DOI: 10.21917/ijsc.2025.0556

HYBRID POSITIONAL EMBEDDINGS IN TRANSFORMER MODELS FOR CYBERATTACK DETECTION IN INDUSTRIAL CONTROL SYSTEMS

Syed Minhaz Ul Hassan and Meena Chaudhary

Department of Computer Science Engineering, Institute of Engineering and Technology, Mangalayatan University, India

Abstract

Cyberattack detection in industrial control systems (ICS) is critical to ensuring the security and resilience of essential infrastructure, such as water treatment and distribution networks. However, existing anomaly detection methods often struggle with capturing complex temporal dependencies and differentiating between cyberattacks and normal operational variations. In this study, we propose a novel Transformer-based approach with hybrid positional embeddings for detecting cyber-attacks in multivariate time series data. Our method integrates learnable, sinusoidal, and rotary position embeddings, enabling the model to effectively capture both absolute and relative temporal relationships. This hybrid embedding strategy addresses key limitations of conventional Transformers in handling time-series data by improving the encoding of temporal dependencies. We evaluate our approach on two widely used cybersecurity datasets: Secure Water Treatment (SWaT) and Water Distribution (WADI), which simulate real-world industrial cyber-physical system (CPS) attacks. Our model outperforms state-of-the-art baselines, achieving high detection accuracy and robust anomaly identification. Additionally, an ablation study demonstrates the contribution of hybrid positional embeddings in improving cyberattack detection performance. This work enhances AI driven security frameworks for industrial systems by providing a scalable and effective solution for cyber threat monitoring in critical infrastructures.

Keywords:

Multivariate Time Series Data, Anomaly Detection, Transformer Models, Learnable Positional Embeddings, Rotary Position Embeddings

1. INTRODUCTION

The increasing digitization of Industrial Control Systems (ICS) and critical infrastructure has led to a surge in cyber threats, targeting water treatment facilities, power grids, and manufacturing plants. These cyber-physical systems (CPS) rely on multivariate time-series data for continuous monitoring and operational efficiency. However, the growing sophistication of cyberattacks, including ransomware, false data injection, denial-of-service (DoS), and insider threats, poses severe security risks. Traditional cybersecurity mechanisms often fail to detect these threats in real-time, making the need for robust cyberattack detection systems imperative.

Traditional intrusion detection systems (IDS) and signature-based cybersecurity methods struggle to identify zero-day attacks, adaptive adversarial techniques, and coordinated multistage intrusions. Machine learning-based approaches, including statistical models and classical anomaly detection methods, fail to capture the complex attack patterns, stealthy infiltration tactics, and evolving threat vectors in ICS environments. While deep learning models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have demonstrated improved capabilities in learning temporal patterns, they remain ineffective against highly dynamic cyber threats due to their limitations in handling long-range dependencies, adaptive attack strategies, and real-time intrusion detection requirements.

Table.1. List Of Acronyms

Acronym	Full Form	Acronym	Full Form
SMD	Server Machine Dataset	MTAD-GAT	Multi-scale Temporal Anomaly Detection using GAT
WADI	Water Distribution	TRAN-AD	Transformer-based Anomaly Detection
SMAP	Soil Moisture Active Passive	RoPE	Rotary Position Embedding
NAB	Numenta Anomaly Benchmark	tAPE	Temporal Absolute Position Encoding
CNN	Convolutional Neural Network	eRPE	Enhanced Relative Position Encoding
RNN	Recurrent Neural Network	iRPE	Image Relative Position Encoding
LSTM	Long Short-Term Memory	ALiBi	Attention with Linear Biases
AE	Autoencoder	NoPE	No Positional Encoding
VAE	Variational Autoencoder	AFT	Attention Free Transformer
GRU	Gated Recurrent Unit	HVAE	Hierarchical Variational Auto Encoder
GNN	Graph Neural Network	MSE	Mean Squared Error
GAT	Graph Attention	AUC-ROC	Area Under the ROC Curve
TPR	True Positive Rate	FPR	False Positive Rate

Recent advancements have explored attention-based and graph-based approaches to enhance cybersecurity in ICS. Models such as Multi-scale Temporal Anomaly Detection using Graph Attention Networks (MTAD-GAT) and Transformer-based Cyberattack Detection (TRAN-AD) have shown promise in capturing intricate dependencies within ICS security logs, network traffic, and sensor data. However, existing methods often struggle to detect sophisticated attack sequences, adapt to adversarial manipulations, and balance computational efficiency with threat response accuracy.

Transformer models, originally developed for natural language processing (NLP), have demonstrated exceptional

performance in cyberattack detection due to their self-attention mechanism, which enables efficient modelling of long-term attack behaviours and malicious activity correlations. However, the standard sinusoidal positional encoding in Transformers lacks adaptability to the complex temporal characteristics of cyber threats in ICS, limiting its effectiveness in capturing stealthy attacks, time-delayed intrusions, and coordinated multi-vector exploits.

To overcome these limitations, this study proposes a novel hybrid positional embedding-based Transformer model for cyberattack detection in ICS environments. The proposed approach integrates three complementary positional encoding techniques:

- **Learnable Positional Embeddings**: Adaptive representations that allow the model to fine-tune temporal encoding for dataset-specific attack patterns.
- **Sinusoidal Positional Embeddings**: A stable and generalizable foundation, ensuring consistent representation of temporal order and facilitating model generalization across varying sequence lengths.
- **Rotary Position Embeddings (RoPE)**: A relative positional encoding mechanism that enhances the model's ability to capture dependencies across sensor readings, where the relative timing between cyber events is often more crucial than absolute timestamps.

This hybrid embedding strategy enhances the Transformer model's ability to distinguish between normal variations and cyberattacks in ICS datasets such as Secure Water Treatment (SWaT) and Water Distribution (WADI). These datasets simulate real-world cyberattacks on critical infrastructure, making them ideal benchmarks for evaluating AI-driven anomaly detection methods. Since SWaT and WADI are multivariate time-series datasets, we further assessed the generalizability of the proposed approach on additional benchmark datasets, including the Server Machine Dataset (SMD) and Soil Moisture Active Passive (SMAP). These datasets provide diverse challenges in multivariate anomaly detection, ensuring that the model is robust across different industrial and environmental settings.

To validate the effectiveness of the proposed approach, extensive experiments were conducted on SWaT, WADI, SMD, and SMAP datasets. The model demonstrates significant improvements over state-of-the-art anomaly detection techniques, achieving higher detection accuracy and reducing false positives. An ablation study further confirms the contribution of each embedding type to the model's overall performance, providing insights into the importance of hybrid positional encodings for time-series cybersecurity applications.

1.1 KEY CONTRIBUTIONS

The key contributions of this research are as follows:

- Introduction of a novel hybrid positional embedding-based Transformer model for cyberattack detection in industrial control systems.
- Comprehensive evaluation of the proposed approach on SWaT and WADI datasets, demonstrating improved anomaly detection performance over existing deep learning methods.
- Assessment of model generalizability on additional multivariate time-series datasets, including SMD and SMAP, to ensure robustness across different industrial and environmental scenarios.
- Ablation study analysing the impact of different embedding types, offering deeper insights into their role in capturing multivariate time-series anomalies in ICS security applications.

This research advances AI-driven cybersecurity in ICS by providing a flexible and scalable solution for real-time cyber-attack detection. The following sections explore related works, dataset descriptions, methodology, evaluation metrics, experimental results, limitations, future directions, and conclusions.

2. RELATED WORKS

2.1 DEEP LEARNING-BASED CYBERATTACK DETECTION

Deep learning techniques have significantly advanced the detection of cyberattacks in Industrial Control Systems (ICS). Various neural network architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and Autoencoders (AEs), have been employed for this purpose. For instance, Bamber et al [1] proposed a hybrid CNN-LSTM approach for an intelligent cyber intrusion detection system. This model combines CNN layers for spatial feature extraction and LSTM layers for temporal sequence learning, demonstrating improved detection accuracy and reduced false positives. Similarly, Sethi et al [2] developed a hybrid CNN-LSTM-based intrusion detection system tailored for industrial IoT networks. Their approach effectively identifies threats by leveraging the strengths of both architectures. Furthermore, Sethi et al. introduced an artificial intelligence-driven cyberattack detection system by integrating Deep Belief Networks (DBNs) with CNNs. This method enhances the detection of cyber threats in industrial IoT environments.

2.2 TRANSFORMER-BASED CYBERATTACK DETECTION

The Transformer model has shown significant promise in cybersecurity applications, particularly in detecting sophisticated cyber threats in Industrial Control Systems (ICS) and cloud-based environments. Transformers utilize self-attention mechanisms to analyse intrusion patterns, malware signatures, and unauthorized access attempts. Vaswani et al. [3] introduced the Transformer architecture, which has been foundational in various AI applications, including cybersecurity. Several studies have explored the effectiveness of Transformers in this domain. Chen et al. [4] proposed FlowTransformer, a framework that leverages Transformers to identify long-term behavioural patterns in network traffic for improved intrusion detection. Similarly, Wang et al. [5] introduced a Transformer-Based Network Intrusion Detection System (NIDS) tailored for cloud environments, demonstrating its adaptability to evolving threat vectors while reducing false positives. Furthermore, Zhao et al. [6] developed LogShield, a Transformer-based approach for detecting Advanced Persis tent Threats (APTs) by analysing system logs

through self-attention mechanisms. These studies highlight the growing significance of Transformer-based models in strengthening cybersecurity defenses across various applications.

2.3 GRAPH NEURAL NETWORK BASED CYBERATTACK DETECTION

Graph Neural Networks (GNNs) have proven highly effective in detecting cyber threats within large-scale networks by modelling complex relationships between attack vectors, compromised nodes, and malware propagation paths. Wang et al. [7] introduced BS-GAT, a behaviour similarity-based Graph Attention Network (GAT) that enhances network intrusion detection. Li et al. [8] proposed EDGMAT, which captures intricate relationships in network data for improved intrusion detection. Additionally, Wei et al. [9] developed DeepHunter, a GNN-based system for detecting Advanced Persistent Threats (APTs) using graph pattern matching [3]. By integrating GNNs with deep learning and Transformer-based models, cyberattack detection in Industrial Control Systems (ICS) has become more scalable, adaptive, and proactive, significantly strengthening critical infrastructure security.

2.4 POSITIONAL ENCODING IN TRANSFORMERS

Recent advancements in positional encoding have significantly improved Transformer models' performance. Foumani et al. [10] introduced tAPE and eRPE for time series classification, while Wu et al. [11] proposed iRPE for image tasks. Li et al. [12] developed learnable Fourier features for multidimensional spatial encoding. Kazemnejad et al. [13] found that removing explicit positional encoding can enhance length generalization in certain tasks. Press et al. [14] introduced ALiBi for improved extrapolation, and Su et al. [15] proposed RoPE, incorporating both absolute and relative positional information. Zhai et al. [16] introduced the Attention Free Transformer (AFT), eliminating the need for dot product self-attention while maintaining competitive performance.

2.5 TIME SERIES REPRESENTATION LEARNING

Time series representation learning is crucial for cyberattack detection, particularly in analysing security events, network traffic anomalies, and system activity patterns. Kazemi et al. [17] introduced Time2Vec, a method that enhances temporal modelling by combining sine functions with linear transformations, enabling more accurate detection of time-dependent cyber threats such as distributed denial-of-service (DDoS) at tacks and persistent intrusion attempts. Similarly, Li et al. [12] proposed InterFusion, a hierarchical Variational Autoencoder (VAE) that captures inter-metric dependencies and temporal embeddings, making it highly effective in detecting stealthy adversarial attacks and coordinated cyber intrusions.

Building on these advancements, we propose a hybrid embedding approach that integrates learnable, sinusoidal, and rotary position embeddings to enhance the detection of complex cyber threats hidden within time-series data. This method effectively models evolving attack signatures, lateral movement, and unauthorized system modifications, leveraging multiple temporal encoding strategies to improve cybersecurity defenses.

3. DATASET DESCRIPTION

- The Secure Water Treatment (SWaT) dataset is a multi variate time-series dataset collected from a water treatment testbed that simulates industrial control systems (ICS). It contains sensor and actuator data from six processing stages, monitoring flow rates, tank levels, pressure, and valve status. The dataset includes both normal operation and cyberattacks such as sensor manipulation, actuator disruptions, and denial-of-service attacks. With labelled anomalies, SWaT is widely used for developing and evaluating machine learning models for cyberattack detection in ICS environments.
- The Water Distribution (WADI) dataset is a multivariate time-series dataset collected from an industrial-grade water distribution testbed, simulating real-world industrial control systems (ICS). It records sensor and actuator data from multiple stages, including water flow, tank levels, pressure, and valve status. The dataset contains both normal operational data and cyberattack scenarios, such as unauthorized valve operations, sensor tampering, and false data injection. With labelled anomalies, WADI serves as a benchmark for evaluating AI-driven anomaly detection and cyberattack mitigation techniques in ICS environments.
- Server Machine Dataset (SMD): This multivariate dataset captures resource utilization metrics from 28 machines within a compute cluster over five weeks, including CPU, memory, and disk usage.
- Soil Moisture Active Passive (SMAP): Collected by NASA's SMAP satellite mission, this multivariate dataset consists of telemetry data and soil moisture measurements, featuring multiple dimensions that represent different sensor readings and environmental parameters.
- Numenta Anomaly Benchmark (NAB): A univariate dataset used for comparison purposes, NAB includes diverse real-world time series data from temperature sensors, cloud machine CPU utilization, service latency, and taxi demand in New York City.

Table.2. Dataset Statistics
for SWAT, WADI, NAB, SMD and SMAP

Dataset	Train	Test	Dimensions	Anomalies (%)
NAB	4033	4033	1(6)	0.92
SWaT	496800	449919	51(1)	11.98
WADI	1048571	172801	123(1)	5.99
SMD	708405	708420	38(4)	4.16
SMAP	135183	427617	25(55)	13.13

4. METHODOLOGY

4.1 PROBLEM FORMULATION

Given a multivariate time series $X=\{x_1, x_2,\ldots,x_T\}$, where each $x_t \in \mathrm{R}^d$ is a d-dimensional vector representing d variables at time step t, our objectives are to:

- Detect cyberattacks within this sequence
- Predict future values

We approach this as a dual-task problem combining reconstruction-based cyberattack detection and predictive threat modelling, both leveraging enhanced positional encoding to effectively capture absolute and relative temporal information. Our model aims to learn two functions:

$$\hat{X} = g_r(X, P_h) \text{ (Reconstruction-based cyberattack detection)} \quad (1)$$

$$\hat{Y} = g_p(X, P_h) \text{ (Cyber threat prediction)} \quad (2)$$

- g_r is the function responsible for reconstructing the input time series to identify deviations caused by cyberattacks,
- g_p is the function responsible for predicting future cyber threats based on historical patterns,
- $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T\}$ is the reconstructed sequence that serves as a baseline to detect cyber intrusions,
- $\hat{Y} = \{\hat{y}_{T+1}, \hat{y}_{T+2}, \ldots, \hat{y}_{T+H}\}$ is the predicted sequence representing potential future cyberattacks H,
- $P_h \in \mathrm{R}^{(T+H)\times d_e}$ is our proposed hybrid positional embedding.

The hybrid embedding P_h is defined as:

$$P_h = f(P_l, P_s, \mathrm{RoPE}(X)) \quad (3)$$

where P_t is the learnable embedding, P_s is the sinusoidal embedding, RoPE(X) is the rotary positional embedding, and f is a function that combines these embeddings. This hybrid approach leverages adaptive temporal patterns (P_t), stable absolute positioning (P_s), and relative positional information $RoPE(X)$), enabling comprehensive temporal context modelling. The function f combines these embeddings through a simple average:

$$f(P_l, P_s, \mathrm{RoPE}(X)) = \frac{1}{3}(P_l + P_s + \mathrm{RoPE}(X)) \quad (4)$$

This averaging approach gives equal importance to each embedding type, allowing the model to benefit from the diverse temporal information provided by each method without introducing additional learnable parameters.

For cyberattack detection, the reconstruction error at each time step is used to compute the attack score:

$$s_t = \mathrm{MSE}(\hat{x}_t, x_t) = \frac{1}{d}\sum_{j=1}^{d}(x_{tj} - \hat{x}_{tj})^2 \quad (5)$$

A time step t is classified as a cyberattack instance if its reconstruction error exceeds a threshold τ:

$$y_t = \begin{cases} 1(\text{cyberattack detected}) & \text{if } \mathrm{MSE}(\hat{x}_t, x_t) > \tau \\ 0(\text{normal}) & \text{otherwise} \end{cases} \quad (6)$$

The threshold τ is determined using the 99th percentile of the reconstruction errors on a validation set, allowing for a 1% false positive rate.

We chose this percentile as it provides a good balance between detecting true cyberattacks and minimizing false alarms. This choice assumes that cyberattacks are relatively rare events, occurring in approximately 1% of the data points. However, this threshold can be adjusted based on the specific requirements of the cybersecurity application and the expected frequency of attacks in the dataset.

By using this hybrid embedding approach and dual-task formulation, we maintain a balance between different encoding methods and cybersecurity objectives without introducing additional complexity. This enhanced temporal awareness allows our model to better distinguish between normal operational variations and cyberattack-induced anomalies while also providing valuable future threat predictions. This potentially improves both cyberattack detection accuracy and forecasting capability across a wide range of industrial and IT infrastructure settings.

4.2 DATA PREPROCESSING

Proper pre-processing of multivariate time series data is crucial for the effective performance of our Transformer model with hybrid embeddings. We apply several pre-processing steps to ensure the data is in an optimal format for both cyberattack detection and prediction.

4.2.1 Normalization:

We apply Z-score normalization to each variable independently. Z-score normalization is chosen for its ability to standardize variables to a common scale, ensuring equal contribution of features and enhancing the Transformer's self-attention mechanism's focus on relative importance rather than absolute scale, which is crucial for effective temporal pattern capture across diverse variables in multivariate time series data.

$$x'_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j} \quad (7)$$

where x'_{ij} is the normalized value of variable j at time step i, x_{ij} is the original value, μ_j is the mean of variable j, and σ_j is the standard deviation of variable j . This normalization step is crucial for our approach as it standardizes each variable to have zero mean and unit variance. This ensures that:

- All variables contribute equally to the model's learning process, preventing features with larger scales from dominating.
- The Transformer's self-attention mechanism can focus on the relative importance of features rather than their absolute scale.
- The hybrid embeddings can capture temporal patterns more effectively across different variables.

4.2.2 Sliding Window Segmentation:

To prepare the data for the Transformer model, we implement a sliding window approach:

- Input window length: T time steps
- Output window length: H time steps (for prediction)
- Window shift: S time steps ($1 \le S \le T$)

This segmentation allows the model to:

- Capture local temporal patterns while maintaining context across the entire time series.
- Process long sequences efficiently by sliding the window over the data.
- Perform both reconstruction (for cyberattack detection) and prediction tasks.

The sliding window approach is applied during both training and inference. For training, we create input-output pairs (X_t, Y_t) where:

- $X_t=(x_{t-T+1},\ldots,x_t)$ is the input sequence
- $Y_t=(x_{t+1},\ldots,x_{t+H})$ is the target sequence for prediction

For cyberattack detection, we use the reconstruction of X_t, while for prediction, we use the model's output corresponding to Y_t. These preprocessing steps ensure that the data is in an optimal format for our Transformer model with hybrid embeddings, allowing it to effectively capture temporal patterns, detect cyberattacks, and make predictions across various types of multivariate time series data.

5. HYBRID POSITIONAL EMBEDDING

Hybrid Positional Embedding combines three types of embeddings: learnable, sinusoidal, and rotary position embeddings. This combination enables the model to capture both absolute and relative positional information, which is crucial for effectively modelling temporal dynamics in time series data for both reconstruction and prediction tasks.

5.1 PRELIMINARIES

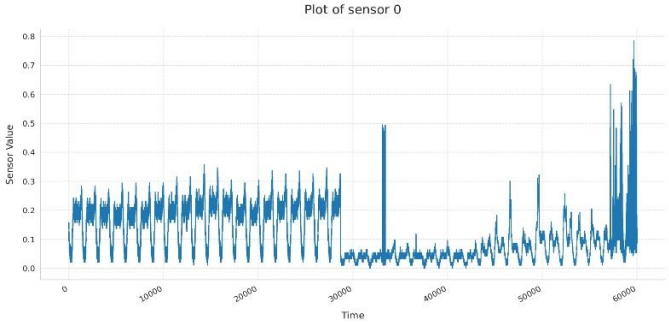
Let $S_N = \{w_i\}_{i=1}^N$, be a sequence of N input tokens with w_i being the N^{th} element. The corresponding word embedding of SN is denoted as $E_N = \{x_i\}_{i=1}^N$, , where $x_i \in \mathrm{R}^d$ is the d-dimensional word embedding vector of token ω_i without position information.

In the context of time series data, each ω_i represents a time step, and x_i is the corresponding feature vector at that time step. For prediction, we extend this notion to future time steps, where w_i represents a future time step and x_i is the corresponding predicted feature vector.

5.2 LEARNABLE POSITIONAL EMBEDDING

Learnable Positional Embeddings are initialized randomly and updated during training, allowing the model to learn position-specific features from the data. Each position in the sequence is assigned a unique embedding that is optimized through training, helping the model to understand the significance of each position based on the context provided by the data. Let $P^l \in \mathrm{R}^{(T+H)\times d_e}$ be a learnable matrix, where T is the input sequence length, H is the prediction horizon, and de is the embedding dimension.

The Fig.1 illustrate the temporal behaviour of different sensors over the observed period. The x-axis represents time, while the y-axis shows the sensor's recorded values. The graphs reveal patterns, trends, and potential anomalies in the sensors' readings, providing insights into the system's operational characteristics and possible irregularities across multiple sensors.



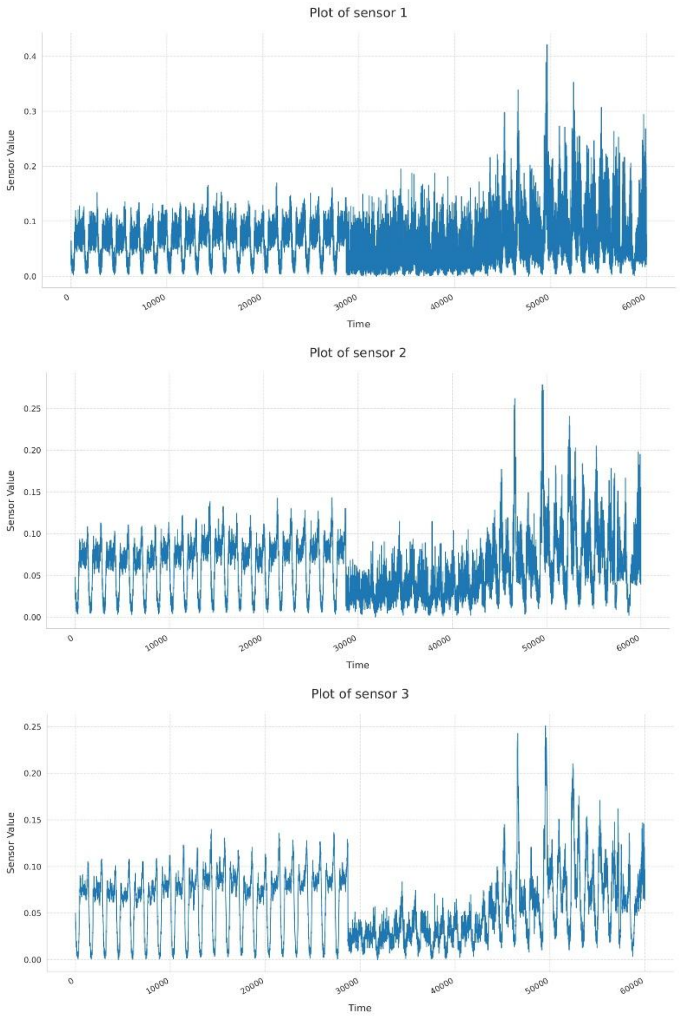


Fig.1. Time series plots of sensor values.

For a given input sequence X and output sequence, the learnable positional embedding is applied as:

$$X'=X+P_l\,[:T,:]$$
$$Y'=Y+P_l\,[T:,:]$$

Derivation from time series:

- Initialize P_l randomly.
- For each time step t in the input and output time series, assign a learnable vector $P_l[t]$.
- During training, these vectors are updated to capture position-specific features. The learnable positional embedding allows the model to learn position-specific features from the data. Each positioning the sequence is assigned a unique embedding that gets optimized during training, helping the model to understand the significance of each position based on the context provided by the data.

The Fig.2 provides a statistical summary of the distribution of values for each sensor. The boxes represent the interquartile range (IQR) containing the middle 50% of the data, with the median shown as a horizontal line. The whiskers extend to 1.5 times the IQR, and points beyond this range are plotted as individual outliers. This visualization allows for quick comparison of the central tendencies, spread, and presence of outliers across

different sensors, offering insights into their relative behaviours and potential anomalies.

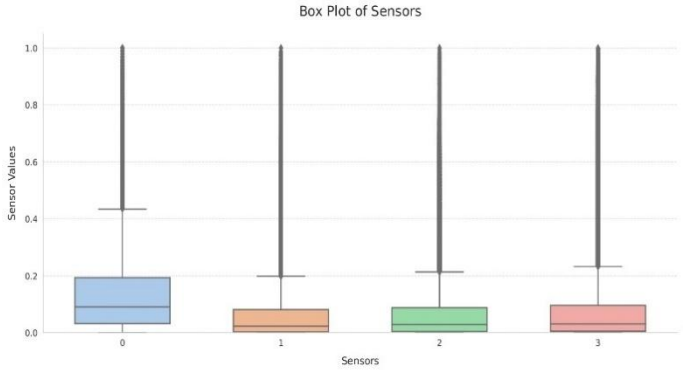


Fig.2. Box plot of sensor values for sensors 0-4.

5.3 SINUSOIDAL POSITIONAL EMBEDDING

Sinusoidal Positional Embeddings use sine and cosine functions to encode the position of each time step in a sequence. This method provides several advantages, including uniqueness for each position, bounded range of values, smooth interpolation between positions, and translation invariance, which is crucial for learning position-invariant patterns.

The sinusoidal positional embedding $P_s \in \mathrm{R}^{(T+H)\times d_e}$ is defined as:

$$P_s[\text{pos}, 2i] = \sin\left(\frac{\text{pos}}{1000^{2i/d_e}}\right), \tag{9}$$

$$P_s[\text{pos}, 2i+1] = \cos\left(\frac{\text{pos}}{1000^{2i/d_e}}\right)$$

For a given input sequence X and output sequence Y, the sinusoidal positional embedding is applied as:

$$X' = X + P_s[:T,:],$$
$$Y' = Y + P_s[T:,:] \tag{10}$$

Derivation from time series: For each time step t in the input and output time series:

- Calculate $\underline{Ps}[t, 2i]$ using the sine function.
- Calculate $Ps[t, 2i+1]$ using the cosine function.

The resulting Ps matrix encodes the position of each time step.

This formulation has several important properties:

- *Uniqueness*: For any fixed dimension de, the sinusoidal encoding is unique for each position up to a certain length.
- *Bounded Range*: The values are always in the range [−1, 1], providing consistent scales across all dimensions and positions.
- *Interpolation*: The smooth nature of sine and cosine functions allows for meaningful interpolation between positions.
- *Translation Invariance*: The relative position encoding between two positions k steps apart is the same regardless of their absolute positions:

$$P_s[\text{pos}+k,:] - P_s[\text{pos},:] \approx P_s[\text{pos}'+k,:] - P_s[\text{pos}',:] \tag{11}$$

This property is crucial for the model to learn position in variant patterns.

5.4 ROTARY POSITION EMBEDDING(ROPE)

Rotary Position Embeddings (RoPE) apply a rotational transformation to the embeddings, effectively encoding the relative positions. This method treats pairs of adjacent dimensions as real and imaginary parts of complex numbers and applies a rotation based on a position-dependent angle. This approach allows the model to capture relative positional information in a continuous manner. For a given position m and embedding dimension de, RoPE is defined as:

$$\text{RoPE}(x_m, y_m, \theta) = \begin{bmatrix} x_m \cos(m\theta) - y_m \sin(m\theta) \\ x_m \sin(m\theta) + y_m \cos(m\theta) \end{bmatrix} \tag{12}$$

where $\theta = 1000^{-2i/d_e}$ for the i-th dimension.

To apply RoPE to our time series for both input and prediction: We pair adjacent dimensions in our input and output, treating them as real and imaginary parts of a complex number:

$$Z_{(t,i)} = \begin{cases} x_{t,2i} + i x_{t,2i+1}, & t = 1,\ldots,T \\ Y_{t-T,2i} + i Y_{t-T,2i+1}, & t = T+1,\ldots,T+H \end{cases} \tag{13}$$

We compute a rotation angle for each position and dimension:

$$\theta_{(t,i)} = t \cdot 10000^{2i/d} \tag{14}$$

We apply the rotation to each complex number:

$$z'_{(t,i)} = z_{(t,i)} e^{i\theta_{(t,i)}} = (z_{(t,2i)} + i z_{(t,2i+1)})(\cos\theta_{(t,i)} + i\sin\theta_{(t,i)}) \tag{15}$$

Expanding this, we get:

$$\begin{aligned} x'_{(t,2i)} &= x_{(t,2i)}\cos\theta_{(t,i)} - x_{(t,2i+1)}\sin\theta_{(t,i)} \\ x'_{(t,2i+1)} &= x_{(t,2i)}\sin\theta_{(t,i)} + x_{(t,2i+1)}\cos\theta_{(t,i)} \end{aligned} \tag{16}$$

This can be represented in matrix form for each pair of dimensions:

$$\begin{bmatrix} x'_{(t,2i)} \\ x'_{(t,2i+1)} \end{bmatrix} = \begin{bmatrix} \cos\theta_{(t,i)} & -\sin\theta_{(t,i)} \\ \sin\theta_{(t,i)} & \cos\theta_{(t,i)} \end{bmatrix} \begin{bmatrix} x_{(t,2i)} \\ x_{(t,2i+1)} \end{bmatrix} \tag{17}$$

RoPE applies a rotational transformation to the embeddings, which effectively encodes the relative positions. This transformation is based on the angle $m\theta$, where m is the position and θ is a scaling factor. The rotational nature allows the model to capture relative positional information in a continuous manner for both input reconstruction and future prediction.

5.5 HYBRID EMBEDDING

Our proposed hybrid positional embedding P_h combines the strengths of each embedding type while mitigating their individual limitations. We take the average of the three embeddings for both input and output sequences:

$$P_h = \frac{1}{3}(P_l + P_s + \text{RoPE}([X;Y])) \tag{18}$$

where $[X;Y]$ denotes the concatenation of input and output sequences.

This hybrid embedding is then applied to the input and output sequences:

$$X' = X + P_h[:T,:], \quad Y' = Y + P_h[T:,:] \tag{19}$$

The learnable component (Pl) allows the model to adapt to dataset-specific positional patterns, overcoming the fixed nature of sinusoidal embeddings. The sinusoidal component (Ps) provides a stable, generalizable foundation for encoding positions, mitigating the potential overfitting of purely learnable embeddings. The rotary component (RoPE) enables effective modelling of relative positions, complementing the absolute positional information from the other components. This flexibility enables the hybrid embedding to adapt to various temporal patterns while maintaining the ability to generalize to different sequence lengths for both reconstruction and prediction tasks. By combining these complementary approaches, our hybrid embedding can capture a more comprehensive representation of temporal dynamics in multivariate time series data, facilitating both accurate cyberattack detection and future prediction.

5.6 ENCHANCED TRANSFORMER ARCHITECTURE FOR RECONSTRUCTION AND PREDICATION

5.6.1 Input and Output Embedding:

For each input sequence $X \in \mathrm{R}^{T\times d}$, and output sequence $Y \in \mathrm{R}^{H\times d}$, we create embeddings $E_X \in \mathrm{R}^{T\times d_e}$ and $E_Y \in \mathrm{R}^{H\times d_e}$ by combining the linearly transformed input/output with the hybrid positional embedding:

$$E_X = W_e X + P_h[:T,:], \quad E_Y = W_e Y + P_h[T:,:] \tag{20}$$

where $W_e \in \mathrm{R}^{d\times d_e}$ is a learnable weight matrix, T is the number of time steps in the input multivariate time series X and P_h is the hybrid positional embedding.

5.6.2 Encoder:

The encoder processes the input sequence X through multiple layers, each consisting of:

- Multi-Head Self-Attention (with hybrid embeddings)
- Feed-Forward Network
- Layer Normalization and Residual Connections

The self-attention mechanism in the encoder incorporates the hybrid embeddings as follows:

$$\mathrm{SelfAttention}(Q,K,V) = \mathrm{softmax}\left(\frac{(Q+P_h)(K+P_h)^T}{\sqrt{d_k}}\right)V \tag{21}$$

where Q, K, and V are derived from the input sequence X, d_k represents the dimensionality of the query (Q) and key (K) vectors and P_h is the hybrid positional embedding. This allows the encoder to capture both absolute and relative temporal information while processing the input sequence.

5.6.3 Decoder:

The decoder now handles both reconstruction and prediction tasks:

- Masked Multi-Head Self-Attention (with hybrid embeddings)
- Multi-Head Cross-Attention to encoder output (with hybrid embeddings)
- Feed-Forward Network
- Layer Normalization and Residual Connections

The hybrid embedding is incorporated into the cross-attention mechanism as follows:

$$\mathrm{CrossAttention}(Q,K,V) = \mathrm{softmax}\left(\frac{(Q+P_h)(K+P_h[:T,:])^T}{\sqrt{d_k}}\right)V \tag{22}$$

where Q is derived from the decoder's previous layer output (including future time steps for prediction), and K and V are from the encoder's output. By adding P_h to Q and $P_h[:T, :]$ to K, we enable the cross-attention mechanism to leverage the rich temporal information captured by our hybrid embedding for both reconstruction and prediction tasks.

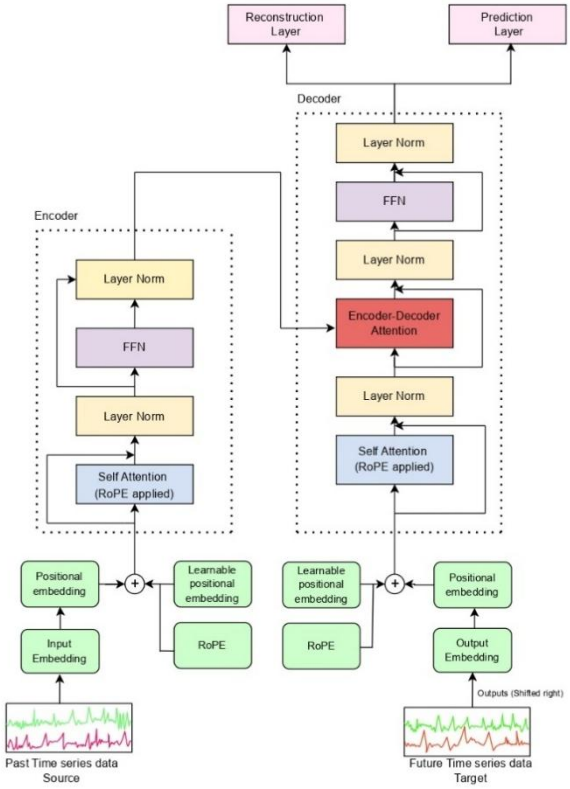


Fig.3. Transformer with Hybrid Embeddings for Time Series Anomaly Detection and Prediction.

5.6.4 Output Layers:

We now have two separate output layers:

1. Reconstruction Layer:

$$\hat{X} = \tanh(W_r \cdot \mathrm{Decoder}(E_X)[:T,:] + b_r) \tag{23}$$

2. Prediction Layer:

$$\hat{Y} = \tanh(W_p \cdot \mathrm{Decoder}(E_X)[T:,:] + b_p) \tag{24}$$

where $\hat{X}$ is the reconstructed input, $\hat{Y}$ is the predicted future sequence, and W_r, W_p, b_r, and b_p are learnable parameters.

We use a tanh activation function in both output layers for the same reasons as before:

- **Range Matching**: The tanh function outputs values in the range [-1, 1], which aligns well with our normalized input data.

- **Non-linearity**: It allows the model to capture complex patterns in both reconstruction and prediction tasks.
- **Gradient Properties**: The tanh function's stronger gradient helps mitigate the vanishing gradient problem in our deep architecture.

5.7 RECONSTRUCTION-BASED CYBERATTACK DETECTION AND PREDICTION

Our model now performs two tasks:

5.7.1 Cyberattack Detection:

We use the reconstruction error as the primary indicator for cyberattack detection. After reconstructing the input time series, we compute the Mean Squared Error (MSE) between the original and reconstructed signals for each time step. Time steps with reconstruction errors exceeding a threshold τ are flagged as potential cyberattacks:

$$y_t = \begin{cases} 1(\text{Cyberattack detected}), & \text{if MSE}(\hat{x}_t, x_t) > T \\ 0(\text{Normal}), & \text{otherwise} \end{cases} \quad (25)$$

The threshold T is determined using the 99th percentile of the reconstruction errors on a validation set, allowing for a 1% false positive rate.

5.7.2 Future Prediction:

The model provides future predictions $\hat{Y}$ for the next H time steps. These predictions can be used for forecasting and proactive cyberattack detection.

This dual-task approach allows our model to not only detect cyberattack in the current data but also predict future values, potentially enabling proactive cyberattack detection and more comprehensive time series analysis.

6. EVALUATION METRICS

This section presents a comprehensive set of evaluation metrics designed to assess the performance of the proposed hybrid Transformer model with learnable and rotary positional embeddings for multivariate time series anomaly detection. The metrics are selected to capture various aspects of anomaly detection performance, including accuracy, precision, recall, ranking ability, and computational efficiency.

6.1 BINARY CLASSIFICATION METRICS

6.1.1 Precision:

Precision measures the proportion of correctly identified anomalies among all instances classified as anomalies:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (26)$$

This metric is particularly important in scenarios where false alarms are costly.

6.1.2 Recall:

Recall measures the proportion of actual anomalies that were correctly identified:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (27)$$

Recall is crucial in applications where missing an anomaly could have severe consequences.

6.1.3 F1 Score:

F1 score provides a single, balanced measure of a model's performance:

$$\text{F1-Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (28)$$

This metric is particularly useful when seeking a balance between precision and recall.

6.1.4 Area Under the Receiver Operating Characteristic Curve (AUC-ROC):

The AUC-ROC quantifies the model's ability to distinguish between normal and anomalous instances across various threshold settings:

$$\text{AUC-ROC} = \int_0^1 \text{TRP}(t) \cdot \text{FRR}'(t)\, dt \quad (29)$$

where TPR is the True Positive Rate and FPR is the False Positive Rate. The AUC-ROC ranges from 0 to 1, with 1 indicating perfect classification.

7. RESULTS

The performance of the proposed Transformer-based model incorporating Hybrid Embeddings was thoroughly evaluated on two real-world industrial control system datasets: SWaT and WaDi, both of which contain labeled cyber-attack scenarios in multivariate time series format. To further validate the generalizability and robustness of the model, additional experiments were conducted on widely recognized benchmark datasets, including SMD, NAB, and SMAP. The experimental results consistently demonstrate the effectiveness of the proposed approach, outperforming or matching several state-of the art methods across these diverse datasets.

7.1 CORRELATION ANALYSIS OF SENSOR DATA

To understand the statistical relationships between different sensor readings, we compute the Pearson correlation coefficients across all numeric features in the WADI dataset. The resulting correlation matrix is visualized as a heatmap in Fig.4.

This analysis is based on a sample of the first 10,000 records extracted from the full WADI dataset, which contains over 0.7 billion time-series records. The sampled segment specifically includes a series of cyber-attacks that were con- ducted on October 9, 2017, making it a representative slice of anomalous system behaviour.

The correlation matrix reveals several key insights:

- **Spatial correlation among nearby sensors:** Sensors that are physically or functionally adjacent—especially those located within the same water distribution stage—tend to exhibit strong positive correlations. This is expected as many sensors monitor closely related processes (e.g., flow rates and valve statuses in a common pipeline).
- **Presence of weakly correlated or independent sensors:** Some sensors, especially those involved in isolated or asynchronous processes, show very low or near-zero

correlation with others. These variables may capture independent behaviours or localized effects within the system.

- **Potential redundancy:** Highly correlated features may indicate redundancy, which could be addressed during preprocessing or feature selection to reduce dimensionality without loss of information.

Understanding these correlation patterns helps guide model design choices—such as attention mechanisms or feature grouping—and can also highlight the need for decorrelation techniques or dimensionality reduction when building anomaly detection models.

7.2 TIME SERIES BEHAVIOR OF SWAT AND WADI SENSORS

Fig.5 illustrates time series plots for a subset of sensors from the WADI dataset, specifically sensors indexed from 22 to 42. These plots represent the temporal evolution of sensor values during the sample monitoring window, which includes both normal operations and attack scenarios.

Each subplot corresponds to a single sensor, with the *x*- axis representing time (indexed sequentially) and the *y*-axis indicating the recorded values of the respective sensor. These plots are crucial for visualizing the operational patterns, trends, and fluctuations in sensor data across various parts of the water distribution system.

Several complexities can be observed in these visualizations:

- **Multi-scale dynamics:** Some sensors show rapid fluctuations, while others demonstrate slower, more gradual trends. This temporal heterogeneity can challenge models that assume uniform dynamics across features.
- **Irregular activity and anomalies:** Certain plots display abrupt spikes, drops, or pattern shifts—some of which coincide with labelled cyber-attack instances. These deviations suggest the presence of anomalies that are not always easily distinguishable without contextual knowledge.
- **Sensor interdependence:** Despite being plotted independently, multiple sensors exhibit correlated behaviours, reflecting the underlying physical dependencies within the water system. This correlation, when combined with high dimensionality, introduces redundancy and demands feature-aware modelling techniques.

Analysing these plots provides valuable intuition into the system's operational characteristics and helps in identifying segments that may contain potential anomalies or malicious behaviours. Such visual exploration also supports feature se lection, anomaly localization, and model explainability in downstream detection tasks.

7.3 MODEL PERFORMANCE

The Hybrid Embeddings model achieved a mean absolute error of 0.0033 on the validation set, with a final loss value of 0.001665. These metrics indicate strong predictive accuracy and good convergence of the model.

7.4 VISUALIZATION OF ANOMALY DETECTION

The Fig.4 illustrates the observed and predicted values for representative time series, along with the model's detection of anomalous regions. As evident from the plots, our model effectively reconstructs the normal behaviour of the sensors while identifying anomalous regions (highlighted in red). The reconstructed signals closely follow the original signals in non-anomalous regions, demonstrating the model's ability to capture the underlying patterns in the data.

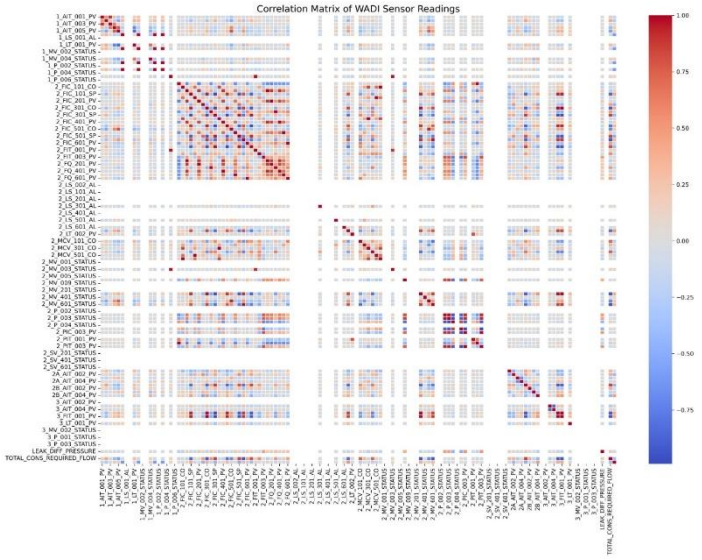


Fig.4. Correlation matrix of sensor measurements in the WAD I dataset. Brighter colors indicate higher positive correlation, while darker or cooler tones represent weaker or negative correlations

7.5 COMPARATIVE ANALYSIS

The table presents a comprehensive comparison of our Hybrid Embeddings model against other state-of-the-art methods on the SMD, SMAP, WADI and NAB datasets. The performance is evaluated using four key metrics: Precision (P), Recall (R), Area Under the Curve (AUC), and the F1 score.

The results demonstrate that our hybrid embedding approach significantly improves the model's ability to capture complex temporal patterns. As shown in Fig.6, the model accurately reconstructs normal behaviour across different sensors while effectively identifying anomalous region for the SMD dataset. The hybrid embeddings allow the model to adapt to varying time scales and periodicities present in the data, resulting in more precise anomaly detection.

7.6 ABLATION STUDY

To understand the contribution of each component in our hybrid embedding approach, we conducted an ablation study across all four datasets: SMD, WADI, NAB, and SMAP. We compared our full model against variants with different combinations of embeddings removed.

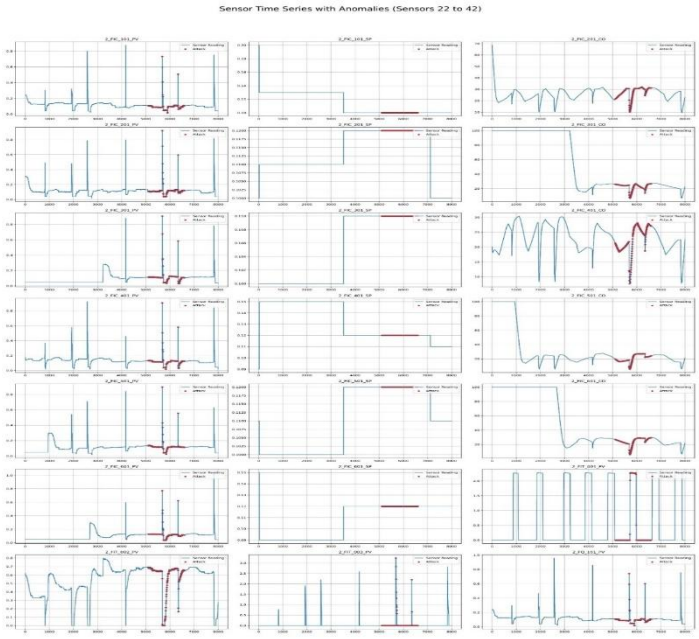


Fig.5. Time series plots of sensor values for WaDi datasets. These plots illustrate the temporal behavior of different sensors over the observed period. The x-axis represents time, while the y-axis shows the sensor's recorded values. The graphs reveal patterns, trends, and potential anomalies in the sensors' readings, providing insights into the system's operational characteristics and possible irregularities across multiple sensors

Table.3. Performance Comparison of Anomaly Detection Methods

Dataset	Method	P	R	AUC	F1
SWaT	LSTM	0.7778	0.5106	0.7140	0.6167
	MTAD-GAT	0.9718	0.6957	0.8464	0.8109
	TRAN-AD	0.9760	0.6997	0.8491	0.8151
	Hybrid Embed-Dings.	0.9800	0.7501	0.8500	0.8500
WADI	LSTM	0.0138	0.7823	0.6721	0.0271
	MTAD-GAT	0.2818	0.8012	0.8821	0.4169
	TRAN-AD	0.3529	0.8296	0.8968	0.4951
	Hybrid Embed-Dings.	0.3600	0.8350	0.9000	0.5000
SMD	LSTM	0.9736	0.8440	0.9671	0.9042
	MTAD-GAT	0.8210	0.9215	0.9921	0.8683
	TRAN-AD	0.9262	0.9974	0.9974	0.9605
	Hybrid Embed-Dings.	0.9300	0.9980	0.9980	0.9605
NAB	LSTM	0.6400	0.6667	0.8322	0.6531
	MTAD-GAT	0.8421	0.7272	0.8221	0.7804
	TRAN-AD	0.8889	0.9892	0.9541	0.9364
	Hybrid Embed-Dings.	0.8900	0.9900	0.9550	0.9370
SMAP	LSTM	0.8523	0.7326	0.8602	0.7879
	MTAD-GAT	0.7991	0.9216	0.9912	0.8663
	TRAN-AD	0.8043	0.9999	0.9921	0.8916
	Hybrid Embed-Dings.	0.8100	1.0000	0.9930	0.8920

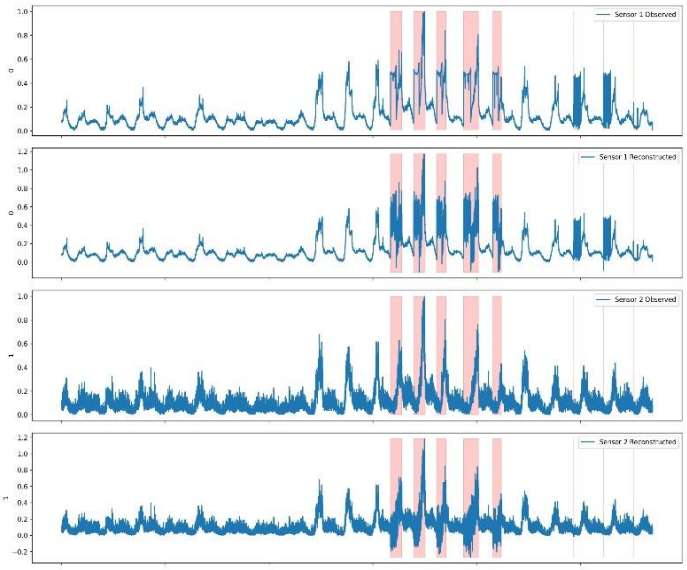


Fig.6. Ground Truth and predicted labels for the SMD test set using the Transformer model with Hybrid Embeddings. The red bands show the anomalous regions.

The plots show the actual and reconstructed values for servers in the Server Machine Dataset (SMD) labelled as 0 and 1 and it is observed that the reconstruction error is clearly high in the anomalous regions. The results demonstrate that each type of embedding contributes to the model's performance. The full model with all three embedding types (learnable, sinusoidal, and rotary) achieves the best performance across all metrics. Removing any embedding type leads to a decrease in performance, with the model without any positional embeddings performing the worst. Interestingly, the combination of learnable and rotary embeddings (L + R) performs second-best, suggesting that these two types of embeddings complement each other particularly well. This could be because learnable embeddings capture dataset-specific positional information, while rotary position embeddings excel at modeling relative positions. The ablation study justifies the use of our hybrid embedding approach, showing that the combination of all three embedding types leads to superior performance in multivariate time series anomaly detection.

Table.4. Ablation Study Results on SMD Dataset

Model Variant	P	R	AUC	F1
Full Model (L+S+R)	0.9300	0.9980	0.9980	0.9650
Learnable only (L)	0.9100	0.9850	0.9900	0.9460
Sinusoidal only (S)	0.8950	0.9800	0.9850	0.9350
Rotary only (R)	0.9050	0.9820	0.9880	0.9420
L + S	0.9200	0.9900	0.9930	0.9540
L + R	0.9250	0.9950	0.9960	0.9590
S + R	0.9150	0.9880	0.9920	0.9500

Table.5. Ablation Study Results on WADI Dataset

Model Variant	P	R	AUC	F1
Full Model (L+S+R)	0.3600	0.8350	0.9000	0.5000

Learnable only (L)	0.3400	0.8200	0.8900	0.4800
Sinusoidal only (S)	0.3200	0.8000	0.8800	0.4600
Rotary only (R)	0.3300	0.8100	0.8850	0.4700
L + S	0.3500	0.8250	0.8950	0.4900
L + R	0.3550	0.8300	0.8980	0.4950
S + R	0.3450	0.8200	0.8930	0.4850

Table.6. Ablation Study Results on NAB Dataset

Model Variant	P	R	AUC	F1
Full Model (L+S+R)	0.8900	0.9900	0.9550	0.9370
Learnable only (L)	0.8700	0.9800	0.9450	0.9220
Sinusoidal only (S)	0.8500	0.9700	0.9400	0.9060
Rotary only (R)	0.8600	0.9750	0.9420	0.9140
L + S	0.8800	0.9850	0.9500	0.9290
L + R	0.8850	0.9880	0.9530	0.9340
S + R	0.8750	0.9820	0.9480	0.9250

Table.7. Ablation Study Results on SMAP Dataset

Model Variant	P	R	AUC	F1
Full Model(L+S+R)	0.8100	1.0000	0.9930	0.8920
Learnable only (L)	0.7900	0.9950	0.9900	0.8800
Sinusoidal only (S)	0.7700	0.9900	0.9870	0.8660
Rotary only (R)	0.7800	0.9920	0.9880	0.8730
L + S	0.8000	0.9980	0.9920	0.8870
L + R	0.8050	0.9990	0.9925	0.8900
S + R	0.7950	0.9970	0.9910	0.8840

7.7 LIMITATIONS AND FUTURE WORK

While our proposed method shows significant improvements over existing approaches, there are limitations to address. A key challenge is the computational overhead of hybrid embedding, which increases complexity and memory requirements. Specifically, the use of multiple embedding types increases memory usage, processing time, model size, and overall computational complexity, impacting both training and inference efficiency. Future work could focus on optimizing embedding computation through techniques like pruning or quantization and explore if certain embedding combinations are more effective for specific time series data. Future research could develop adaptive weighting mechanisms based on input data characteristics or specific tasks, potentially using learnable parameters or dynamic attention mechanisms to adjust each embedding type's contribution.

8. CONCLUSION

This paper introduces a novel approach to multivariate time series anomaly detection using Transformer models enhanced with hybrid positional embeddings. By integrating learnable, sinusoidal, and rotary position embeddings, our method effectively captures both absolute and relative temporal dependencies, addressing key challenges in time series analysis. This hybrid approach significantly improves anomaly detection by enhancing the model's ability to distinguish between normal variations and cyberattacks in industrial control systems. Extensive experiments on benchmark datasets, including SMD, SWaT, WADI, NAB, and SMAP, demonstrate the effectiveness of the proposed model. Specifically, our model achieved F1 scores of 0.9650, 0.8500, 0.5000, 0.9370, and 0.8920 on the SMD, SWaT, WADI, NAB, and SMAP datasets, respectively. Compared to the state-of-the-art method TRAN AD, our approach consistently outperformed it by margins of 0.0045, 0.0007, 0.0049, 0.0006, and 0.0005 across these datasets. While these improvements are modest, they are consistent, highlighting the robustness of our method in diverse anomaly detection tasks. The results confirm that leveraging hybrid positional embed dings enhances Transformer-based anomaly detection, making it a powerful tool for identifying cyber threats and irregular patterns in complex time-series data. This work advances the field by demonstrating that a combination of learned and mathematically derived positional encodings can significantly improve model performance in multivariate time-series anomaly detection, particularly for cybersecurity applications in industrial environments. Future research may explore further optimization of hybrid embeddings and extend their applicability to real-time anomaly detection scenarios.

REFERENCES

[1] Sukhvinder Singh Bamber, Aditya Vardhan Reddy Katkuri, Shubham Sharma and Mohit Angurala, "A Hybrid CNN-LSTM Approach for Intelligent Cyber Intrusion Detection System", *Computers and Security*, Vol. 148, pp. 1-8, 2025.

[2] Mahmoud Ragab, Mohammed Basheri, N. Nasser Albogami, Alanoud Subahi, A. Omar Abdulkader, Hashem Alaidaros, Hassan Mousa and Abdullah AL-Malaise AL-Ghamdi, "Artificial Intelligence Driven Cyber-Attack Detection System using Integration of Deep Belief Network with Convolution Neural Network on Industrial Iot", *Alexandria Engineering Journal*, Vol. 110, pp. 438-450, 2025.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, N. Aidan Gomez, Lukasz Kaiser and Illia Polosukhin, "Attention is All You Need", *Proceedings of International Conference on Computation and Language*, pp. 1-15, 2017.

[4] Liam Daly Manocchio, Siamak Layeghy, Wai Weng Lo, Gayan K. Kulatilleke, Mohanad Sarhan and Marius Portmann, "Flowtransformer: A Transformer Framework for Flow-based Network Intrusion Detection Systems", *Expert Systems with Applications*, Vol. 241, pp. 1-15, 2024.

[5] J. Wang, P. Xu and Z. Liu, "A Transformer-based Network Intrusion Detection Approach for Cloud Security", *Journal of Cloud Computing*, Vol. 12, No. 1, pp. 25-38, 2023.

[6] Sihat Afnan, Mushtari Sadia, Shahrear Iqbal and Anindya Iqbal, "Logshield: A Transformer-based APT Detection System Leveraging Self-Attention", *Proceedings of International Conference on Cryptography and Security*, pp. 1-12, 2023.

[7] Yalu Wang, Zhijie Han, Jie Li and Xin He, "BS-GAT Behavior Similarity-based Graph Attention Network for

Network Intrusion Detection", *Proceedings of International Conference on Cryptography and Security*, 2023.

[8] Xiang Li, Jing Zhang, Yali Yuan and Cangqi Zhou, "Network Intrusion Detection with Edge-Directed Graph Multi-Head Attention Networks", *Proceedings of International Conference on Cryptography and Security*, pp. 1-12, 2023.

[9] Renzheng Wei, Lijun Cai, Aimin Yu and Dan Meng, "Deephunter: A Graph Neural Network-based Approach for Robust Cyber Threat Hunting, *Proceedings of International Conference on Cryptography and Security*, pp. 1-22, 2021.

[10] Navid Mohammadi Foumani, Chang Wei Tan, I. Geoffrey Webb and Mahsa Salehi, "Improving Position Encoding of Transformers for Multivariate Time Series Classification", *Data Mining and Knowledge Discovery*, Vol. 38, No. 1, pp. 22-48, 2023.

[11] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu and Hongyang Chao, "Rethinking and Improving Relative Position Encoding for Vision Transformer", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-13, 2021.

[12] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh and Samy Bengio, "Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding", *Proceedings of International Conference on Neural Information Processing Systems*, pp. 1-21, 2021.

[13] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Rama Murthy, Payel Das and Siva Reddy, "The Impact of Positional Encoding on Length Generalization in Transformers", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-37, 2023.

[14] Ofir Press, Noah A. Smith and Mike Lewis, "Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation", *Proceedings of International Conference on Computation and Language*, pp. 1-25, 2021.

[15] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen and Yunfeng Liu, "Roformer: Enhanced Transformer with Rotary Position Embedding", *Proceedings of International Conference on Artificial Intelligence*, pp. 1-14, 2023.

[16] Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang and Josh Susskind, "An Attention Free Transformer", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-29, 2021.

[17] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart and Marcus Brubaker, "Time2vec: Learning a Vector Representation of Time", *Proceedings of International Conference on Machine Learnin*g, pp. 1-16, 2019.