text{fft}}$ and $\overline{D}_{\text{wave}}$, serving as inputs for the subsequent contrastive structure.

D. Contrastive Structure

The core principle of DRIFT is that normal points exhibit a consistent pattern in both global and local periodic features, whereas anomalous points fail to maintain this consistency. To enhance the consistency of normal data and make anomalous data more conspicuous for more efficient anomaly detection, DRIFT proposes a contrastive structure. This structure introduces two loss functions, L_F and L_W, tailored to the two periodic distributions of the input:

$$\begin{aligned} L_F &= \text{KL}(\overline{D}_{\text{fft}}, \text{Stop}(\overline{D}_{\text{wave}})) + \text{KL}(\text{Stop}(\overline{D}_{\text{wave}}), \overline{D}_{\text{fft}}), \\ L_W &= \text{KL}(\overline{D}_{\text{wave}}, \text{Stop}(\overline{D}_{\text{fft}})) + \text{KL}(\text{Stop}(\overline{D}_{\text{fft}}), \overline{D}_{\text{wave}}), \end{aligned} \tag{7}$$

where $\text{KL}(\cdot,\cdot)$ denotes the Kullback–Leibler divergence, which calculates the distance between two representations to reflect similarity between global and local periodic representations. The $\text{Stop}(\cdot)$ operation (referred to as ‘stop gradient’) enables asynchronous training of the two branches.

Specifically, after normalizing local periodicity distribution to eliminate numerical fluctuations, L_F aligns the fine-grained periodicity distribution of the wavelet branch with the global periodicity distribution of the Fourier branch, preventing excessive fragmentation of fine-grained periodicity and ensuring consistency across multi-granularity periodicities. L_F aims to minimize the divergence between two distributions, reinforcing periodic consistency in normal data. However, due to overfitting, a single loss function fails to yield optimal results. Consequently, L_W is employed to prevent the global periodicity from overly constraining the fine-grained periodicity, preserving the wavelet branch's ability to capture random fluctuations and preventing the homogenization of varying-granularity periodicity. The combined training of both losses ultimately yields the global-local periodicity discrepancy kl_loss. Simultaneously, during training, the periodicity discrepancy in anomalies proves harder to reduce and more prone to amplification, making the disruption to consistency caused by anomalies increasingly pronounced.

E. Anomaly Criterion

After obtaining the global-local periodicity discrepancy, based on the core principle that normal data exhibits consistent periodic patterns while anomalies disrupt such consistency, the model employs periodicity discrepancy as the primary criterion to detect anomalies. Concurrently, the encoder output undergoes linear projection and is compared with the original time series X_{in} to yield a reconstruction error rec_loss. This serves as a supplementary measure to the periodicity discrepancy, yielding an anomaly score:

$$\text{AS}(X) = \text{KL}(\overline{D}_{\text{fft}}, \overline{D}_{\text{wave}}) + \text{KL}(\overline{D}_{\text{wave}}, \overline{D}_{\text{fft}}) + \text{MSE}(X_{\text{in}}, X_{\text{out}}), \tag{8}$$

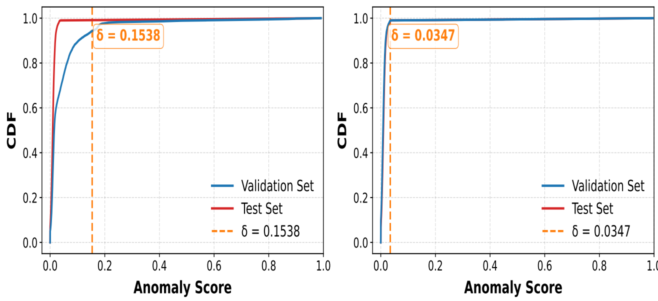


Fig. 2. CDF of anomaly scores on PSM validation and test sets for TimesNet (Left) and DRIFT (Right).

where $\mathrm{MSE}(\cdot,\cdot)$ denotes the mean squared error, and X_{out} denotes the sequence output. When the anomaly score of a sample exceeds the set threshold δ, the sample is classified as anomalous. The detected label y_t can be expressed as:

$$y_t = \begin{cases} 1 & \text{if } \mathrm{AS}(x_t) \geq \delta \\ 0 & \text{if } \mathrm{AS}(x_t) < \delta \end{cases} . \tag{9}$$

IV. EXPERIMENTS

A. Experimental Setup

We use four publicly real time-series datasets, including MSL [11], PSM [22], NAB [23], and UCR [24]. Among these, MSL and PSM are multivariate datasets, while NAB and UCR (UCR comprises multiple sub-datasets, and we employed a subset of them) are univariate datasets. We select precision, recall, and F_1 score as evaluation metrics. Results are compared with methods including LSTM-NDT [11], DAGMM [25], MSCRED [19], CAE-M [26], MERLIN [27], MTAD-GAT [10], GDN [12], Anomaly Transformer [28] (AnoTran for short), DCdetector [17], MtsCID [21], and DConAD [18].

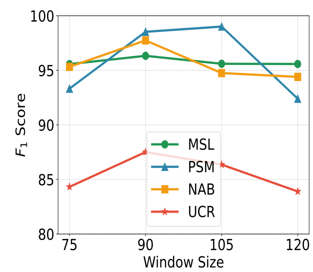
The Adam optimizer [29] is used to train the model with a learning rate of 0.0001 and a batch size of 256 over 10 training epochs. The Transformer is configured with 3 layers. The threshold δ is pre-determined by identifying r of the data as anomalies. Specifically, $r = 1.0\%$ is set for MSL and PSM, 0.5% for UCR, and 0.7% for NAB. All experiments are conducted in a PyTorch 2.7.1 environment using a single NVIDIA GeForce RTX 4090 GPU.

B. Main Results

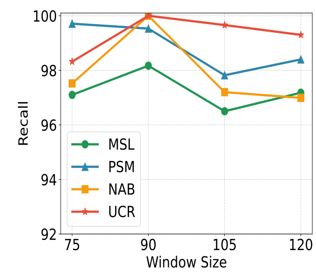
We evaluate the DRIFT model against eleven competitive baselines across four datasets (see Table I). We select precision, recall, and F_1 score as evaluation metrics to fairly assess the performance of anomaly detection methods. On both the NAB and UCR datasets, DRIFT outperforms the other 11 baseline methods. On the MSL and PSM datasets, DRIFT maintains optimal performance on two metrics while achieving excellent levels on the remaining metrics. Conducting significance tests, DRIFT achieves an average ranking of 1.0000, indicating a statistically significant difference in the rankings across methods and demonstrating DRIFT's significant superiority (see Table II). Then we perform a case study using the PSM dataset. As illustrated in Fig. 2, cumulative distribution function (CDF) of TimesNet on the validation and test sets show a clear gap caused by shifts. However, CDF generated by DRIFT on the validation and test sets are always similar. This experiment substantiates that our criterion can effectively alleviate time series distribution shifts. Extensive experiments conducted across real-world time series datasets demonstrate the effectiveness of DRIFT.

TABLE III
ABLATION RESULTS OF DRIFT. THE PRECISION (P), RECALL (R), AND F_1 SCORE (F_1) ARE IN %.

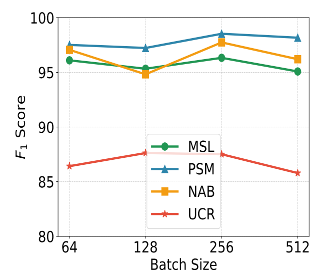
Datasets	Metrics	DRIFT	w/o FB	w/o WB	w/o contrast
MSL	P	**94.55**	92.67	93.24	74.88
	R	**98.17**	88.69	95.93	90.42
	F_1	**96.33**	91.58	94.05	85.64
PSM	P	**97.82**	96.35	95.38	94.89
	R	**99.53**	96.54	97.50	96.27
	F_1	**98.53**	96.94	96.85	96.51
NAB	P	**95.57**	94.53	93.50	92.91
	R	**100.00**	99.99	99.99	99.99
	F_1	**97.74**	96.98	96.64	96.59
UCR	P	**77.79**	74.79	75.61	73.55
	R	**100.00**	99.99	99.99	99.99
	F_1	**87.51**	85.58	85.91	84.76



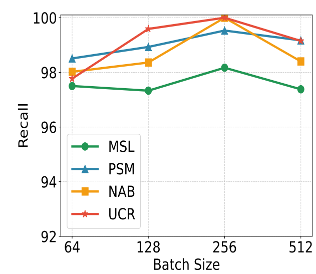
(a) Window size (F_1 score)



(b) Window size (Recall)



(c) Batch size (F_1 score)



(d) Batch size (Recall)

Fig. 3. Hyper-parameter studies of DRIFT.

C. Ablation Studies

To verify the impact of capturing only coarse-grained or fine-grained periodic patterns on anomaly detection, we conducted ablation experiments on the dual-branch attention. The ablation experiments in Table III highlight the importance of capturing multi-granularity periodic patterns, as removing either the FB or WB leads to deterioration in detection performance. Concurrently, through ablation experiments on the contrastive structure, we validated the effectiveness of employing periodicity discrepancy as the criterion in alleviating the impact of distribution shifts.

D. Hyper-Parameter Sensitivity Analysis

Fig. 3 illustrates the results of the model using different hyperparameters. Specifically, the experiment primarily investigates the impact of different window sizes and batch sizes on DRIFT. The window sizes and batch sizes are selected from the sets [75, 90, 105, 120] and [64, 128, 256, 512], respectively. Results in Fig. 3 (a) and (b) indicate that DRIFT achieves optimal performance at a window size of 90. Both excessively large or small window sizes lead to decreased F_1 scores and recall rates. Results in Fig. 3 (c) and (d) indicate that DRIFT achieves peak detection performance at a batch size of 256. When the batch size increases or decreases beyond this threshold, performance on the four datasets exhibits varying degrees of decline.

V. CONCLUSION

This paper proposes DRIFT, which employs a dual-branch attention to extract varying-granularity periodicity in time series exhibiting complex patterns. Then it uses a contrastive structure to differentiate periodicity distributions, replacing traditional reconstruction errors with global-local periodicity discrepancy to alleviate the impact of distribution shifts. Experimental results across real-world datasets demonstrate the model's effectiveness in detecting anomalies on both univariate and multivariate time-series data. Meanwhile, in materials science, it enables researchers to analyse time-series data more efficiently. In future work, we will adapt DRIFT to other time series tasks, such as time series prediction and classification.

REFERENCES

[1] M. A. Belay, A. Rasheed, and P. S. Rossi, "Self-supervised modular architecture for multi-sensor anomaly detection and localization," in *Proc. IEEE Conf. Artif. Intell.*, 2024, pp. 1278–1283.

[2] C. Fjellström, "Long short-term memory neural network for financial time series," in *Proc. IEEE Int. Conf. Big Data*, 2022, pp. 3496–3504.

[3] T. Choi, D. Lee, Y. Jung, and H.-J. Choi, "Multivariate time-series anomaly detection using seqvae-cnn hybrid model," in *2022 International Conference on Information Networking (ICOIN)*. IEEE, 2022, pp. 250–253.

[4] A. Ukil, S. Bandyoapdhyay, C. Puri, and A. Pal, "Iot healthcare analytics: The importance of anomaly detection," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, 2016, pp. 994–997.

[5] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, "The challenges of modeling and forecasting the spread of covid-19," *Proc. Natl. Acad. Sci.*, vol. 117, no. 29, pp. 16 732–16 738, 2020.

[6] S. Liu, B. Zhou, Q. Ding *et al.*, "Time series anomaly detection with adversarial reconstruction networks," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4293–4306, 2023.

[7] J. Paparrizos, P. Boniol, Q. Liu, and T. Palpanas, "Advances in time-series anomaly detection: Algorithms, benchmarks, and evaluation measures," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. V.2*, 2025, pp. 6151–6161.

[8] 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 *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. V.2*, 2025, pp. 4623–4634.

[9] D. Bhattacharya, S. Mukherjee, C. Kamanchi, V. Ekambaram, A. Jati, and P. Dayama, "Towards unbiased evaluation of time-series anomaly detector," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2025, pp. 1–5.

[10] 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," in *Proc. IEEE Int. Conf. Data Min.*, 2020, pp. 841–850.

[11] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2018, pp. 387–395.

[12] A. Deng and B. Hooi, "Graph neural network-based anomaly detection in multivariate time series," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35. AAAI, 2021, pp. 4027–4035.

[13] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: deep transformer networks for anomaly detection in multivariate time series data," in *Proc. VLDB Endow.*, 2022, pp. 1201–1214.

[14] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–23.

[15] K. Kingsbury and P. Alvaro, "Elle: inferring isolation anomalies from experimental observations," *Proc. VLDB Endow.*, vol. 14, no. 3, pp. 268–280, 2020.

[16] P. Boniol, J. Paparrizos, T. Palpanas, and M. J. Franklin, "Sand: Streaming subsequence anomaly detection," *Proc. VLDB Endow.*, vol. 14, no. 10, pp. 1717–1729, 2021.

[17] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proc. ACM SIGKDD*, 2023, pp. 3033–3045.

[18] W. Zhang, X. Lin, W. Yu, G. Yao, J. Zhong, and Y. Li, "Dconad: A differencing-based contrastive representation learning framework for time series anomaly detection," in *Proc. IEEE Int. Joint Conf. Neural Netw.* IEEE, 2025.

[19] 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 *Proc. AAAI Conf. Artif. Intell.*, vol. 33. AAAI, 2019, pp. 1409–1416.

[20] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, H. He, D. Lian, N. An, L. Cao, and Z. Niu, "Frequency-domain mlps are more effective learners in time series forecasting," in *Proc. Conf. Neural Inf. Process. Syst.*, 2023.

[21] Y. Xie, H. Zhang, and M. Babar, "Multivariate time series anomaly detection by capturing coarse-grained intra- and inter-variate dependencies," in *Proc. ACM Web Conf.*, 2025, pp. 697–705.

[22] A. Abdulaal, Z. Liu, and T. Lancewicki, "Practical approach to asynchronous multivariate time series anomaly detection and localization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2021, pp. 2485–2494.

[23] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms-the numenta anomaly benchmark," in *Proc. Int. Conf. Mach. Learn. Appl.*, 2015, pp. 38–44.

[24] R. Wu and E. J. Keogh, "Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2421–2429, 2021.

[25] 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 *Proc. Int. Conf. Learn. Represent.*, 2018.

[26] Y. Zhang, Y. Chen, J. Wang, and Z. Pan, "Unsupervised deep anomaly detection for multi-sensor time-series signals," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[27] T. Nakamura, M. Imamura, R. Mercer *et al.*, "Merlin: Parameter-free discovery of arbitrary length anomalies in massive time series archives," in *Proc. IEEE Int. Conf. Data Min.*, 2020, pp. 1190–1195.

[28] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv Preprint arXiv:2110.02642*, 2021.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Preprint arXiv:1412.6980*, 2014.