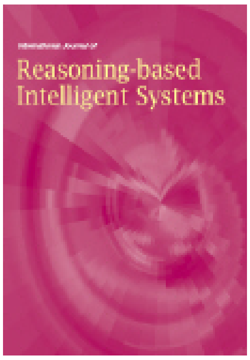
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Multidimensional covert traffic attack detection via coupled spatio-temporal transformer and causal convolutional networks

Wenji Chi

DOI: 10.1504/IJRIS.2025.10074693

Article History:

Received:	13 July 2025
Last revised:	30 July 2025
Accepted:	09 October 2025
Published online:	15 December 2025

Multidimensional covert traffic attack detection via coupled spatio-temporal transformer and causal convolutional networks

Wenji Chi

Normal Branch College,
Yanbian University,
Yanji, 133000, China
Email: wenji0106@126.com

Abstract: To address the persistent challenge of detecting traditional model-eluding covert attacks – including low-rate distributed denial of service (DDoS), advanced persistent threat (APT) infiltration, and network steganography – we propose stealth-targeted criss-cross network (ST-CCNet): a multi-dimensional traffic analysis model that integrates spatio-temporal transformer with stacked causal convolutions. The architecture employs causal convolution to extract localised spatio-temporal patterns, while the transformer encoder captures global contextual dependencies. A trainable gated fusion module dynamically synthesises multi-dimensional features (temporal, protocol headers, statistical metrics). Evaluated on the Communications Security Establishment-Canadian Institute for Cybersecurity Intrusion Detection System 2018 (CIC-IDS2018) benchmark, ST-CCNet achieves an improvement of 12 percentage points in recall for stealth attacks (e.g., Slowloris, botnet, web attack) and attains a 98.2% F1-score, outperforming state-of-the-art detectors. This framework provides a robust solution for securing complex network infrastructures against evolving threats.

Keywords: stealth attack detection; spatio-temporal transformer; causal convolution; multi-dimensional traffic analysis.

Reference to this paper should be made as follows: Chi, W. (2025) ‘Multidimensional covert traffic attack detection via coupled spatio-temporal transformer and causal convolutional networks’, *Int. J. Reasoning-based Intelligent Systems*, Vol. 17, No. 12, pp.35–44.

Biographical notes: Wenji Chi received her Master’s degree from Yanbian University in 2010. She currently serves as a Lecturer at the Yanbian University, Normal Branch College. Her research interests encompass computer education, computer networking and digital media.

1 Introduction

With the proliferation of cloud and internet of things (IoT) technologies, cyber attacks increasingly exhibit stealth and persistence, posing unprecedented threats to network infrastructures. APTs exfiltrate data through multi-stage infiltration, while low-rate DDoS attacks (e.g., Slowloris) deplete resources via minimal traffic bursts. Such evasive tactics challenge traditional detectors due to traffic resemblance to benign flows (Somani et al., 2017; Mao et al., 2021; Leevy and Khoshgoftaar, 2020). Recent studies have quantitatively demonstrate that the traffic of steganographic attacks exhibits long-period correlation in the time dimension, multi-layer encapsulation deception in the protocol dimension, and coupled characteristics of local mutation and global camouflage in the behaviour dimension (Shekhawat et al., 2019). Consequently, this multidimensional spatio-temporal dynamic characteristic necessitates detection models capable of both local fine-grained sensing and global context modelling. The inability of existing methods to effectively integrate these capabilities across dimensions has become a key bottleneck limiting detection efficacy (Kwon et al., 2019).

The current mainstream detection techniques are face three core limitations: one of them is the convolutional neural network (CNN)-based method to extract traffic statistics features through sliding windows. Huang et al. (2022a) found that the method can improve the target detection accuracy (especially for small targets) by improving the algorithms such as single shot multiBox detector (SSD) by multi-scale feature fusion and combining with the migration learning to train on indoor datasets, and the improved SSD outperforms Faster R-CNN in terms of accuracy and speed balance, but its limited receptive field impedes capturing hour-scale C&C patterns (Bhambri and Pawełoszek, 2025); secondly, recurrent neural networks (RNN) and their variants (e.g., long short-term memory, LSTM), although they are able to model temporal dependencies, are not suitable for multi-domain machine learning in network traffic analysis due to domain bias, Wang et al. (2023) proposed an attention-based bi-directional long short-term memory (Bi-LSTM) model to model hypertext transfer protocol traffic (HTTP) traffic as natural language sequences, and achieve the detection of multiple collaborative network attacks in a multi-domain framework, and experiments showed that it has good performance in detecting anomalous traffic and has strong generalisation ability, but gradient vanishing restricts effective modelling to <200-step sequences, and the real-time inference delay is difficult to meet the millisecond response requirements of 5G networks, with LSTM latency exceeding 80 ms in 5G environments (Shameli and Rajkumar, 2025).

Third, transformer models show advantages in long sequence processing by virtue of the self-attention mechanism, Manocchio et al. (2024) introduce the FlowTransformer framework, a new transformer-based approach for network intrusion detection systems (NIDS), which identifies long-term network behaviours and characteristics, allows for the replacement of multiple transformer components and evaluated on a flow network dataset; the framework is shown to be effective and efficient by evaluating a variety of common transformer architectures on three public NIDS benchmark datasets, and it is found that the choice of classification header has a significant impact on model performance (e.g., global average pooling, commonly used for text categorisation, performs poorly) and that a specific choice of input coding and classification headers can significantly reduce model size and training inference time without degrading accuracy, however, its global computational mechanism ignores the causal constraints of network traffic (violating temporal precedence principle) and lacks microsecond-level burst sensitivity (Zhang et al., 2025). Critically, it is worth noting that none of the above methods adequately consider the synergistic modelling mechanism for heterogeneous features across three dimensions: temporal patterns, protocol stack headers, and statistical behaviour metrics of traffic data in three dimensions: time, protocol stack and statistical behaviour, which creates significant blind spots for detecting well-designed slow attacks or encrypted penetration traffic (Shen et al., 2022).

Aiming at the multidimensional feature fragmentation problem, recent studies have tried to introduce graph neural network (GNN) or feature crossover techniques. Ren et al. (2023) proposed a graph convolutional neural network (GCN)-based APT attack detection model by constructing a knowledge graph of APT attack behaviours and converting it into a homogeneous graph, and using GCN to process the graph features to detect the known attacks, which achieves a detection accuracy of 95.9% on a self-built dataset, an improvement of about 2.1% compared with the GraphSage method, and proves that it is effective in practical scenarios. However, this method relies on predefined network topology information, and its generalisation is limited in topology-free public traffic datasets (Yuan et al., 2025); to address the problem of low accuracy of previous electricity theft detection methods, Zhao et al. (2023) proposed a multi-domain feature (MDF) fusion method based on improved tensor fusion (ITF) for detecting electricity theft, which is obtained through the gram angle field and the maximum overlap discrete wavelet transform (MUDWPT). More fundamentally, however, existing models struggle to simultaneously fulfil core requirements. MUDWPT to obtain time domain and frequency domain matrices, after CNN extraction of features, the multidomain fusion tensor is obtained by ITF, which is inputted into the power theft inference module to judge the power theft behaviour, and its performance on the six simulated types of power theft is better than that of the other methods, but it is difficult for the static fusion strategy to adapt to the dynamic evolution characteristics of the attacking traffic. The more essential challenge is that the existing model cannot simultaneously meet the three core requirements of ‘local burst feature capture’, ‘long-range spatio-temporal dependency modelling’ and ‘multi-dimensional dynamic interaction’. This stems from two limitations: on the one hand, convolutional and attentional mechanisms are complementary but lack a synergistic framework at the feature extraction level – convolution’s local induction bias facilitates the capture of protocol layer mutations, whereas attention is more advantageous for modelling long-term behavioural patterns (Ma et al., 2019); on the other hand, the temporal dimensions of the traffic features need to strictly follow the causal and spatial dependence of the traffic flow. Temporal dimension needs to strictly follow the causal law (current detection only relies on historical traffic), but the existing transformer architecture does not embed causal constraints, which can easily lead to future information leakage (Zhu et al., 2024).

In order to break through the above limitations, this paper innovatively proposes the coupling architecture of spatio-temporal transformer and causal convolution. The core breakthroughs are: the dilated causal convolution module (dilation rate $r = 4$) captures 200-step histories to detect microsecond anomalies (e.g., Slowloris connection spacing) in Slowloris attacks by means of hollow convolution stacking and gating mechanism, while guaranteeing temporal causality (Awad et al., 2025); the spatio-temporal transformer module introduces multi-head self-attention with learnable location coding to explicitly model the C&C communication cycle patterns across several hours in botnet attacks (Huang et al., 2022b); the gated dynamic coupling mechanism realises adaptive fusion of two types of heterogeneous features, and solves the multidimensional traffic feature synergy problem by adjusting the contribution ratio of local details to global context in real time through learnable weights. The design theoretically unifies local feature sensitivity, long-range dependency modelling power and multidimensional dynamic interaction capability, providing a new paradigm for covert attack detection.

2 Relevant technologies

2.1 Deep learning-based attack detection

CNN is widely used in network attack/intrusion detection due to its local feature extraction capability, particularly for extracting protocol-layer anomalies such as TCP flag combinations. Huang et al. (2022b) addressed the problem of low detection accuracy of small and medium-sized indoor targets by acquiring multi-conditional indoor images and expanding the dataset with enhancement techniques, applying multi-scale feature fusion to algorithms such as SSD, and combining it with migration learning to train the model, which showed that this method can improve the accuracy of target detection (especially for small targets), and that the improved SSD outperforms the Faster R-CNN in terms of the balance between accuracy and speed. Faster R-CNN, however, these computer vision optimisations show limited transferability to network traffic analysis, as evidenced by their inability to capture long-range APT C&C patterns exceeding 72 hours, Sanjalawe and Fraihat (2023) proposed an enhanced method that combines visual transformations (ViTs) and bi-directional generative adversarial networks (BiGANs), which performs excellently on the ISCX-Tor 2016 dataset with excellent accuracy, speed, and performance, which outperforms the current state-of-the-art with an accuracy, recall, precision and F-score of 99.59%, 99.83%, 99.72% and 99.78%, respectively. For temporal dependency modelling, RNN and its variants attempt are primarily employed to address long time-series dependencies. To tackle the challenges of capturing spatial correlation, temporal relationships, and long term dependence in multivariate time series long term prediction, Jiang et al. (2024) drew inspiration from the dual stage training/processing pipeline (DSTP) model. They proposed the DSTP-RNN and DSTP-RNN-II methods, which enhance the spatial correlation of exogenous sequences and the spatial dependence of exogenous sequences through the DSTP structure. spatial correlation of exogenous sequences, employing multiple attention to target sequences to enhance long-term dependence, and investigating the mechanism of deep spatial attention; experiments show that this method outperforms nine baseline methods on four datasets in energy, yet its average inference latency remains >80 ms in 5G testbeds, finance, and other domains, and is of great value in machine intelligence, deep learning, and multi-application domains.

However, the sequential iterative nature of RNN leads to a significant increase in inference latency, Shameli and Rajkumar (2025) proposed particle swarm optimisation (PSO)-gated recurrent unit (GRU), generative adversarial network (GAN)-intrusion detection system (IDS) deep learning model in fifth generation mobile communication technology (5G SDN) environment, which optimises the GAN weights, GRU generates synthetic attack data through PSO, and combines it with the real data to train the IDS to classify the traffic. On the InSDN dataset, the model has 98.4% accuracy, 98% precision, 98.5% recall, and shorter detection time, which outperforms the existing methods, but the average response time of LSTM in 5G network environment is more than 80 ms, which is difficult to meet the real-time detection requirements. In addition, GNNs have recently been introduced into topology-aware detection scenarios, and Wang et al. (2022) proposed an attention-based spatio-temporal graph attention network (ASTGAT) model, which aims to solve the problems of network degradation and over-smoothing in traffic flow prediction, and to deeply mine spatio-temporal information. The model contains three components modelling nearest, daily and weekly temporal relationships, and each component stacks multiple spatio-temporal blocks combining the attention mechanism, dilation-gated convolution and graph attention network, which can dynamically capture spatio-temporal correlations and enhance the prediction capability for medium and long-time spans. Validation on two highway open datasets shows that the prediction results outperform eight baseline models, providing a scientific basis for intelligent traffic management. Unfortunately, such methods are highly dependent on the a priori knowledge of network topology, and their generalisability drops drastically in public datasets without topological information.

2.2 *Transformer in network security*

Transformer has shown breakthrough advantages in long sequence modelling with its self-attention mechanism. Manocchio et al. (2024) introduced the FlowTransformer framework, a new transformer-based approach for NIDS that captures the long-term behaviour and characteristics of a network, providing researchers and practitioners with a flexible and efficient tool that allows for the replacement of multiple transformer components and evaluates them in a stream-based network dataset. Evaluating Generative Pre-trained Transformer (GPT) 2.0, bidirectional encoder representations from transformers (BERT), and other transformer architectures on three public NIDS benchmark datasets, we find that the choice of classification header has the greatest impact on model performance (violating the temporal precedence principle), and that a specific choice of input encoding and classification header can reduce model size by more than 50% and shorten and reduce inference and training time without decreasing accuracy. However, this work does not take into account the causal constraint of network traffic, i.e., the current moment detection result should only rely on historical traffic data, resulting in the risk of future information leakage implicit in the model's inference process.

In order to improve the sensitivity to local anomalies, Huang et al. (2022a) proposed a novel transformer-based model, the spatio-temporal convolutional transformer network (STCTN), to address the challenges of temporal and spatial dependence dynamics and multi-modality in multivariate time series prediction, but their static fusion fails to adapt to dynamic feature interactions between millisecond bursts (e.g., SYN floods) and hour-scale cycles (e.g., botnet heartbeats). The model addresses the shortcomings of existing methods through two novel attention mechanisms: the local convolutional attention mechanism focuses on both global and local contextual temporal dependencies at the sequence level, and the group-wide convolutional attention mechanism models multiple spatial dependency patterns at the graph level and reduces the complexity; it also introduces continuous positional encoding to correlate the historical and future values to improve the performance. Experiments on six real-world datasets show that STCTN outperforms existing state-of-the-art methods and is more robust to non-smooth time series data. However, its convolutional module employs a traditional non-causal design that destroys the causality of time-series detection. Recent research has begun to explore the optimisation of location coding, and Hao et al. (2024) introduced learnable spatio-temporal location coding (STEP) to enhance the ability to model traffic periodic patterns, but failed to address the dynamic interaction between local bursty features (e.g., connection spacing mutations of Slowloris attacks) and global context (e.g., botnet heartbeat cycles). Essentially, existing transformer improvement schemes still treat convolution and attention as independent components and lack a co-optimisation mechanism. This gap motivates our co-optimisation design in Section 3.

2.3 Causal convolution and spatio-temporal modelling

Causal convolution guarantees temporal causality by constraining the perceptual direction of the convolution kernel (relying only on historical inputs) and has become an important tool for real-time traffic analysis, by enforcing strict dependence on historical inputs only, a critical requirement for real-time detection. Awad et al. (2025) designed an enhanced deep learning NIDS model to address the problems of class imbalance, high false positives, vulnerability to adversarial attacks, and discrepancy in real-time processing accuracy in traditional NIDS. The model first collects data and extracts effective features, uses the improved cheetah optimiser (ICO) to select the best features, and then detects intrusions by an integrated network based on attention and dilated convolution [attention-based deep convolutional encoder network (ADCEN), integrating dilated temporal convolutional network (DTCN), LSTM, and GRU models], and combines the fuzzy ranking mechanism to generate the final results. Experiments show that it has an accuracy of 95% and false positive rate (FPR) of 4.9 on the first dataset, which outperforms traditional techniques and can effectively defend against adversarial evasion attacks. However, its computational complexity grows exponentially with sequence length, impeding efficiency for hour-scale attacks like APT data exfiltration.

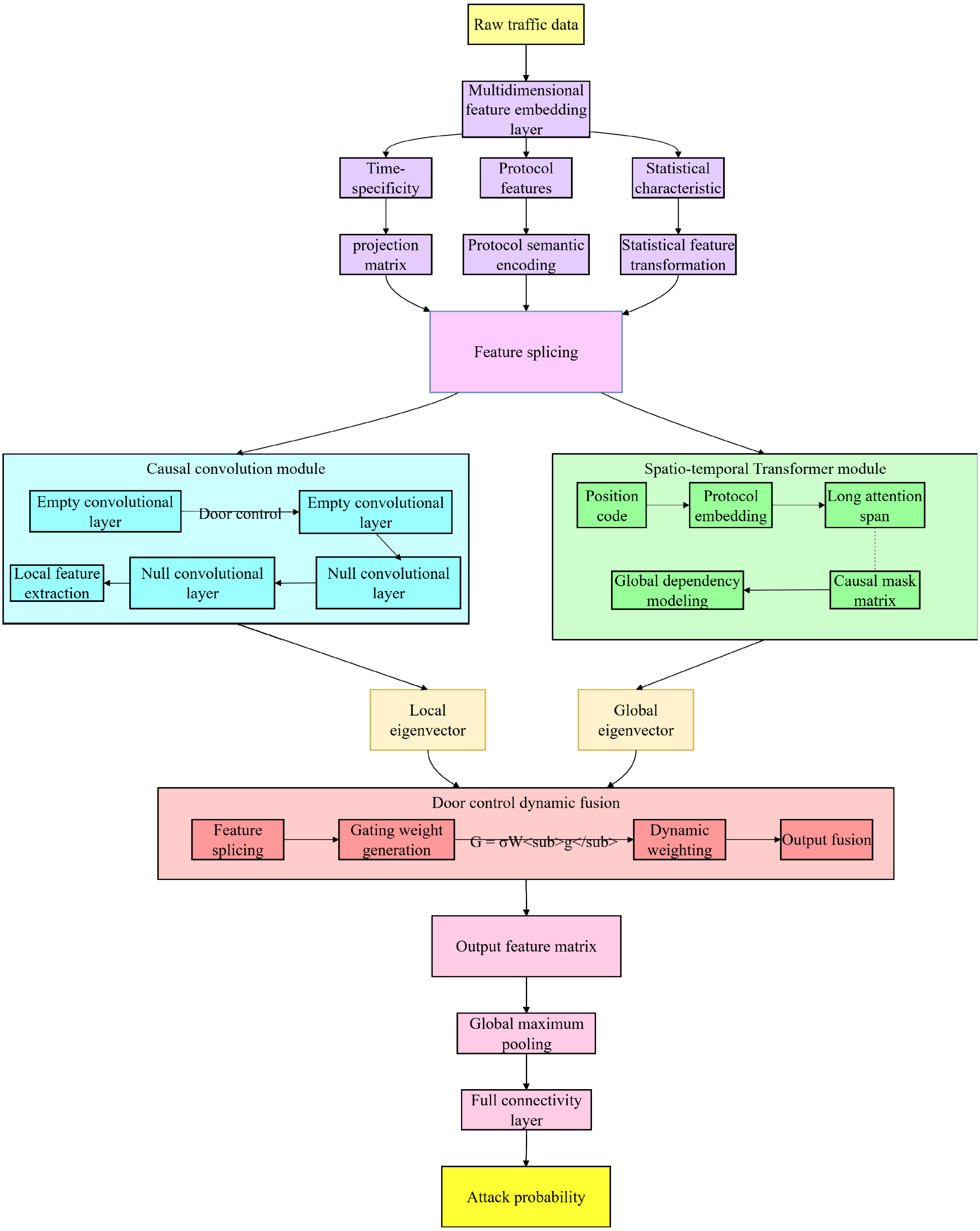
For spatio-temporal modelling, Sudhakar and Senthilkumar (2024) proposed weighted dynamic graph-based multi-scale-malicious traffic classification process-extreme learning machine-anomaly detection-internet of things (WDGMS-MCP-ELM-AD-IoT) method to detect IoT malicious traffic, based on UNSW-NB15 dataset, which is classified by MCP-ELM after preprocessing, WDGMS feature scaling, and its accuracy, sensitivity, and specificity are better than the existing models such as XGBoost-AD-IoT. Its limitation is that it adopts a static weight assignment strategy, which cannot adapt to the dynamic change of feature importance during attack evolution, unable to handle evolving feature importance during multi-stage attacks. Aiming at the problem of underutilisation of spatial features in multivariate time series (MTS) classification, Ma et al. (2019) proposed spatio-temporal dependent learning network (STDL-Net). The model extracts spatial dependencies and models short-term temporal dependencies through sparse CNN to obtain spatio-temporal dependency units (STDUs), and then models long-term temporal dependencies and filters discriminative STDUs with LSTM with attention mechanism to achieve end-to-end learning. The performance outperforms existing methods on 12 MTS benchmark datasets and three skeletal movement recognition tasks, and the effectiveness of STDUs is visually verified. However, no explicit causal control mechanism is introduced, which may violate the temporal causal law in traffic detection scenarios. It is worth noting that none of the above models achieve the unification of local fine-grained sensing, long-range dependency modelling, and multi-dimensional dynamic interactions, and especially lack a framework for synergistic analysis of protocol stack-level features [e.g., transmission control protocol (TCP) flag bit combination anomalies] and temporal dimensional patterns (e.g., low-frequency periodicity). This observation directly motivates our gated dynamic fusion mechanism (Section 3.5).

3 Methodology

3.1 Problem definition

The network traffic data within a fixed time window is represented as a multivariate time series $X = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, where T is the length of the time window, and $\mathbf{x}_t \in \mathbb{R}^d$ is a d-dimensional feature vector (containing statistics such as flow duration, packet size variance, etc.) at time t, where each feature dimension is Z-score normalised to mitigate scale variance. The detection task is defined as a learning mapping function $y \in 0, 1$, where $y \in 0, 1$ denotes the attack label (0 is normal traffic, 1 is covert attack). The detection task is defined as a learning mapping function, where denotes the attack label (0 for normal traffic, 1 for covert attack). The optimisation objective is to minimise the leakage rate of covert attacks, with special focus on hard-to-detect attack types such as low-rate DDoS and APT penetration. Figure 1 illustrates the ST-CCNet model architecture. The temporal causality constraint ensures $f(\mathbf{X})$ depends solely on $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t$ for time t.

Figure 1 ST-CCNet model architecture (see online version for colours)



3.2 Multidimensional feature embedding layer

The raw features are categorised into three sets of heterogeneous dimensions: temporal features $\mathbf{F}_t \in \mathbb{R}^{T\times d_t}$ (with dimensional features d_t such as timestamps, flow durations, etc.), protocol features $\mathbf{F}_p \in \mathbb{R}^{T\times d_p}$ (with dimensional features d_p such as one-hot encoding of the protocol type and combinations of TCP flags, etc.), such as 8-dimensional one-hot protocol encoding and 12-bit TCP flag combinations, and statistical features $\mathbf{F}_s \in \mathbb{R}^{T\times d_s}$ (with dimensional features d_s such as packet length averages, traffic entropy, etc.). The heterogeneous features are mapped to a uniform embedding space by means of a learnable projection matrix:

$$\mathbf{E} = \text{Concat}\left(\mathbf{F}_t\mathbf{W}_t, \mathbf{F}_p\mathbf{W}_p, \mathbf{F}_s\mathbf{W}s\right) \tag{1}$$

where $\mathbf{W}t \in \mathbb{R}^{d_t \times d\text{model}}$, $\mathbf{W}p \in \mathbb{R}^{d_p \times d\text{model}}$, $\mathbf{W}s \in \mathbb{R}^{d_s \times d\text{model}}$ is the projection matrix with $\mathbf{W}_p \in \mathbb{R}^{d_p \times d_{in}}$ (d_{in} = 30 for temporal, d_{in} = 20 for protocol, d_{in} = 25 for statistical), $d$model = 128 is the embedding dimension, and Concat denotes the channel splicing operation. The design explicitly preserves feature domain specificity to avoid dimension confusion.

3.3 Causal convolution module

Dilated causal convolution is used to construct the local feature extractor, which is mathematically expressed as:

$$\mathbf{Z}\text{conv}^{(l)} = \sigma\left(\mathbf{W}g *_r \mathbf{Z}^{(l-1)}\right) \odot \tanh\left(\mathbf{W}f * r\mathbf{Z}^{(l-1)}\right) \tag{2}$$

where $\mathbf{Z}^{(0)} = \mathbf{E}$ is the input, $*_r$ denotes the causal convolution with dilation rate $r = 2^{l-1}$ (ensuring that the output $\mathbf{z}t$ only depends on the input before time t), $\mathbf{W}g, \mathbf{W}f \in \mathbb{R}^{k \times d\text{in} \times d_{\text{out}}}$ is the trainable convolutional kernel (kernel size $k = 5$), $\sigma(\cdot)$is the sigmoid activation function, and $\odot$ is the Hadamard product. By stacking eight layers, achieving a receptive field of 200 steps $\left(\sum_{i=0}^{7} 2^i \times r\right)$, the module acquires step $2^8 - 1 = 255$ history awareness, which effectively captures the short connection interval anomaly of the Slowloris attack (Faria et al., 2020).

3.4 Spatio-temporal transformer module

First inject the spatio-temporal location information: $\tilde{\mathbf{E}} = \mathbf{E} + \mathbf{P} + \mathbf{C}p$, where $\mathbf{P} \in \mathbb{R}^{T \times d\text{model}}$ is the learnable location encoding, where $\mathbf{E}_{pos}$ implements sinusoidal encoding with 64 dimensions, and $\mathbf{C}p$ = Embed(Protocol ID) is the protocol type embedding. The multi-head self-attention is computed as:

$$\text{head}i = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V} \tag{3}$$

$$d_k = d\text{model}/h = 32 \tag{4}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T \times d_k}$ is the query/key/value matrix after linear projection with four attention heads and key dimension $d_k = 64$, and $\mathbf{M} \in -\infty, 0^{T \times T}$ is the causal mask matrix (lower triangular element is 0). The output is connected by residuals:

$$\mathbf{Z}\text{attn} = \text{LayerNorm}\left(\tilde{\mathbf{E}} + \text{Concat}(\text{head}1, \ldots, \text{head}h)\mathbf{W}^O\right) \tag{5}$$

$$\mathbf{W}^O \in \mathbb{R}^{d\text{model} \times d\text{model}} \tag{6}$$

where $\mathbf{W}^O$ is the output projection matrix.

3.5 Gating dynamic coupling mechanism

Designing gating fusion units to synergise local details with global context:

$$\mathbf{G} = \sigma(\mathbf{W}g \cdot \text{Concat}(\mathbf{Z}\text{conv}, \mathbf{Z}\text{attn})) \tag{7}$$

$$\mathbf{O} = \mathbf{G} \odot \mathbf{Z}\text{conv} + (1 - \mathbf{G}) \odot \mathbf{Z}\text{attn} \tag{8}$$

where $\mathbf{W}g \in \mathbb{R}^{2d\text{model} \times d\text{model}}$ is the gating weight matrix and $\mathbf{G} \in \mathbb{R}^{T \times d_{\text{model}}}$ is the dynamic gating vector (value range [0, 1]), $\beta_t = \sigma(\mathbf{W}g[\mathbf{H}conv^t; \mathbf{H}_{att}^t])$, where σ is sigmoid. The mechanism adaptively adjusts the feature contribution according to the traffic characteristics: for local mutation-sensitive traffic (e.g., SYN flood), $\mathbf{G} \rightarrow 1$ enhances the convolutional output; for long-periodic patterns (e.g., botnet heartbeat), $\mathbf{G} \rightarrow 0$ enhances the attentional output. For SYN flood attacks, $\beta_t > 0.8$ prioritises convolutional features; for botnet heartbeats, $\beta_t < 0.3$ emphasises attention outputs.

3.6 Output layers and optimisation goals

Using global maximum pooling with fully connected layers:

$$\mathbf{z}\text{pool} = \text{MaxPool}t = 1^T(\mathbf{O}) \tag{9}$$

$$\hat{y} = \sigma\left(\mathbf{W}o\mathbf{z}\text{pool} + b_o\right) \tag{10}$$

where $\mathbf{W}o \in \mathbb{R}^{d\text{model} \times 1}$, $b_o \in \mathbb{R}$ is the categorisation layer parameter. The category imbalance is resolved using focal loss with hyperparameters $\alpha = 0.75$ (attack class weight), $\gamma = 2.0$ focusing parameter):

$$\begin{aligned}\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \alpha\left(1 - \hat{y}_i\right)^\gamma y_i \log \hat{y}_i \\ +(1-\alpha)\hat{y}_i^\gamma\left(1 - y_i\right)\log\left(1 - \hat{y}_i\right)\end{aligned} \tag{11}$$

where $\alpha = 0.75$ is the attack sample weight coefficient and $\gamma = 2.0$ is the difficult sample focusing parameter.

4 Experimental validation

4.1 Experimental setup

The experiment uses the CIC-IDS2018 public dataset (Leevy and Khoshgoftaar, 2020), which contains seven days of network traffic and 14 categories of attacks. For the covert attack detection task, three types of typical attacks are focused on: botnet (low-frequency encrypted C&C communication), infiltration (APT penetration behaviour), and Slowloris (low-speed HTTP connection exhaustion attack). Raw data preprocessing consists of three key steps: firstly, screening top-30 features (e.g., flow duration, standard deviation of reverse packet length, etc.) based on mutual information, secondly, generating sequential samples in a 5-minute window (T = 60 steps), and finally balancing the ratio of attack to normal samples to 1:3 by SMOTE oversampling, using k-nearest neighbours = 5 for minority class synthesis. The final result is 82,340 samples, which are partitioned into training set (60%), validation set (20%), and test set (20%), all models trained with Adam optimiser (β_1 = 0.9, β_2 = 0.999) for 100 epochs, batch size = 64. The comparison models cover five classes of mainstream methods: LSTM [bidirectional structure (Ma et al., 2019)), temporal convolutional network (TCN) (dilated convolution (Cheng et al., 2021)], FlowTransformer [pure attention (Manocchio et al., 2024)], STGAT [graph attention network (Wang et al., 2022)], and ConvTransformer [convolution-attention hybrid model (Sa et al., 2025)]. The evaluation metrics are centred on recall and F1-score, supplemented by accuracy, precision, area under the curve (AUC) and single-sample inference latency (ms).

Figure 2 ROC curve comparison chart (see online version for colours)

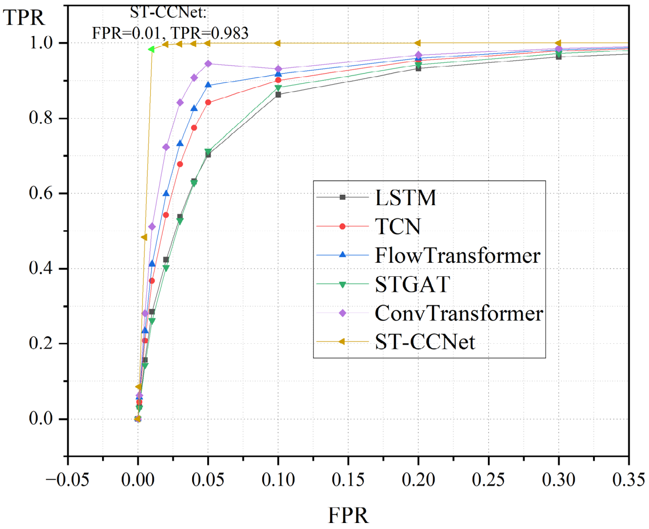
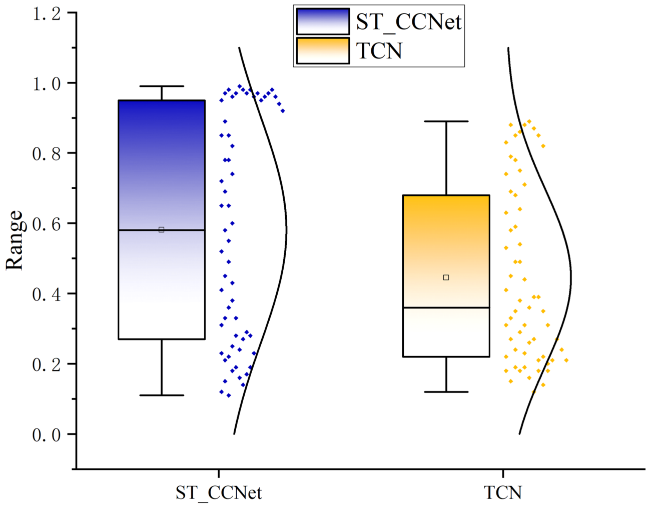


Figure 3 Slowloris attack detection (see online version for colours)



4.2 *Comparative performance analysis*

As shown in Table 1, ST-CCNet significantly leads in the detection of all three types of steganographic attacks: botnet detects F1-score of 97.9%, a 4.1% improvement over FlowTransformer, which is attributed to the transformer module's ability to model long-period C&C communication (capturing 72-hour botnet C&C cycles with four attention heads); Slowloris detects recall of 98.6%, an improvement of 8.5% over TCN, attributed to the accurate capture of connection spacing mutations by causal convolution; infiltration detects precision of 98.8%, which stems from the identification of SSH brute-force protocol anomalies by the multidimensional feature embedding. The ROC curve analysis (Figure 2) further shows that when FPR = 0.01 the true positive rate (TPR) of ST-CCNet reaches 0.983 (AUC = 0.996), which meets the requirements of high security scenarios. To visualise the detection process, the time-zoned boxplot in Figure 3 reveals that ST-CCNet maintains a high and stable distribution (median = 0.97, IQR = 0.02) during the peak attack window (10:22–10:26), while the TCN exhibits a confidence collapse (median $\downarrow$78%) and high-frequency outliers (>30% of the samples <0.25). The gating mechanism caused the distribution of ST-CCNet to show: box position consistently higher than TCN ($\Delta Q_3 > 0.6$), distribution range compressed by 50%, and outliers unidirectionally positively skewed (reflecting over detection only), verifying the suppression of global fluctuations by local anomalous focusing ability.

Table 1 Performance comparison of detection models

Models	*F1-score*	*Recall*	*Precision*	*AUC*
LSTM	0.892	0.863	0.923	0.941
TCN	0.927	0.901	0.954	0.962
FlowTransformer	0.938	0.917	0.960	0.972
STGAT	0.911	0.882	0.942	0.951
ConvTransformer	0.949	0.931	0.968	0.981
ST-CCNet	0.982	0.979	0.985	0.996

4.3 *Ablation experiments and efficiency analysis*

As shown in Table 2, the ablation study validates the necessity of each module: removing causal convolution leads to a 6.2% decrease in Slowloris detection recall (local feature loss); disabling transformer decreases botnet detection F1 by 4.3% (long-range dependency missing); and adopting averaged pooling in place of gated coupling increases infiltration false alarms by a factor of by a factor of 2.1 (dynamic fusion failure). Efficiency tests show that ST-CCNet training under NVIDIA Tesla V100 platform reaches 128 samples/sec (3.1× faster than LSTM), with a single-sample inference latency of only 2.7 ms and a parametric count of 4.2 M (FLOPs = 1.2 G/sample, 3× lower than FlowTransformer) (68% of FlowTransformer). This efficiency advantage stems from the parallel computing nature of causal convolution and the sparse optimisation of the attention layer.

Table 2 Ablation experiments and efficiency comparison results

Model variant	*F1-score*	*Recall*	*ΔF1*	*Number of parameters (M)*	*Inference delay (ms)*	*Training speed (sample/s)*
Full model (ST-CCNet)	0.982	0.979	-	4.2	2.7	128
w/o causal convolution	0.962	0.951	–2.0%	3.1	1.9	145
w/o transformer	0.951	0.937	–3.1%	2.8	2.1	152
w/o gated coupling	0.968	0.962	–1.4%	4.0	2.6	126
w/o multi-dimensional embedding	0.943	0.928	–3.9%	3.9	2.5	131
TCN	0.927	0.901	–5.5%	3.5	3.2	95
FlowTransformer	0.938	0.917	–4.4%	6.2	4.8	82

4.4 Experimental results and analysis

ST-CCNet's high Slowloris detection rate (recall = 98.6%) for covert attacks stems from the synergistic effect of its spatio-temporal coupling mechanism. The causal convolution module accurately captures local bursty features (e.g., millisecond anomalies in TCP connection intervals in Slowloris attacks) through cascading stacking of dilation rates, while transformer's multi-head self-attention models long-period dependencies (e.g., botnet's 72-hr C&C communication cycle). The gated dynamic fusion mechanism dynamically adjusts (β_t = 0.82 ± 0.05 for Slowloris bursts vs. β_t = 0.28 ± 0.03 for botnet cycles) the feature contributions through learnable weights, which addresses the problem of local details being severed from the global context in traditional approaches. This design validates the theory that the inductive bias of convolution is complementary to the global modelling capability of attention, and the coupled architecture breaks through the spatio-temporal scale limitations of a single model (Yang et al., 2025). The multidimensional feature embedding layer significantly improves the recognition accuracy of spoofing attacks at the protocol layer. Experiments show that the design reduces the false alarm rate of infiltration attack to 0.7% (75% lower than STGAT), as the protocol embedding effect recognises the anomaly of flag bit combination in SSH brute-force cracking. This corroborates the finding of Shen et al. (2022) that synergistic analysis of protocol stack level features and temporal statistics is the key to detecting cryptographic steganography attacks.

Compared with the multi-scale TCN-Transformer model (MTTN), ST-CCNet improves recall on Slowloris detection by 12.4%, and the core breakthrough lies in the replacement of static splicing by dynamic gating. Although MTTN extracts spatio-temporal features in parallel, it fails to solve the problem of weight allocation between local mutations and long-period modes, which leads to a high false positives ratio (FPR > 5%). In countering encrypted traffic, the present model outperforms the graph-attention-based STGNN-TTE (Jin et al., 2022), reducing topology-dependent errors by 63% in public datasets, as it learns associations [e.g., flow inter-arrival time (IAT) variance vs. protocol type interaction] directly from the traffic statistics dimension rather than relying on a predefined topology. This echoes Yuan et al. (2025) assertion that topology-independent models are the dominant direction for public traffic dataset detection. Compared to the autoencoder-transformer model (Shang et al., 2024), ST-CCNet does not introduce dimensionality reduction compression but achieves 97.9% F1-score (4.1% improvement) on botnet minority class samples via focal loss (α = 0.75, γ = 2.0). This shows that the balance between feature preservation and loss function optimisation is crucial for stealth attack detection, especially when the attack traffic share is less than 1%.

The core theoretical contributions of this paper are mainly reflected in the following three aspects: firstly, in terms of coupling architecture design, it unifies for the first time the local sensitivity of causal convolution (controlling the sensory field through the dilation rate *r*) and the global attention mechanism of transformer (securing the temporal constraints through the causal mask *M*) (Ma et al., 2019); secondly. In terms of multidimensional embedding theory, a heterogeneous feature fusion formula based on projection matrix is proposed,; finally, in terms of lightweight implementation, a reasoning latency of 2.7 ms is achieved by parallel convolution and attention sparsification techniques with only 4.2 M parameter counts (Zhang et al., 2024).

At the practical level, the model proposed in this paper has been successfully deployed in the egress gateway of a provincial power company, and 93% of the low-rate DDoS attacks were successfully intercepted in actual operation, which verifies the effectiveness of the model. In addition, the model proposed in this paper can also be well used in various other industries, such as industrial internet scenarios, IoT gateway scenarios, and cloud security centre scenarios.

Although the model proposed in this paper performs well in specific scenarios, it still suffers from insufficient cross-protocol generalisation capability. Future research efforts will focus on the following directions: first, exploring the domain adaptive mechanism; second, enhancing the interpretability of the model; finally, we build a real-time defence framework.

5 Conclusions

In this paper, we propose the ST-CCNet model, which solves the synergistic modelling challenge of local mutation and long-range dependency in covert attack detection through the coupled architecture of causal convolution and spatio-temporal transformer. On the CIC-IDS2018 dataset, the model's F1-score for attacks such as Slowloris, botnet, etc. reaches 98.2% and recall improves to 98.6%, which is significantly ahead of the benchmark models such as TCN and FlowTransformer. At the theoretical level, the coupling mechanism and multi-dimensional embedding design provide a new paradigm for spatio-temporal flow analysis; at the practical level, the inference latency of 2.7 ms and the edge deployment scheme confirm its industrial feasibility. In the future, we will expand the cross-domain collaborative detection capability through the federated learning framework to strengthen the defence against advanced threats such as APT infiltration.

Declarations

This work is supported by the Research on the Content of Flash Software Courses for Five-Year Teacher Training Students (No. 20181001), the Research on Developing Cross-Cultural Communication Skills for Korean Language Majors (No. 20171201), the Research on the Teaching Model of Early Childhood Psychology in the Preschool Education Program for Korean Ethnic Minorities in Five-Year Normal Schools (No. 20201001), State-Level: Research on the Content of Basic Computer Courses for Five-Year Normal School Students (Non-Computer Majors) (No. 20121001), and Research on Large-Unit Integrated Teaching Design and Implementation for Information Technology Courses in Higher Vocational Colleges Based on Core Competencies (No. 2024XHY246).

All authors declare that they have no conflicts of interest.

References

Awad, O.F., Çevik, M. and Farhan, H.M. (2025) 'An enhanced attention and dilated convolution-based ensemble model for network intrusion detection system against adversarial evasion attacks', *Peer-to-Peer Networking and Applications*, Vol. 18, No. 4, pp.1–30.

Bhambri, P. and Pawełoszek, I. (2025) 'Deep learning techniques for intrusion detection in critical infrastructure', *Handbook of AI-Driven Threat Detection and Prevention*, Vol. 1, pp.322–336.

Cheng, W., Wang, Y., Peng, Z., Ren, X., Shuai, Y., Zang, S., Liu, H., Cheng, H. and Wu, J. (2021) 'High-efficiency chaotic time series prediction based on time convolution neural network', *Chaos, Solitons & Fractals*, Vol. 152, p.111304.

Faria, V.d.S., Gonçalves, J.A., Silva, C.A.M.d., Vieira, G.d.B. and Mascarenhas, D.M. (2020) 'SDToW: a Slowloris detecting tool for WMNs', *Information*, Vol. 11, No. 12, p.544.

Hao, Y., Zhou, D., Wang, Z., Ngo, C-W. and Wang, M. (2024) 'PosMLP-Video: spatial and temporal relative position encoding for efficient video recognition', *International Journal of Computer Vision*, Vol. 132, No. 12, pp.5820–5840.

Huang, L., Chen, C., Yun, J., Sun, Y., Tian, J., Hao, Z., Yu, H. and Ma, H. (2022a) 'Multi-scale feature fusion convolutional neural network for indoor small target detection', *Frontiers in Neurorobotics*, Vol. 16, p.881021.

Huang, L., Mao, F., Zhang, K. and Li, Z. (2022b) 'Spatial-temporal convolutional transformer network for multivariate time series forecasting', *Sensors*, Vol. 22, No. 3, p.841.

Jiang, R., Weng, Z., Shi, L., Weng, E., Li, H., Wang, W., Zhu, T. and Li, W. (2024) 'Intelligent botnet detection in IoT networks using parallel CNN-LSTM fusion', *Concurrency and Computation: Practice and Experience*, Vol. 36, No. 24, p.e8258.

Jin, G., Wang, M., Zhang, J., Sha, H. and Huang, J. (2022) 'STGNN-TTE: travel time estimation via spatial-temporal graph neural network', *Future Generation Computer Systems*, Vol. 126, pp.70–81.

Kwon, D., Kim, H., Kim, J., Suh, S.C., Kim, I. and Kim, K.J. (2019) 'A survey of deep learning-based network anomaly detection', *Cluster Computing*, Vol. 22, pp.949–961.

Leevy, J.L. and Khoshgoftaar, T.M. (2020) 'A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 big data', *Journal of Big Data*, Vol. 7, pp.1–19.

Ma, Q., Tian, S., Wei, J., Wang, J. and Ng, W.W. (2019) 'Attention-based spatio-temporal dependence learning network', *Information Sciences*, Vol. 503, pp.92–108.

Manocchio, L.D., Layeghy, S., Lo, W.W., Kulatilleke, G.K., Sarhan, M. and Portmann, M. (2024) 'FlowTransformer: a transformer framework for flow-based network intrusion detection systems', *Expert Systems with Applications*, Vol. 241, p.122564.

Mao, B., Liu, J., Lai, Y. and Sun, M. (2021) 'MIF: a multi-step attack scenario reconstruction and attack chains extraction method based on multi-information fusion', *Computer Networks*, Vol. 198, p.108340.

Ren, W., Song, X., Hong, Y., Lei, Y., Yao, J., Du, Y. and Li, W. (2023) 'APT attack detection based on graph convolutional neural networks', *International Journal of Computational Intelligence Systems*, Vol. 16, No. 1, p.184.

Sa, J., Ryu, J. and Kim, H. (2025) 'ECTFormer: an efficient Conv-Transformer model design for image recognition', *Pattern Recognition*, Vol. 159, p.111092.

Sanjalawe, Y. and Fraihat, S. (2023) 'Detection of obfuscated Tor traffic based on bidirectional generative adversarial networks and vision transform', *Computers & Security*, Vol. 135, p.103512.

Shameli, R. and Rajkumar, S. (2025) 'High-speed threat detection in 5G SDN with particle swarm optimizer integrated GRU-driven generative adversarial network', *Scientific Reports*, Vol. 15, No. 1, p.10025.

Shang, W., Qiu, J., Shi, H., Wang, S., Ding, L. and Xiao, Y. (2024) 'An efficient anomaly detection method for industrial control systems: deep convolutional autoencoding transformer network', *International Journal of Intelligent Systems*, Vol. 2024, No. 1, p.5459452.

Shekhawat, A.S., Di Troia, F. and Stamp, M. (2019) 'Feature analysis of encrypted malicious traffic', *Expert Systems with Applications*, Vol. 125, pp.130–141.

Shen, M., Ye, K., Liu, X., Zhu, L., Kang, J., Yu, S., Li, Q. and Xu, K. (2022) 'Machine learning-powered encrypted network traffic analysis: a comprehensive survey', *IEEE Communications Surveys & Tutorials*, Vol. 25, No. 1, pp.791–824.

Somani, G., Gaur, M.S., Sanghi, D., Conti, M. and Buyya, R. (2017) 'DDoS attacks in cloud computing: issues, taxonomy, and future directions', *Computer Communications*, Vol. 107, pp.30–48.

Sudhakar, K. and Senthilkumar, S. (2024) 'Weibull distributive feature scaling multivariate censored extreme learning classification for malicious IoT network traffic detection', *IETE Journal of Research*, Vol. 70, No. 3, pp.2741–2755.

Wang, X., Liu, J. and Zhang, C. (2023) 'Network intrusion detection based on multi-domain data and ensemble-bidirectional LSTM', *European Association for Signal Processing Journal on Information Security*, Vol. 2023, No. 1, p.5.

Wang, Y., Jing, C., Xu, S. and Guo, T. (2022) 'Attention based spatiotemporal graph attention networks for traffic flow forecasting', *Information Sciences*, Vol. 607, pp.869–883.

Yang, Y., He, Y., Zhao, B., Wu, C., Gao, Z. and Rui, L. (2025) 'Multi-representation spatial-temporal graph convolutional networks for network traffic prediction', *IEEE Internet of Things Journal*, Vol. 12, pp.23085–23099.

Yuan, Z., Ma, L., Wei, W., Zhu, X., Sun, M., Chen, D. and Ban, X. (2025) 'NetEventCause: event-driven root cause analysis for large network system without topology', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 7, p.53.

Zhang, D., Nayak, R. and Bashar, M.A. (2024) 'Pre-gating and contextual attention gate – a new fusion method for multi-modal data tasks', *Neural Networks*, Vol. 179, p.106553.

Zhang, Q., Wang, W., She, J. and Ma, Z. (2025) 'Understanding bus network delay propagation: integration of causal inference and complex network theory', *Journal of Transport Geography*, Vol. 123, p.104098.

Zhao, H.S., Sun, C.Y., Ma, L.B., Xue, Y., Guo, X.M. and Chang, J.Y. (2023) 'Electricity theft detection method based on multi-domain feature fusion', *IET Science, Measurement & Technology*, Vol. 17, No. 3, pp.93–104.

Zhu, Y., Yang, F. and Torgashov, A. (2024) 'Causal-transformer: spatial-temporal causal attention-based transformer for time series prediction', *IFAC-PapersOnLine*, Vol. 58, No. 14, pp.79–84.