Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079644

ARTICLE

Hierarchical Cyber–Physical Symbiosis with Bidirectional State Space Modeling for IIoT Anomaly Diagnosis

Kelan Wang[1] **and Jianfei Chen**[2,*]

[1]School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China
[2]College of Electronic and Optical Engineering and College of Flexible Electronics (Future Technology), Nanjing University of Posts and Telecommunications, Nanjing, China

*Corresponding Author: Jianfei Chen. Email: chenjf@njupt.edu.cn

Received: 25 January 2026; Accepted: 03 March 2026; Published: 08 May 2026

ABSTRACT: As 6G-enabled Industrial Internet of Things (IIoT) evolves, green and sustainable industrial monitoring increasingly relies on edge AI to deliver low-latency diagnosis under tight resource constraints. Industrial cyber–physical systems increasingly rely on heterogeneous sensing and communication infrastructures, where network-side attacks can propagate into physical processes and appear as coupled anomalies. Reliable diagnosis therefore requires joint learning from time-synchronized cyber and physical telemetry rather than modeling them as independent signals. This paper develops Cyber–Physical Symbiosis Network (CPSNet), a model designed for edge-AI deployment with a dual-stream architecture for fixed-window multiclass cross-domain anomaly diagnosis in IIoT. CPSNet encodes each modality into hierarchical multi-resolution features and refines them with a Multi-Scale Bidirectional-Recursive (MSBR) block. MSBR couples multi-kernel temporal convolutions with a gated bidirectional state space pathway, capturing transient irregularities while retaining long-range context within the window. Cross-modal dependency is injected at every scale by a symbiosis module that performs bidirectional channel-wise gating and holistic state space fusion to learn unified cross-modal dynamics efficiently. A compact multi-scale pooling head with auxiliary modality supervision preserves discriminative evidence in both streams. On the DataSense benchmark, CPSNet achieves 97.18% Accuracy and 99.04% AUC on Multiclass-8, and 89.07% Accuracy and 94.28% AUC on Multiclass-50, showing consistent improvements over single-modality and multi-modal baselines. Ablation and efficiency analyses further suggest complementary gains from multi-scale refinement and explicit coupling with a favorable accuracy–runtime trade-off. These results suggest that hierarchical cross-modal coupling with state space temporal modeling can improve robust, fine-grained IIoT diagnosis for 6G edge-AI monitoring.

KEYWORDS: Industrial Internet of Things; cyber–physical anomaly diagnosis; state space model; multi-scale temporal modeling; network traffic analysis

1 Introduction

Green and sustainable 6G-enabled Industrial Internet of Things (IIoT) is expected to support massive connectivity, ultra-reliable low-latency communication, and pervasive sensing for industrial automation. At the same time, it must meet stringent energy and carbon constraints through edge intelligence and efficient data-plane operation [1]. IIoT infrastructures have become a core substrate for industrial automation, enabling fine-grained monitoring and closed-loop control via heterogeneous sensing, networking, and edge/cloud computing. In such 6G IIoT settings, anomaly diagnosis is increasingly deployed as an edge AI service to reduce backhaul traffic and operational energy while preserving real-time responsiveness.

However, increased connectivity and protocol diversity also enlarge the attack surface and tighten the coupling between cyber incidents and physical processes [2]. As a result, industrial monitoring is increasingly confronted with cyber–physical anomaly diagnosis: abnormal behaviors may be manifested in network traffic such as reconnaissance, exploitation, and malware propagation, while their consequences and correlates appear in physical telemetry like process deviations and abnormal device dynamics [3,4]. Accurate diagnosis therefore requires learning from time-synchronized dual-stream observations and reasoning about their coupled temporal evolution rather than treating cyber and physical evidence as independent signals [5].

A growing body of work has explored deep learning for anomaly detection in IIoT and cyber–physical systems using either cyber traffic or physical telemetry. Cyber-only approaches typically rely on packet/flow statistics and protocol indicators to model anomalous traffic patterns, whereas physical-only approaches focus on sensor time series to identify deviations from nominal dynamics [6]. Nevertheless, cyber–physical anomalies are often subtle and temporally structured across modalities: a traffic-side deviation may induce delayed or weak physical responses, and a physical deviation may be preceded by low-intensity cyber probing. This motivates dual-stream models that preserve modality-specific evidence while enabling cross-modal interaction under strict temporal co-registration, which has begun to be supported by aligned benchmarks such as DataSense [2,7].

Despite this progress, current window-based diagnosis pipelines still exhibit practical limitations that constrain reliable and efficient deployment in green 6G IIoT monitoring. Temporal encoding is frequently implemented as either convolution-dominated extraction or global sequence mixing, which makes it difficult to represent transient irregularities and window-level evolution within a compact fixed-window encoder. This gap limits discrimination when anomaly evidence is distributed across multiple time scales and needs to be captured without deep stacks or heavy attention. Besides temporal modeling, cross-modal dependency is commonly introduced by late fusion or coarse interaction, which provides limited capability for one modality to recalibrate the other as context under heterogeneous noise and missingness. Such weak coupling can degrade robustness when discriminative evidence is dominant in a single stream and requires cross-stream confirmation. Moreover, cross-modal coupling is often applied at a single resolution, while cyber–physical correlations can be scale-dependent and may emerge differently across shallow and deep representations. Without injecting coupling along the hierarchy, complementary evidence can be diluted before aggregation, reducing the benefit of dual-stream observations under fine-grained label settings.

To address these gaps, this paper proposes Cyber–Physical Symbiosis Network (CPSNet), a dual-stream, multi-scale framework for cyber–physical anomaly diagnosis over synchronized windowed sequences. CPSNet strengthens per-stream temporal modeling under fixed windows by integrating multi-kernel local extraction with window-level state space contextualization and explicit gating, enabling compact representation of mixed time-scale signatures. It further enforces cyber–physical dependency through bidirectional cross-context alignment and holistic state space fusion, so that each modality can be calibrated by the other under a shared window-level context. In addition, CPSNet injects cyber–physical coupling across multiple encoder scales before scale-wise pooling, which preserves complementary evidence across resolutions and stabilizes fine-grained diagnosis. In addition, a three-term objective combines task supervision with modality-specific auxiliary losses to preserve discriminative evidence in both branches and stabilize cross-modal learning. Importantly, CPSNet exploits bidirectional contextual dependencies within each fixed window to strengthen representation learning, without claiming strict online causal constraints. The contributions of this paper are summarized as follows:

- A Multi-Scale Bidirectional-Recursive (MSBR) block is proposed for single-modality temporal modeling, which couples a multi-scale convolution bank with a state space context via an explicit gating

mechanism inside a residual structure, enabling efficient sequence modeling over fixed windows while capturing both transient irregularities and window-level long-range evolution.
- A Cyber–Physical Symbiosis (CPS) module is developed to strengthen cross-modal dependency modeling, consisting of symmetric Cyber–Physical Alignment (CPA) blocks for bidirectional cross-context gating and a Holistic State Space Fusion (HSSF) block with Bi-Mamba-based interaction and modality-wise gated injections for controlled fusion.
- A Multi-Scale Feature Aggregation Head is designed to inject CPS at each encoder scale before global pooling, followed by a compact multi-layer perceptron (MLP) classifier, and the training objective integrates task loss with cyber/physical auxiliary supervision to encourage modality-wise discriminativeness alongside fused diagnosis.

The remainder of this paper is organized as follows. Section 2 reviews recent deep learning approaches for IIoT intrusion/anomaly detection and cyber–physical representation learning, with an emphasis on temporal modeling and state space-based sequence encoders. Section 3 formulates the cyber–physical anomaly diagnosis problem and describes the DataSense benchmark. Section 4 details the proposed CPSNet, including MSBR, CPS (CPA and HSSF), the multi-scale aggregation head, and the loss function. Section 5 presents experimental settings and quantitative results, followed by ablation and analysis. Section 6 discusses the advantage of CPSNet according to the experimental results. Section 7 concludes the paper and outlines future directions.

2 Related Works

2.1 Deep Learning for Cyber-Physical Anomaly Diagnosis

Cyber–physical anomaly diagnosis in IIoT systems aims to identify and categorize abnormal behaviors by jointly exploiting time-synchronized cyber observations such as packet/flow statistics and protocol indicators and physical observations like sensor/Message Queuing Telemetry Transport (MQTT) telemetry and device-state logs [8]. Compared with single-stream settings, the key difficulty is that anomaly evidence can be distributed across modalities and manifest with different temporal characteristics, which motivates dual-stream learning frameworks with explicit cross-modal interaction under strict window-level co-registration, as supported by aligned benchmarks such as DataSense [2].

Existing studies span cyber-only intrusion/anomaly detection using Convolutional Neural Network (CNN)/Recurrent Neural Network (RNN)/Transformer variants on traffic-derived features [5,6,9] and physical-side fault/anomaly diagnosis using deep feature extractors for sensor time series under non-stationary regimes [10–12]. Recent cyber–physical approaches further integrate both streams through hybrid pipelines or structured representations, including hybrid monitoring that combines system-state and traffic cues [7] and graph-based CPS anomaly/intrusion analysis [13]. Overall, these works suggest that strong within-modality temporal modeling and explicit cross-modal coupling are both essential. Simple late concatenation or decision-level fusion can be insufficient when discriminative evidence is unevenly distributed across modalities.

2.2 State Space Models and Mamba

State space models (SSMs) parameterize sequence dynamics through latent state evolution and readout, offering linear-time recurrence and strong inductive bias for long-range dependency modeling. Structured SSMs enable efficient implementations that scale favorably with sequence length compared with attention-based Transformers, making them attractive for industrial temporal data with long-horizon dependencies.

Representative advances include HiPPO-based history compression [14], S4 structured state spaces [15], and simplified variants such as S5 [16]. Mamba further introduces selective state spaces with input-dependent parameterization, enabling content-adaptive sequence modeling while retaining linear-time scaling [17]. Related engineering efforts on IO efficiency, such as FlashAttention, contextualize practical trade-offs between attention and alternative long-sequence backbones [18,19]. Recent extensions apply Mamba-style designs to multivariate time series and robust classification settings [20], indicating that SSM/Mamba backbones are well-suited for window-based cyber–physical diagnosis where efficient long-context modeling is beneficial.

All related models in Section 2 are concluded in Table 1.

Table 1: Summary of surveyed models in Section 2.

Category	Strengths	Weaknesses
Cyber only [5,6,9]	Temporal feature	No physical semantics
Physical only [10–12]	Sensor feature	No cyber semantics
Streams fusion [7]	Semantics integration	Coupling coarse
Graph-based CPS [13]	Models relations	Construction cost
Transformer [18,19]	Attention implementation	Quadratic in length
Structured SSM [14–16]	Efficient long-context	Parameter sensitivity
Selective SSM [17,20]	Input adaptive	Task-level coupling design

3 Problem Formulation

DataSense is collected from a realistic IIoT testbed comprising diverse industrial sensors and common IoT devices interconnected through a dual-band Wi-Fi access point, a managed switch, and a centralized MQTT broker hosted on a Raspberry Pi [2]. DataSense provides two time-synchronized data sources collected from the same IIoT testbed: packet traces captured via continuous monitoring (cyber stream), and IIoT sensor logs collected from the MQTT broker and indexed in the logging backend (physical stream) [2]. Let $\Delta = 10$ s be the dataset-defined temporal resolution, and denote the t-th non-overlapping window as

$$\mathcal{W}_t = [t\Delta, (t+1)\Delta). \tag{1}$$

For each $\mathcal{W}_t$, all raw cyber packets and physical sensor records whose timestamps fall within the window are collected and sorted by time. A lightweight feature extraction procedure maps the raw observations in each window into fixed-length vectors $\mathbf{x}_t^c \in \mathbb{R}^{d_c}$ and $\mathbf{x}_t^p \in \mathbb{R}^{d_p}$. The concrete parsing, encoding, normalization, and split protocol are specified in Section 5.1. This yields an aligned cyber–physical instance $(\mathbf{x}_t^c, \mathbf{x}_t^p)$ directly constructed from the Canadian Institute for Cybersecurity (CIC) IIoT 2025 traces.

Benign data are recorded under normal device operation without interference, producing 12 h of benign traffic. For evaluation, a one-hour benign subset is selected, with an initial 5-min profiling segment reserved for device profiling and excluded from the evaluation set. Anomalous data are produced through controlled execution of 50 realistic attacks spanning seven major categories: reconnaissance (Recon), denial of service (DoS), distributed denial of service (DDoS), web exploitation (Web), man-in-the-middle (MITM) and spoofing, Bruteforce, and malware (Mirai).

This work studies cyber–physical anomaly diagnosis as supervised classification over aligned cyber–physical sequences under a fixed-window setting. Given a history length L, the input to CPSNet at time t is

$$\mathbf{X}_t^c = [\mathbf{x}_{t-L+1}^c; \ldots; \mathbf{x}_t^c] \in \mathbb{R}^{L \times d_c}, \qquad \mathbf{X}_t^p = [\mathbf{x}_{t-L+1}^p; \ldots; \mathbf{x}_t^p] \in \mathbb{R}^{L \times d_p}. \tag{2}$$

The learning objective is to estimate a parametric mapping

$$\hat{\mathbf{y}}_t = f_\theta\left(\mathbf{X}_t^c, \mathbf{X}_t^p\right), \tag{3}$$

where $\mathbf{y}_t$ denotes the ground-truth anomaly label under a specified granularity. In DataSense, two standard label sets are defined: multiclass-8 (benign and seven anomaly categories corresponding to attack families), and multiclass-50 (benign and 49 fine-grained anomaly types corresponding to specific attacks). In this paper, inference is performed on a fixed window $\mathbf{X}_t$ and the model is allowed to exploit bidirectional contextual dependencies within the window. This section defines the task formulation and notation, while the next section specifies how $\mathbf{x}_t^c, \mathbf{x}_t^p$ and the aligned sequences are instantiated in practice. The proposed dual-stream architecture operates on $\mathbf{X}_t^c$ and $\mathbf{X}_t^p$ using modality-specific convolutional stems and hierarchical refinement, while cross-stream interaction modules require strict temporal co-registration at each stage; therefore, the dataset processing pipeline must preserve window-level synchronization and scale-consistent alignment.

Table 2 summarizes the key notations used in this paper.

Table 2: Table of mathematical notations.

Notation	Description
Δ	Window duration defined by the dataset, $\Delta = 10$ s
$\mathcal{W}_t$	The t-th non-overlapping time window $[t\Delta, (t+1)\Delta)$
L	Sequence length in windows for fixed-window inference
d_c, d_p	Feature dimensions of cyber and physical window-level vectors
$\mathbf{x}_t^c \in \mathbb{R}^{d_c}$	Cyber window-level feature vector extracted from $\mathcal{W}_t$
$\mathbf{x}_t^p \in \mathbb{R}^{d_p}$	Physical window-level feature vector extracted from $\mathcal{W}_t$
$\mathbf{X}_t^c \in \mathbb{R}^{L \times d_c}$	Cyber input sequence formed by the last L windows up to t
$\mathbf{X}_t^p \in \mathbb{R}^{L \times d_p}$	Physical input sequence formed by the last L windows up to t
$\mathbf{y}_t$	Ground-truth anomaly label at time t
$f_\theta(\cdot)$	CPSNet with parameters θ
$\hat{\mathbf{y}}_t$	Predicted label distribution produced by $f_\theta(\mathbf{X}_t^c, \mathbf{X}_t^p)$
$\hat{y}$	Predicted class probability vector from the classifier head
K	Number of classes in the classification setting
$\mathcal{L}_{\text{task}}$	Task loss for the fused prediction
$\mathcal{L}_{\text{cyber}}, \mathcal{L}_{\text{physical}}$	Auxiliary losses for cyber and physical branches
$\mathcal{L}$	Overall training objective

4 Methodology

4.1 Overview

This work targets cyber–physical anomaly diagnosis from synchronized cyber–physical observations, where network-traffic streams and physical-sensor streams exhibit both modality-specific temporal patterns and strong cross-modal dependency. The proposed CPSNet shown in Fig. 1 follows a dual-stream design: each modality is first encoded into a hierarchical set of multi-resolution features, and the resulting multi-scale representations are then fused and aggregated into a compact representation for anomaly-label prediction.

Two principles guide the design. First, temporal modeling should be efficient and lightweight for fixed-window diagnosis and practical deployment. Second, cyber and physical cues should be coupled explicitly rather than being merged only at the end, so that cross-modal context can calibrate modality-specific evidence at multiple scales. In particular, CPSNet operates on a fixed-length window and is allowed to exploit bidirectional contextual dependencies within the window to strengthen representation learning.

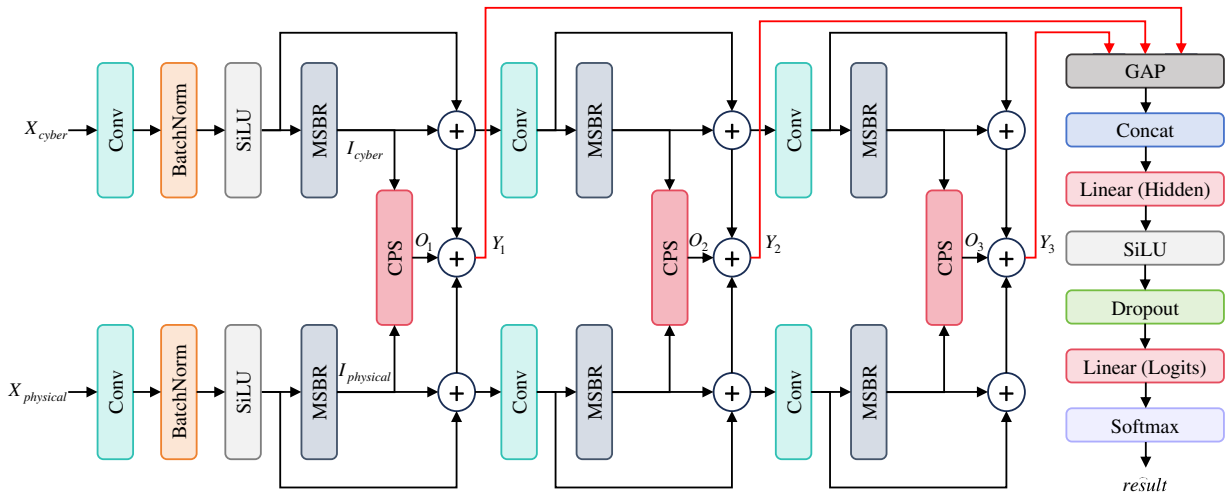


Figure 1: Architecture of the proposed CPSNet for cyber–physical anomaly diagnosis.

Given aligned windowed inputs from the cyber stream and the physical stream, the CPSNet encoder produces three scale-specific feature pairs $\{(Y_s^{\text{cyber}}, Y_s^{\text{physical}})\}_{s=1}^{3}$, where $Y_s^{(\cdot)} \in \mathbb{R}^{B\times L_s\times D_s}$. These features are fed to the multi-scale aggregation head. For each scale, a lightweight adapter is applied to normalize the feature space; the highest-resolution branch uses a stronger Conv–BatchNorm–SiLU adapter, while the others use a lighter Conv-based adapter as shown in Fig. 1. Each modality is then refined by the proposed MSBR block to capture transient local irregularities and long-range evolution within the fixed window. After modality-wise refinement, the CPS module is invoked to inject cross-context information and to produce a fused representation that explicitly encodes cyber–physical interaction at that scale. Finally, the fused representations from all scales are globally pooled, concatenated, and projected by a compact MLP to obtain the anomaly-diagnosis logits.

Formally, the end-to-end inference can be summarized as

$$\tilde{Y}_s^m = \text{MSBR}\big(\phi_s(Y_s^m)\big), \quad Z_s = \text{CPS}\left(\tilde{Y}_s^{\text{cyber}}, \tilde{Y}_s^{\text{physical}}\right), z = \text{Concat}\left(\text{GAP}(Z_1), \text{GAP}(Z_2), \text{GAP}(Z_3)\right), \tag{4}$$

$$\hat{y} = \text{Softmax}\left(\text{Linear}_{\text{logits}}\left(\text{Dropout}\left(\text{SiLU}\left(\text{Linear}_{\text{hid}}(z)\right)\right)\right)\right), \tag{5}$$

where $\phi_s(\cdot)$ denotes the scale-specific adapter, MSBR performs single-modality temporal refinement, and CPS performs cyber–physical coupling and fusion, GAP reflects to global average pooling. During training, the overall objective combines the task loss on $\hat{y}$ with modality-specific auxiliary supervision in Section 4.5 to encourage both branches to preserve discriminative evidence rather than relying solely on late fusion.

4.2 Multi-Scale Bidirectional-Recursive Block

The MSBR block is a single-modality temporal modeling unit for streams. Recursive refers to the latent state update of the bidirectional state space pathway, which is computed as a discrete-time recurrence, while the coupling between the local and contextual paths is realized by gated residual fusion rather than an

additional recursion across branches. Its design targets the coexistence of transient local irregularities and long-range system evolution in cyber–physical anomaly diagnosis sequences. As shown in Fig. 2, MSBR adopts a two-branch structure: a parallel multi-scale convolution path for local pattern extraction and a state space context path for window-level contextualization. The two paths are coupled through a subtractive context interaction followed by a sigmoid gate inside a residual structure.

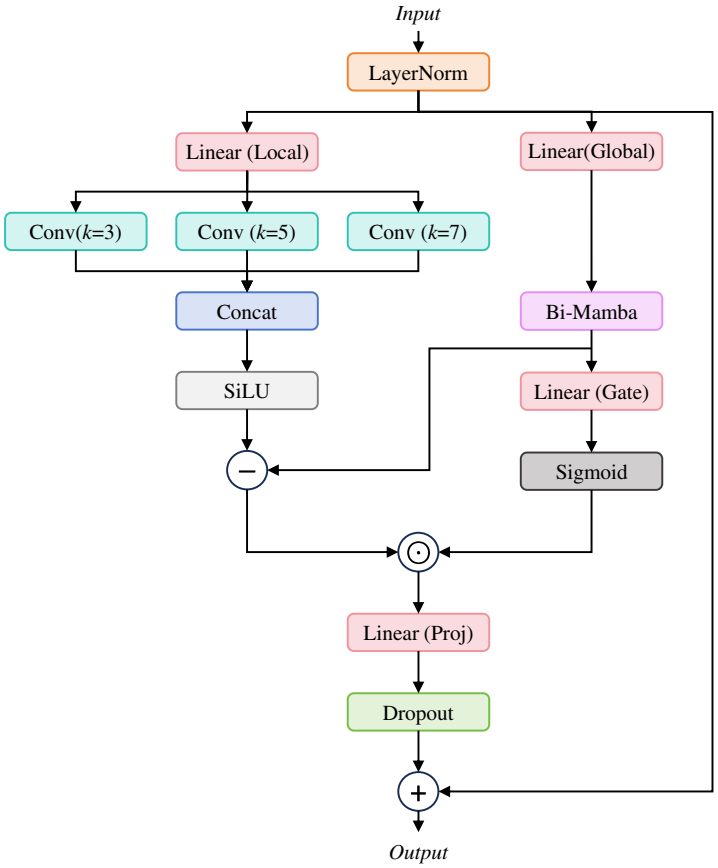


Figure 2: The proposed MSBR block.

Pure Temporal Convolutional Network (TCN) and pure convolutional temporal encoders are efficient for short-lived irregularities, but they often require deep stacks or carefully tuned dilations to cover long-range evolution within a window. This increases complexity and may still under-represent global context. Lightweight Transformer variants offer flexible global mixing, but attention typically scales quadratically with window length and can be sensitive to noisy or missing industrial telemetry. MSBR therefore combines a shallow multi-kernel convolution bank to capture transient local cues with a linear-time bidirectional SSM pathway to model window-level evolution. A subtractive interaction and a sigmoid gate then emphasize deviations from contextual trends in a compact form.

Let the input feature tensor be $X \in \mathbb{R}^{B\times L\times D}$, where B is the batch size, L is the window length, and D is the channel dimension. MSBR first applies LayerNorm to obtain $\tilde{X} = \mathrm{LN}(X) \in \mathbb{R}^{B\times L\times D}$, and then projects $\tilde{X}$ into two branches:

$$X^{\ell} = \mathrm{Linear}_{\ell}(\tilde{X}) \in \mathbb{R}^{B\times L\times D_h}, \qquad X^{g} = \mathrm{Linear}_{g}(\tilde{X}) \in \mathbb{R}^{B\times L\times D_h}, \tag{6}$$

where D_h is the hidden width used for branch interaction. Here, Linear$(\cdot)$ denotes a per-time-step affine projection that mixes channels and maps features into branch-specific subspaces.

Three 1D convolutions with different kernel sizes are applied to X^{ℓ}:

$$H_k = \mathrm{Conv}_k(X^{\ell}) \in \mathbb{R}^{B\times L\times D_k}, \qquad k \in \{3,5,7\}, \tag{7}$$

and the outputs are concatenated along the channel dimension and activated to form the local representation:

$$H_{\mathrm{local}} = \mathrm{SiLU}(\mathrm{Concat}(H_3, H_5, H_7)) \in \mathbb{R}^{B\times L\times D_h}, \qquad D_h = D_3 + D_5 + D_7. \tag{8}$$

Concat$(\cdot)$ concatenates features along the channel dimension to fuse multi-kernel evidence, and SiLU$(\cdot)$ provides a smooth nonlinearity for stabilizing local feature extraction.

The global branch applies a bidirectional state space operator [21] to X^g to produce a contextual sequence:

$$H_{\mathrm{ctx}} = \text{Bi-Mamba}(X^g) \in \mathbb{R}^{B\times L\times D_h}. \tag{9}$$

As illustrated in Fig. 3, Bi-Mamba instantiates two Mamba-style selective state space branches that scan the fused sequence in opposite temporal directions. The module first produces a content stream and a multiplicative gate through two linear projections, where the gate is activated by a sigmoid. The content stream is then locally mixed by a lightweight 1D convolution and passed to a selective SSM, yielding a direction-specific latent sequence. The reverse-direction branch operates on a temporally flipped copy of the input and flips the resulting output back to align with the original order.

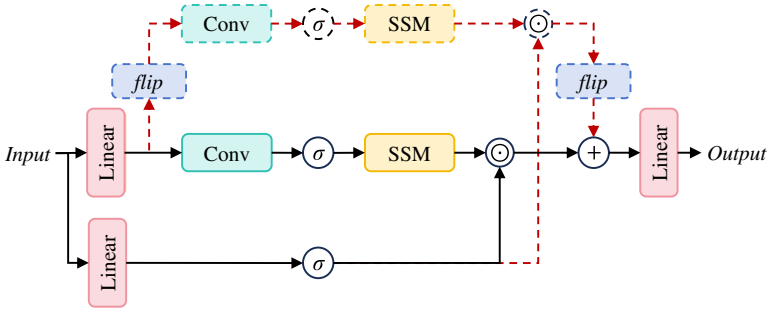


Figure 3: Bi-Mamba architecture.

A sigmoid gate is generated from the contextual sequence via a learnable linear mapping:

$$G = \sigma\big(\mathrm{Linear}_{\mathrm{gate}}(H_{\mathrm{ctx}})\big) \in (0,1)^{B\times L\times D_h}. \tag{10}$$

$\sigma(\cdot)$ denotes the sigmoid function, and $\mathrm{Linear}_{\mathrm{gate}}(\cdot)$ produces channel-wise gating coefficients that modulate the interaction between local and contextual features.

Following the interaction shown in Fig. 2, MSBR first forms a context-compensated feature by subtracting the contextual sequence from the local representation, and then applies the gate by element-wise modulation:

$$H = (H_{\mathrm{local}} - H_{\mathrm{ctx}}) \odot G, \qquad H \in \mathbb{R}^{B\times L\times D_h}. \tag{11}$$

The final output is obtained by projecting back to the input width and adding a residual connection:

$$Y = X + \mathrm{Dropout}\big(\mathrm{Linear}_{\mathrm{proj}}(H)\big), \qquad \mathrm{Linear}_{\mathrm{proj}} : \mathbb{R}^{D_h} \to \mathbb{R}^{D}. \tag{12}$$

For reproducibility, the key tensor shapes are: $X \in \mathbb{R}^{B\times L\times D}$, $X^{\ell}, X^{g} \in \mathbb{R}^{B\times L\times D_h}$, $H_{\mathrm{local}}, H_{\mathrm{ctx}}, G, H \in \mathbb{R}^{B\times L\times D_h}$, and $Y \in \mathbb{R}^{B\times L\times D}$. In the fixed-window setting considered in this paper, the convolution bank provides localized left-to-right temporal sensitivity. The Bi-Mamba pathway supplies bidirectional contextualization within the window to strengthen long-range dependency modeling. This does not claim strict online bidirectional constraints.

4.3 Cyber–Physical Symbiotic Module

The CPS module shown in Fig. 4 operates on synchronized dual-stream inputs, including the cyber stream I_{cyber} and the physical stream I_{physical}. As illustrated in Fig. 4, CPS is composed of two parts: two symmetric CPA blocks that perform bidirectional cross-context gating to enhance each modality, and a HSSF block that aggregates modality-specific updates and learns unified cyber–physical dynamics within the window for anomaly diagnosis.

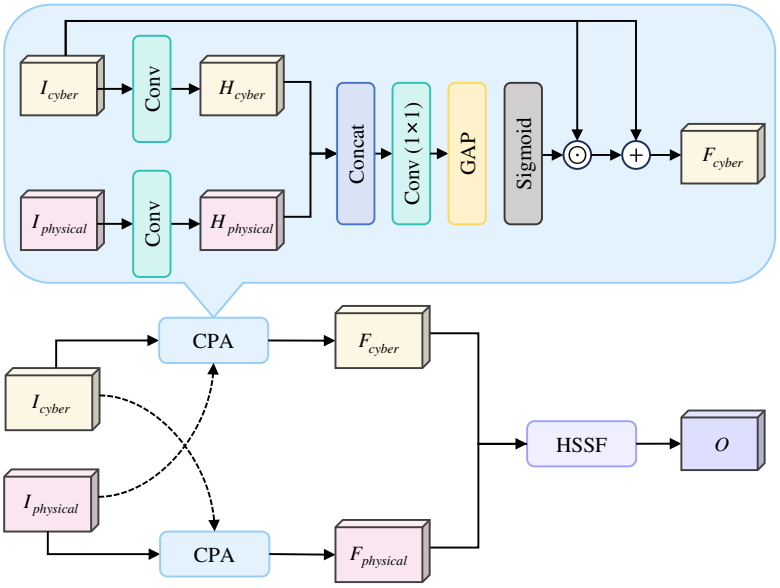


Figure 4: The proposed CPS module (lower part) and CPA (upper part with blue background) block.

Let $I_{\mathrm{cyber}}, I_{\mathrm{physical}} \in \mathbb{R}^{B\times L\times D}$ denote the aligned feature tensors. CPS produces enhanced features $F_{\mathrm{cyber}}, F_{\mathrm{physical}}$ and the fused output O as

$$F_{\mathrm{cyber}} = \mathrm{CPA}(I_{\mathrm{cyber}}, I_{\mathrm{physical}}), \quad F_{\mathrm{physical}} = \mathrm{CPA}(I_{\mathrm{physical}}, I_{\mathrm{cyber}}), \quad O = \mathrm{HSSF}(F_{\mathrm{cyber}}, F_{\mathrm{physical}}). \tag{13}$$

4.3.1 Cyber–Physical Alignment Block

The CPA block shown in upper part of Fig. 4 enhances a main modality using the other modality as cross-modal context via channel-wise gating. For a main input I_a and a context input I_b (both in $\mathbb{R}^{B\times L\times D}$), local features are extracted by 1D convolution:

$$H_a = \mathrm{Conv}(I_a), \qquad H_b = \mathrm{Conv}(I_b). \tag{14}$$

Conv(·) denotes a temporal 1D convolution that extracts local sequential patterns used as evidence for cross-modal calibration. A channel-wise gate is generated from the fused local evidence. Specifically, the local features from the main and context modalities are concatenated along the channel dimension and then passed through a point-wise channel projection (implemented as a 1×1 convolution) to mix and align the fused representation. The resulting sequence is globally averaged over the temporal axis, and a sigmoid activation is applied to obtain the channel-wise gating vector:

$$W_{a \leftarrow b} = \sigma\left(\mathrm{GAP}\left(\mathrm{Conv}(\mathrm{Concat}(H_a, H_b))\right)\right), \qquad W_{a \leftarrow b} \in (0,1)^{B \times 1 \times D}. \tag{15}$$

GAP(·) performs temporal global average pooling to summarize window-level evidence, and Concat(·) fuses the two modalities along channels before generating the gate. The gate is broadcast along the length dimension and applied to the main input with a residual connection:

$$F_a = I_a + I_a \odot W_{a \leftarrow b}. \tag{16}$$

In CPS, the same definition is instantiated twice with swapped roles to obtain F_{cyber} and F_{physical}.

4.3.2 Holistic State Space Fusion Block

The HSSF block shown in Fig. 5 fuses F_{cyber} and F_{physical} through local preprocessing, early cross-modal interaction, shared-state temporal modeling, and modality-wise gated injection followed by lightweight residual refinement. The design targets two requirements: capturing unified cyber–physical evolution at the window level and preserving modality-specific discriminative evidence under heterogeneous noise characteristics.

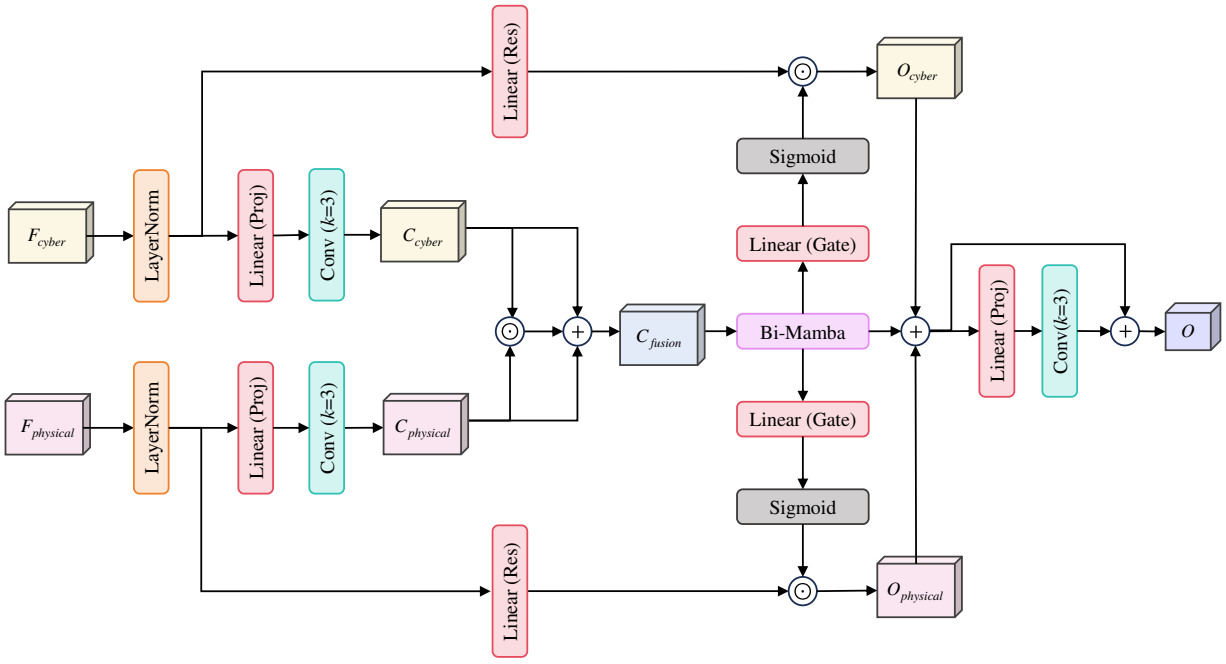


Figure 5: The proposed HSSF block.

Each modality is first normalized and locally encoded by a linear projection and a 1D convolution with kernel size 3:

$$\tilde{F}_m = \mathrm{LN}(F_m), \qquad C_m = \mathrm{Conv}_{k=3}\left(\mathrm{Proj}_m(\tilde{F}_m)\right), \qquad m \in \{\text{cyber}, \text{physical}\}, \tag{17}$$

where $C_{\text{cyber}}, C_{\text{physical}} \in \mathbb{R}^{B\times L\times D}$. $\mathrm{Proj}_m(\cdot)$ denotes a pointwise channel projection that adjusts feature width and mixes channels before temporal convolution, enabling modality-specific local encoding under a unified hidden dimension. A joint interaction state is then constructed by additive aggregation and element-wise interaction:

$$C_{\text{fusion}} = C_{\text{cyber}} + C_{\text{physical}} + \left(C_{\text{cyber}} \odot C_{\text{physical}}\right), \qquad C_{\text{fusion}} \in \mathbb{R}^{B\times L\times D}. \tag{18}$$

The additive term aligns correlated trends across modalities. The element-wise interaction emphasizes co-occurring deviations and suppresses uncorrelated fluctuations. Constructing C_{fusion} before temporal modeling lets the shared operator track coupled dynamics; late fusion would merge features only after separate evolution.

The joint state is modeled by a bidirectional state space operator Bi-Mamba to produce a shared latent sequence:

$$M = \text{Bi-Mamba}(C_{\text{fusion}}), \qquad M \in \mathbb{R}^{B\times L\times D}. \tag{19}$$

Bi-Mamba follows the bidirectional selective state space construction described in Fig. 3, providing linear-time window contextualization in both temporal directions and yielding a single shared state that summarizes coupled evolution.

Conditioned on M, each modality contributes a gated residual injection. Specifically, a residual branch is obtained from the corresponding normalized input, and a modality-specific gate is generated from M:

$$O_m = \mathrm{Res}_m(\tilde{F}_m) \odot \sigma\left(\mathrm{Gate}_m(M)\right), \qquad m \in \{\text{cyber}, \text{physical}\}, \tag{20}$$

$\mathrm{Gate}_m(\cdot)$ denotes a learnable projection that outputs channel-wise gates in $(0,1)$ to control how modality-specific evidence is injected into the shared state. Yielding O_{cyber} and O_{physical} in Fig. 5. The final fused representation is produced by aggregating the shared latent sequence with both gated injections, followed by a projection–convolution residual refinement:

$$S = M + O_{\text{cyber}} + O_{\text{physical}}, \qquad O = S + \mathrm{Conv}_{k=3}\left(\mathrm{Proj}_o(S)\right). \tag{21}$$

The shared-state gates regulate how each modality updates the fused representation under the same contextual reference. This mitigates dominance from a single stream and preserves modality-specific cues.

To validate the necessity of the fusion operator and shared-state gated injections, a parameter-matched alternative replaces Bi-Mamba with a two-layer gated temporal convolution using the same width D and matching pointwise projections, and replaces $\sigma\left(\mathrm{Gate}_m(M)\right)$ with ungated residual addition. This alternative isolates the contribution of state space fusion and shared-state gating under comparable capacity.

4.4 Multi-Scale Feature Aggregation Head

The Multi-Scale Feature Aggregation Head maps the hierarchical encoder outputs into a compact representation for cyber–physical anomaly diagnosis. As shown in the right part of Fig. 1, three multi-resolution fused features $\{Y_s\}_{s=1}^{3}$ are produced by the encoder, where $Y_s \in \mathbb{R}^{B\times L_s\times D_s}$. The head performs

temporal GAP on each scale to obtain fixed-length vectors, concatenates them into a single representation, and applies a compact MLP to output the anomaly-diagnosis logits. Here, $\mathrm{GAP}(\cdot)$ summarizes each scale into a fixed-length descriptor by averaging over time, and $\mathrm{Concat}(\cdot)$ aggregates the scale descriptors into a single fused representation for classification.

For each scale, GAP is applied along the temporal axis:

$$v_s = \mathrm{GAP}(Y_s), \qquad v_s \in \mathbb{R}^{B\times D_s}, \quad s \in \{1,2,3\}. \tag{22}$$

The pooled vectors are concatenated to form the aggregated representation:

$$z = \mathrm{Concat}(v_1, v_2, v_3) \in \mathbb{R}^{B\times(D_1+D_2+D_3)}. \tag{23}$$

Finally, a two-layer projection with SiLU and dropout produces the anomaly-diagnosis logits and the predicted distribution:

$$\hat{y} = \mathrm{Softmax}\left(\mathrm{Linear}_{\mathrm{logits}}\left(\mathrm{Dropout}\left(\mathrm{SiLU}\left(\mathrm{Linear}_{\mathrm{hid}}(z)\right)\right)\right)\right). \tag{24}$$

$\mathrm{Linear}_{\mathrm{hid}}(\cdot)$ and $\mathrm{Linear}_{\mathrm{logits}}(\cdot)$ are fully connected projections that map the fused representation to a hidden space and then to class scores.

By aggregating multi-scale representations via scale-wise pooling and concatenation, the head preserves complementary evidence across resolutions, enabling the classifier to jointly leverage transient signatures from shallow features and longer-horizon dynamics from deeper features within a single diagnostic representation.

4.5 Loss Function

The training objective consists of three components, targeting cyber–physical anomaly diagnosis and the two modality-specific cyber and physical branches. Let y be the ground-truth label and $\mathrm{CE}(\cdot,\cdot)$ denote the cross-entropy loss for the K-class setting.

The task loss supervises the final prediction produced from the fused representation:

$$\mathcal{L}_{\mathrm{task}} = \mathrm{CE}\left(\hat{y},\, y\right). \tag{25}$$

To explicitly enforce modality-wise discriminativeness, two lightweight auxiliary classifiers are attached to the cyber and physical branches using the corresponding modality-specific features from the fusion module. Let $\hat{y}_{\mathrm{cyber}}$ and $\hat{y}_{\mathrm{physical}}$ be the auxiliary predictions derived from pooled modality features like $\mathrm{GAP}(O_{\mathrm{cyber}})$ and $\mathrm{GAP}(O_{\mathrm{physical}})$. The modality-specific losses are defined as

$$\mathcal{L}_{\mathrm{cyber}} = \mathrm{CE}\left(\hat{y}_{\mathrm{cyber}},\, y\right), \qquad \mathcal{L}_{\mathrm{physical}} = \mathrm{CE}\left(\hat{y}_{\mathrm{physical}},\, y\right). \tag{26}$$

The overall objective is a weighted sum:

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{cyber}}\,\mathcal{L}_{\mathrm{cyber}} + \lambda_{\mathrm{physical}}\,\mathcal{L}_{\mathrm{physical}}. \tag{27}$$

This formulation ensures that the fused head is directly optimized for the target diagnosis, while each modality branch is simultaneously encouraged to preserve task-relevant evidence instead of relying solely on cross-modal fusion.

5 Experiments

5.1 Data Preprocessing and Alignment

This subsection instantiates the window-level features $(\mathbf{x}_t^c, \mathbf{x}_t^p)$ and sequence inputs $(\mathbf{X}_t^c, \mathbf{X}_t^p)$ in Section 3 through a leakage-free split protocol, field-wise parsing, normalization, and alignment-preserving multi-scale construction. All processing follows the dataset-defined temporal slicing and alignment rule, but operates directly on the raw cyber and physical streams of the DataSense: CIC IIoT 2025 dataset [2]. Network packets and sensor logs are first segmented into fixed 10 s non-overlapping windows $\{\mathcal{W}_t\}$, and the benign profiling interval is discarded. For each $\mathcal{W}_t$, all raw packet records and sensor messages whose timestamps fall within the window are collected and sorted by time. The dataset integration workflow constructs two standardized benchmark datasets (multiclass-8, multiclass-50) and applies label-stratified partitioning into training and testing subsets (also used for validation) at rate 8:2 before windowing to make sure no training data is exposed to testing, ensuring consistent class proportions across splits. Importantly, because model inputs are constructed as sliding sequences of length $L = 128$ with stride 1, splitting is performed at the base 10 s window index level first and then construct sliding sequences within each split to prevent any overlap of base windows across train and test. This ensures no temporal window used in evaluation is exposed during training.

On the cyber side, each packet in $\mathcal{W}_t$ is parsed into a low-level field vector like transport-layer protocol identifier, source/destination port indices, direction flag, Transmission Control Protocol (TCP) flag bits, payload length, and basic header-size descriptors. Categorical fields are index-encoded and mapped to learnable embeddings, while numerical fields are kept in their raw form. The resulting per-packet vectors are then aggregated within the window to form a fixed-length vector $\mathbf{x}_t^c \in \mathbb{R}^{d_c}$ via a combination of simple temporal pooling (such as counts, sums and means along the packet dimension) and concatenation of embedded categorical codes. No external pre-engineered feature matrix or high-level traffic statistics beyond those directly derived from the raw packet headers and lengths are used.

On the physical side, each sensor log record in $\mathcal{W}_t$ is similarly parsed into a field-level representation, including device or topic identifiers, Quality of Service (QoS) flags, message length, and the numerical payload values when available. Identifiers are index-encoded and embedded, while payload values are treated as numerical fields. These per-record vectors are then aggregated within $\mathcal{W}_t$ using analogous temporal pooling and concatenation to produce a fixed-length physical vector $\mathbf{x}_t^p \in \mathbb{R}^{d_p}$. This yields an aligned pair $(\mathbf{x}_t^c, \mathbf{x}_t^p)$ for every time window on the basis of the original logs.

If a window contains no cyber events or no physical events, the corresponding vector is set to a valid neutral placeholder by filling the numerical fields with zeros and using a special missing-modality index in the categorical fields. A binary modality-availability indicator is appended so that CPSNet can distinguish true low-activity patterns from windows where an entire stream is absent. All numerical fields are normalized using statistics computed on the training split only to avoid leakage. For a numerical feature z, z-score normalization is applied as

$$\hat{z} = \frac{z - \mu}{\sigma + \epsilon}, \tag{28}$$

where μ and σ are the mean and standard deviation over the training data, and ϵ is a small constant for numerical stability.

Finally, model-consistent sequence tensors are assembled by stacking L consecutive windows to obtain $\mathbf{X}_t^c = [\mathbf{x}_{t-L+1}^c; \ldots; \mathbf{x}_t^c] \in \mathbb{R}^{L \times d_c}$ and $\mathbf{X}_t^p = [\mathbf{x}_{t-L+1}^p; \ldots; \mathbf{x}_t^p] \in \mathbb{R}^{L \times d_p}$. Multi-scale inputs required by the hierarchical encoder are constructed via temporally consistent pooling with shared window boundaries across both modalities:

$$\mathbf{X}_t^{c,(s)} = \mathrm{Pool}_s(\mathbf{X}_t^c), \qquad \mathbf{X}_t^{p,(s)} = \mathrm{Pool}_s(\mathbf{X}_t^p), \tag{29}$$

which enforces that all cross-stream fusion modules operate on time-registered cyber and physical representations at every resolution. The resulting per-window vectors have dimensions $d_c = 78$ for the cyber stream and $d_p = 36$ for the physical stream, corresponding to the concatenated embedded categorical fields and normalized numerical fields derived from raw packet headers and sensor payloads. Columns that are constant over the training split are removed, and all remaining numerical fields are standardized by z-score normalization.

Unless otherwise stated, sequences are constructed with $L = 128$ and stride 1 over time, and the same slicing is applied to both streams to preserve strict temporal alignment. The hierarchical encoder consists of $S = 3$ scales. A temporal pooling strategy with stride 2 is applied at the transition between scales, resulting in sequence lengths of $L_1 = 128$, $L_2 = 64$, and $L_3 = 32$ for the respective stages. The hidden dimension is set to $D = 96$, and the expansion factor for the internal Mamba layers is set to 2 with a state dimension $d_{\text{state}} = 16$. The multi-scale convolutional branch in the MSBR block utilizes kernel sizes of $\{3, 5, 7\}$ to capture diverse local granularities. The coefficients of the auxiliary losses are fixed to $\lambda_{\text{cyber}} = \lambda_{\text{physical}} = 0.3$ for all experiments.

5.2 Experimental Setup

All models are implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU. The AdamW optimizer is adopted with an initial learning rate of 1×10^{-3}, a weight decay of 0.05, and a batch size of 64. A cosine-annealing schedule is used with a linear warm-up over the first 10 epochs, followed by gradual decay for a total of 100 epochs. To mitigate overfitting, a dropout rate of 0.2 is applied across all projection layers, and early stopping with a patience of 15 epochs based on validation loss is used to select the final checkpoint. All baselines are trained and evaluated under the same data preprocessing, windowing, and sequence construction protocol, using identical cyber and physical inputs and the same train and validation splits. For a fair comparison, experiments use a unified training budget and optimization setting for all methods, and reports the best validation checkpoint for testing.

Performance is evaluated using Accuracy (Acc), Precision (Prec), and Recall (Rec). For multiclass scenarios, area under the receiver operating characteristic curve (AUC) is also computed using the macro-average one-vs-rest strategy, which averages the AUC of each class against all others so that minority and majority classes are treated equally. Unless otherwise stated, all reported test results correspond to the average over five independent runs with different random seeds.

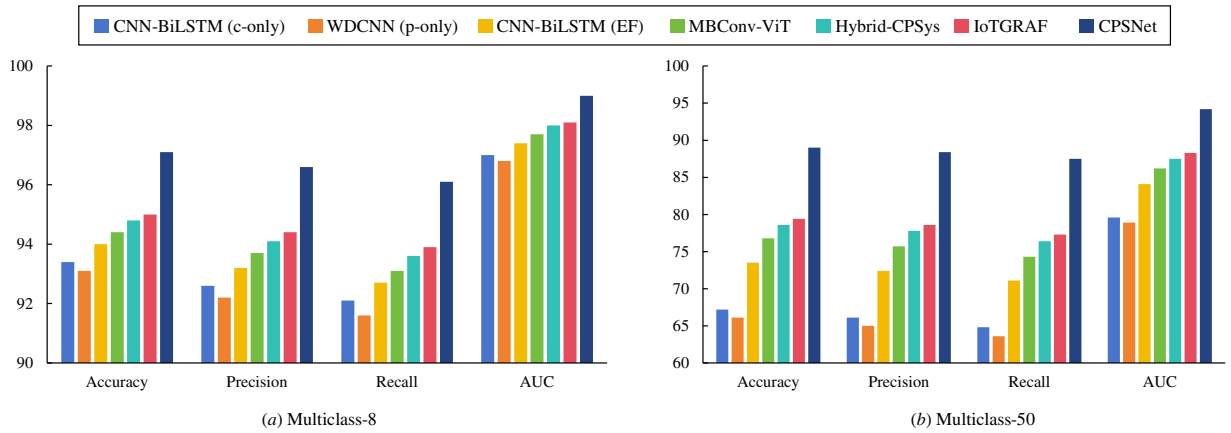
5.3 Comparison Experiments

Table 3 and Fig. 6 report the comparison results on the DataSense benchmark under two label granularities, namely Multiclass-8 and Multiclass-50. The evaluation includes representative baselines that cover single-modality modeling and multi-modal fusion. CNN-BiLSTM is a cyber-stream temporal classifier that models network-side sequential patterns. Deep Convolutional Neural Networks with Wide First-layer Kernel (WDCNN) is a physical-stream baseline that extracts discriminative features from telemetry signals via deep convolutional stacking. CNN-BiLSTM EarlyFusion performs feature-level fusion by concatenating cyber and physical features before temporal encoding. MBConv-ViT combines efficient convolutional feature extraction with global token mixing for improved sequence representation. Hybrid-CPSys adopts a hybrid cyber–physical monitoring design to integrate traffic-side and physical-side cues. IoTGRAF represents graph-based modeling for CPS anomaly diagnosis and captures relational dependencies beyond pure

sequence encoders. CPSNet is evaluated under the same dataset partition and metric definitions as the baselines to ensure comparability.

Table 3: Comparison of CPSNet with representative baselines on the DataSense benchmark under Multiclass-8 and Multiclass-50 settings (%).

Method	Multiclass-8				Multiclass-50			
	Acc	Prec	Rec	AUC	Acc	Prec	Rec	AUC
CNN-BiLSTM cyber-only [5]	93.43	92.61	92.14	97.01	67.21	66.15	64.83	79.66
WDCNN physical-only [10]	93.12	92.23	91.68	96.81	66.18	65.04	63.67	78.97
CNN-BiLSTM EarlyFusion [3]	94.03	93.24	92.77	97.49	73.53	72.48	71.19	84.13
MBConv-ViT [9]	94.44	93.76	93.18	97.79	76.88	75.73	74.34	86.24
Hybrid-CPSys [7]	94.83	94.13	93.62	98.04	78.67	77.80	76.45	87.57
IoTGRAF [13]	95.04	94.47	93.93	98.19	79.41	78.61	77.34	88.37
CPSNet	97.18	96.67	96.18	99.04	89.07	88.48	87.53	94.28



(*a*) Multiclass-8 (*b*) Multiclass-50

Figure 6: Comparison of CPSNet with representative baselines on the DataSense benchmark under Multiclass-8 and Multiclass-50 settings.

CPSNet achieves the best overall performance across both label granularities. The gains are consistent on Accuracy and AUC, and become larger under Multiclass-50, indicating that the proposed cross-modal coupling and hierarchical temporal modeling better support fine-grained diagnosis where anomaly evidence is weak and distributed across cyber and physical streams.

Fig. 6 shows that the improvement trends are aligned across all reported metrics, suggesting that CPSNet strengthens both classification reliability and ranking quality under different label granularities.

To complement diagnostic accuracy, CPSNet is also compared with representative baselines in terms of computational efficiency in Table 4, including parameter count, Floating Point Operations (FLOPs), end-to-end inference latency, and throughput measured as processed windows per second. This evaluation characterizes the practical deployability of each method under resource and real-time constraints. All efficiency numbers are measured on the same RTX 3090 using batch size 64 under FP32, reporting averaged GPU timing over repeated runs with host-to-device transfer and model forward included while excluding offline preprocessing.

Table 4: Efficiency comparison of CPSNet and representative baselines.

Method	Params (M)	FLOPs (G)	Latency (ms)	Throughput (windows/s)
CNN-BiLSTM (cyber-only)	3.6	1.1	7.4	135
MBConv-ViT	21.8	4.7	16.8	60
Hybrid-CPSys	14.2	3.6	13.9	72
IoTGRAF	18.5	5.0	18.1	55
CPSNet (ours)	9.8	2.4	10.3	105

Table 4 indicates that CPSNet maintains moderate parameter count and FLOPs while achieving low latency and high throughput compared with heavier transformer- and graph-based baselines, which facilitates near-real-time deployment without sacrificing accuracy.

Fig. 7 provides complementary evidence for the effectiveness of CPSNet from two perspectives—Fig. 7a: a confusion matrix that reveals how errors distribute across classes, and Fig. 7b: a t-SNE visualization that illustrates the geometric structure of the learned representations. Together, they connect classification outcomes with representation quality, clarifying whether the model not only achieves strong accuracy but also learns discriminative, well-structured fused features that align with cyber–physical anomaly semantics.

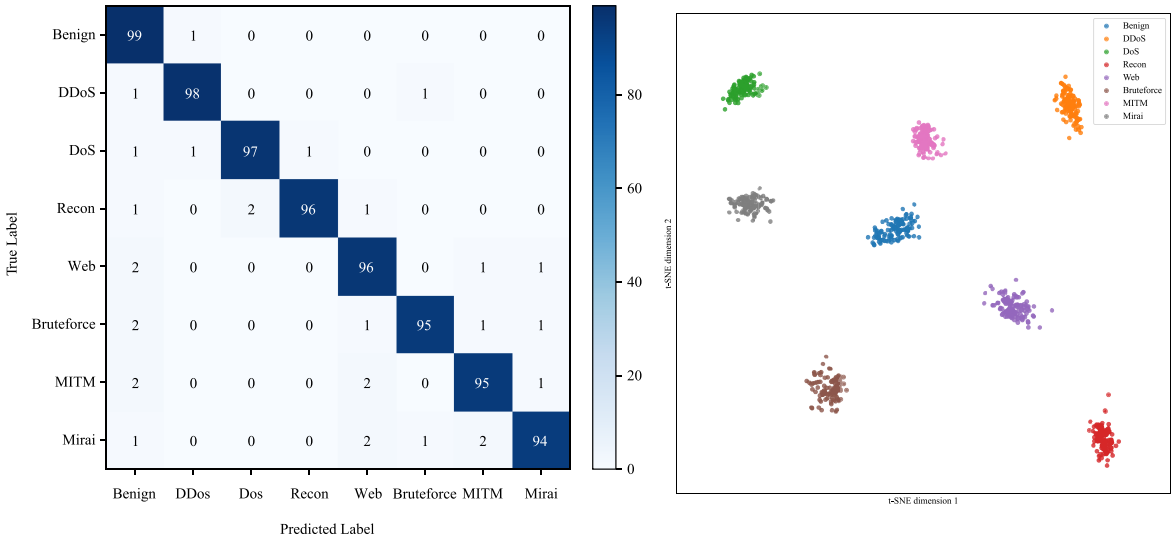


Figure 7: Confusion matrix (**a**) and t-SNE result (**b**) of CPSNet.

The confusion matrix is dominated by diagonal entries, while the t-SNE embedding forms compact clusters with clear separation between normal behavior and attack categories. The remaining confusions are concentrated among a small number of closely related classes, which is consistent with a fused representation that retains class-discriminative cues while reducing spurious cross-modal correlations.

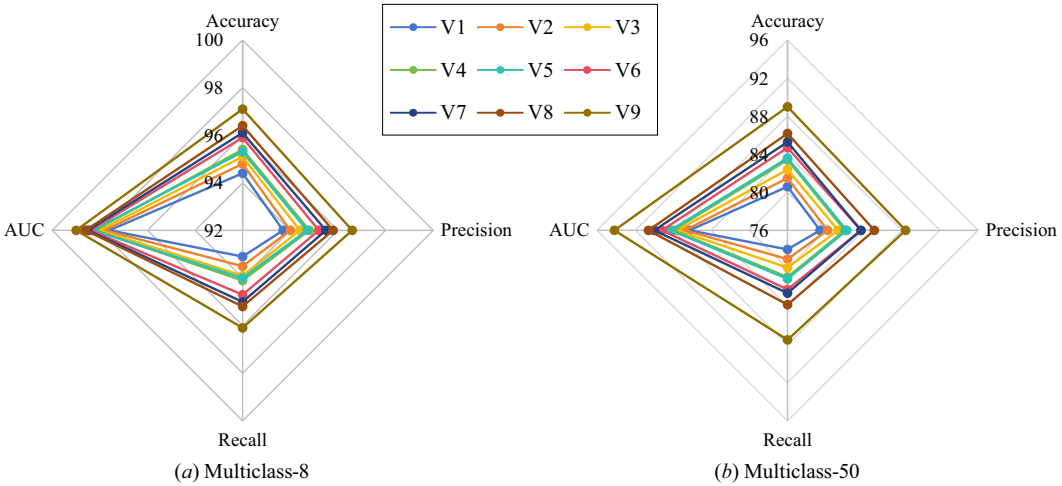
5.4 Ablation Study

To quantify the contribution of each architectural component in CPSNet, a component-wise ablation is conducted on the DataSense benchmark under both Multiclass-8 and Multiclass-50 settings in Table 5 and Fig. 8. Starting from a single-scale CNN baseline, we selectively enable MSBR, CPA, HSSF, and

Bi-Mamba, and evaluate how each module and their combinations affect discrimination performance and ranking capability in terms of Accuracy, Macro-Precision, Macro-Recall, and Macro-AUC.

Table 5: Component-wise ablation of CPSNet on the DataSense benchmark. Checkmarks indicate enabled modules (%).

	Components				Multiclass-8				Multiclass-50			
	MSBR	CPA	HSSF	Bi-Mamba	Acc	Prec	Rec	AUC	Acc	Prec	Rec	AUC
V1					94.4	93.7	93.1	97.6	80.6	79.4	78.0	86.3
V2				✓	94.8	94.0	93.5	97.8	81.5	80.2	79.0	86.8
V3	✓				95.1	94.4	93.9	98.0	82.4	81.3	79.9	87.6
V4	✓			✓	95.4	94.8	94.1	98.2	83.4	82.0	80.9	88.0
V5		✓			95.3	94.7	94.0	98.3	83.6	82.2	81.1	88.3
V6	✓	✓			95.9	95.2	94.7	98.4	84.7	83.7	82.2	89.2
V7		✓	✓		96.1	95.5	95.0	98.5	85.3	83.7	82.6	90.0
V8	✓	✓	✓		96.4	95.8	95.2	98.6	86.2	85.1	83.8	90.6
V9	✓	✓	✓	✓	97.1	96.6	96.1	99.0	89.0	88.4	87.5	94.2



(*a*) Multiclass-8 (*b*) Multiclass-50

Figure 8: Ablation study of CPSNet on Multiclass-8 (**a**) and Multiclass-50 (**b**).

Table 5 and Fig. 8 show consistent improvements when enabling stronger temporal modeling and cross-modal coupling modules, and the full configuration achieves the best results. The improvements are more pronounced on Multiclass-50, indicating that hierarchical coupling and fusion contribute more under fine-grained label fragmentation.

To further isolate the role of modality usage and fusion strategy, a second ablation evaluates single-modality variants and a dual-stream late-fusion baseline against the proposed CPSNet fusion mechanism in Table 6. This study distinguishes whether the observed improvements primarily come from having access to both modalities, or from how the modalities are coupled and fused.

Table 6 shows that using both modalities is beneficial and that structured coupling further improves over late fusion, with larger AUC gains under Multiclass-50, indicating more reliable ranking under fine-grained diagnosis.

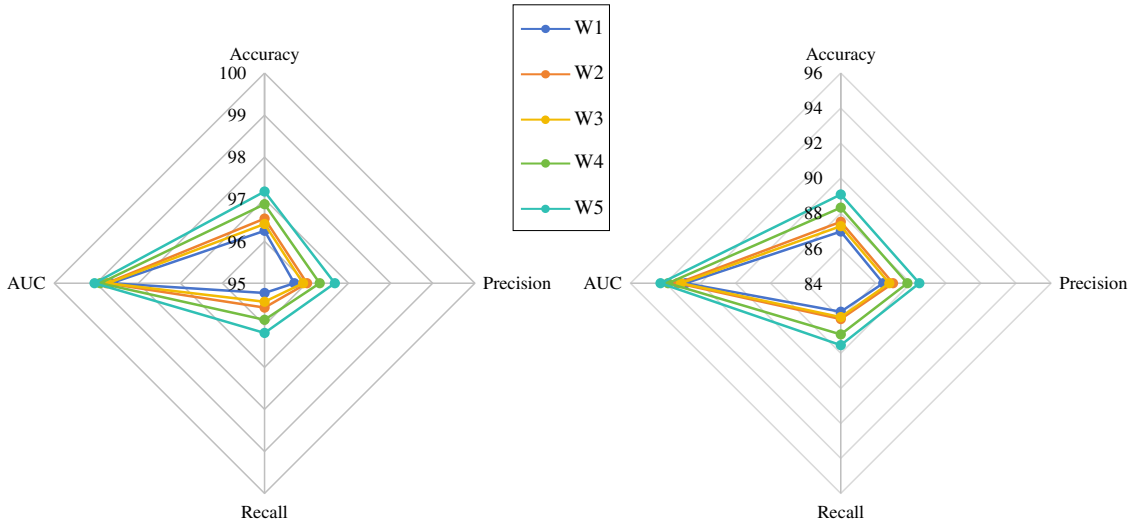
Table 6: Ablation on modality usage and fusion strategy on the DataSense benchmark (%).

Variant	Multiclass-8				Multiclass-50			
	Acc	Prec	Rec	AUC	Acc	Prec	Rec	AUC
Cyber-only (CPSNet-C)	95.6	94.9	94.2	98.3	84.7	83.6	82.1	89.8
Physical-only (CPSNet-P)	95.2	94.4	93.8	98.0	83.9	82.8	81.4	89.1
Dual-stream w/Late fusion	96.7	96.1	95.5	98.8	87.1	86.1	84.9	92.2
Dual-stream w/CPSNet (ours)	97.1	96.6	96.1	99.0	89.0	88.4	87.5	94.2

To further investigate the role of multi-scale cyber–physical symbiosis, we conduct an ablation on the CPS outputs from different encoder scales, as summarized in Table 7 and Fig. 9. Specifically, Y_1, Y_2, and Y_3 shown in Fig. 1 denote the CPS outputs from the first (highest-resolution), second (intermediate), and third (deepest) scales, respectively. Variants W1–W3 use only a single-scale CPS output, W4 aggregates the first two scales ($Y_1 + Y_2$), and W5 corresponds to the full CPSNet using all three scales ($Y_1 + Y_2 + Y_3$) with the same classifier head.

Table 7: Ablation on multi-scale CPS outputs on the DataSense benchmark (%).

Variant	Used CPS Outputs			Multiclass-8				Multiclass-50			
	Y_1	Y_2	Y_3	Acc	Prec	Rec	AUC	Acc	Prec	Rec	AUC
W1	✓			96.25	95.71	95.23	98.65	86.95	86.42	85.63	92.85
W2		✓		96.54	96.01	95.58	98.76	87.51	86.98	86.05	93.31
W3			✓	96.41	95.92	95.44	98.72	87.26	86.77	85.94	93.12
W4	✓	✓		96.88	96.31	95.87	98.91	88.32	87.81	86.92	93.86
W5	✓	✓	✓	97.18	96.67	96.18	99.04	89.07	88.48	87.53	94.28



(*a*) Multiclass-8 (*b*) Multiclass-50

Figure 9: Ablation on multi-scale CPS outputs on the DataSense benchmark.

Table 7 and Fig. 9 indicate that each scale is effective and multi-scale aggregation yields the best performance, confirming that complementary information exists across temporal resolutions and becomes more valuable under Multiclass-50.

6 Discussion

Across comparison and ablation results, the dominant accuracy gains are consistently associated with explicit cyber–physical coupling and hierarchical fusion rather than solely increasing per-stream temporal encoder strength. The margins are larger under Multiclass-50, suggesting that fine-grained diagnosis benefits from multi-scale interaction states that stabilize class separation when inter-class similarity and long-tail effects are amplified. The representation evidence in Fig. 7 is consistent with this trend, where fused features form compact and separable structures and the remaining confusions are concentrated near semantically adjacent categories, indicating that the model learns discriminative coupled dynamics instead of superficial concatenation.

The efficiency evaluation complements the accuracy results by showing that the proposed design achieves competitive latency and throughput with moderate parameter and FLOP budgets compared with heavier fusion backbones. This indicates that the hierarchical coupling and state space based temporal modeling can be deployed under resource constraints while retaining strong diagnostic performance, which is aligned with real-time industrial monitoring requirements where fixed-window inference must meet runtime budgets without compromising fine-grained recognition.

7 Conclusion

This work proposes CPSNet, an efficient window-based dual-stream framework for cyber–physical anomaly diagnosis that couples synchronized network-traffic and physical-sensor evidence via explicit multi-scale interaction. On the DataSense benchmark, CPSNet achieves strong performance under both label granularities, reaching 97.18% Accuracy and 99.04% AUC for Multiclass-8, and 89.07% Accuracy and 94.28% AUC for Multiclass-50, outperforming representative single-modality and multi-modal baselines. The gains are particularly notable for Multiclass-50, where improved AUC suggests more reliable ranking under higher inter-class similarity and long-tail effects. Component-wise ablations substantiate the design: enhanced per-stream temporal modeling yields consistent improvements, while the largest gains come from explicit cyber–physical coupling and hierarchical fusion. The full configuration performs best, indicating that MSBR-based multi-scale refinement and CPS-based interaction provide complementary benefits rather than redundant capacity. Efficiency evaluation further supports deployability, as CPSNet maintains competitive latency and throughput with a substantially smaller compute footprint than heavier graph- or transformer-based alternatives, offering a favorable accuracy–cost trade-off for window-based monitoring. These properties align with 6G IIoT requirements for sustainable edge-AI monitoring under strict latency and resource budgets.

Future work will center on digital-twin-driven industrial security and intelligence, where cyber–physical diagnosis acts as a core perception component of continuously updated industrial twins. A key challenge is to maintain scalable synchronization and strict timing alignment across heterogeneous sensing and wireless links, while supporting low-latency and energy-aware edge inference under noisy or partially missing telemetry. Another open issue is uncertainty-aware fusion under changing modality availability and evolving device populations, so that twin state updates remain reliable when streams become intermittent or distribution shifts occur. From the security perspective, digital twins introduce new attack surfaces and demand consistent cyber–physical reasoning to reduce false alarms under benign operational changes and to prevent stealthy attacks from being masked by natural process variation. Addressing these challenges

requires attack-aware twin dynamics modeling, continual adaptation to evolving workloads and network configurations, and interpretable cyber–physical evidence that can be mapped to actionable mitigation for safety-critical deployments.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Kelan Wang and Jianfei Chen; methodology, Kelan Wang and Jianfei Chen; software, Kelan Wang; validation, Kelan Wang; formal analysis, Kelan Wang, Kelan Wang and Jianfei Chen; resources, Jianfei Chen; data curation, Jianfei Chen; writing—original draft preparation, Kelan Wang and Jianfei Chen; visualization, Kelan Wang and Jianfei Chen; supervision, Jianfei Chen; project administration, Jianfei Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This study used publicly available data from the Canadian Institute for Cybersecurity DataSense: CIC IIoT dataset 2025 repository at https://www.unb.ca/cic/datasets/iiot-dataset-2025.html. The dataset is described in [2]. The designed architecture of CPSNet is available at https://github.com/KelanWang2002/CMC2026CPSNet.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wisanwanichthan T, Thammawichai M. A lightweight intrusion detection system for IoT and UAV using deep neural networks with knowledge distillation. Computers. 2025;14(7):291. doi:10.3390/computers14070291.
2. Firouzi A, Dadkhah S, Maret SA, Ghorbani AA. DataSense: a real-time sensor-based benchmark dataset for attack analysis in IIoT with multi-objective feature selection. Electronics. 2025;14(20):4095.
3. Zhang C, Li J, Wang N, Zhang D. Research on intrusion detection method based on transformer and CNN-BiLSTM in internet of things. Sensors. 2025;25(9):2725. doi:10.3390/s25092725.
4. Hu X, Zhang H, Cao J, Huang Y, Zhang X, Wang H, et al. PSRONet: a deep reinforcement learning-based sensor configuration framework in railway point machines fault diagnosis. IEEE Trans Instrum Meas. 2026;75:2500513.
5. Wang J, Si C, Wang Z, Fu Q. A new industrial intrusion detection method based on CNN-BiLSTM. Comput Mater Contin. 2024;79(3):4297–318. doi:10.32604/cmc.2024.050223.
6. Odeh A, Taleb A. Robust network security: a deep learning approach to intrusion detection in IoT. Comput Mater Contin. 2024;81(3):4149–69. doi:10.32604/cmc.2024.058052.
7. He J, Zhang W, Liu X, Liu J, Yang G. Toward intrusion detection of industrial cyber-physical system: a hybrid approach based on system state and network traffic abnormality monitoring. Comput Mater Contin. 2025;84(1):1227–52. doi:10.32604/cmc.2025.064402.
8. Hu X, Jiang C, Huang Y, Peng D, Su H, He Y, et al. SMNet: a novel compositional generalization model for industrial robot multi-joint fault diagnosis. IEEE Internet Things J. 2026:1. doi:10.1109/JIOT.2026.3652582.
9. Du C, Guo Y, Zhang Y. A deep learning-based intrusion detection model integrating convolutional neural network and vision transformer for network traffic attack in the internet of things. Electronics. 2024;13(14):2685. doi:10.3390/electronics13142685.
10. Zhang W, Peng G, Li C, Chen Y, Zhang Z. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals. Sensors. 2017;17(2):425. doi:10.20944/preprints201701.0132.v1.
11. Chen F, Zhao Z, Hu X, Liu D, Yin X, Yang J. Intelligent transformation in the operational maintenance of pumped storage units: hydraulic-mechanical multi-scenario fault diagnosis based on tensor feature extraction indicators. Adv Eng Inform. 2026;69(2):103894. doi:10.1016/j.aei.2025.103894.

12. Yang Z, Mao R, Ye L, Liu Y, Hu X, Li Y. VSC-ACGAN: bearing fault diagnosis model applied to imbalanced samples. Meas Sci Technol. 2025;36(3):036212. doi:10.1088/1361-6501/adb872.
13. Yasaei R, Moghaddas Y, Al Faruque MA. IoT-GRAF: IoT graph learning-based anomaly and intrusion detection through multi-modal data fusion. In: 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE). Piscataway, NJ, USA: IEEE; 2024. p. 1–6.
14. Gu A, Dao T, Ermon S, Rudra A, Ré C. Hippo: recurrent memory with optimal polynomial projections. Adv Neural Inf Process Syst. 2020;33:1474–87.
15. Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. arXiv:2111.00396. 2021.
16. Smith JT, Warrington A, Linderman SW. Simplified state space layers for sequence modeling. arXiv:2208.04933. 2022.
17. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. arXiv:2312.00752. 2024.
18. Dao T, Fu D, Ermon S, Rudra A, Ré C. FlashAttention: fast and memory-efficient exact attention with IO-awareness. Adv Neural Inf Process Syst. 2022;35:16344–59.
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. p. 1–11.
20. Feng J, Zhang J, Cao G, Liu Z, Ding Y. DecMamba: mamba utilizing series decomposition for multivariate time series forecasting. Comput Mater Contin. 2025;82(1):1049–68.
21. Liang A, Jiang X, Sun Y, Shi X, Li K. Bi-mamba+: bidirectional mamba for time series forecasting. arXiv:2404.15772. 2024.