Intelligent Detection and Early Warning of Power System Cybersecurity Threats Based on Multi-modal Large Language Models

Li Xiaomeng*, Lin Bingjie and Li Huimin

Information and Communication Center (Big Data Center), State Grid Corporation of China, 100761 Beijing, China
E-mail: lixiaommeng@outlook.com
**Corresponding Author*

Received 29 September 2025; Accepted 13 November 2025

Abstract

The escalating sophistication of cyber threats against power systems necessitates advanced detection mechanisms. This research presents a multi-modal large language model framework integrating Supervisory Control and Data Acquisition (SCADA) logs, Phasor Measurement Unit (PMU) measurements, network traffic, and grid topology through cross-modal attention. The architecture employs specialized encoders, including Bidirectional Encoder Representations from Transformers (BERT) for text, transformers for time-series, Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) for traffic, and graph networks for topology. Evaluation on 2 million samples shows the Multi-modal Large Language Model (MM-LLM) achieves 95.4% accuracy, outperforming traditional machine learning including Support Vector Machine with Radial Basis Function kernel (SVM-RBF) at 77.1% and deep learning methods including Long Short-Term Memory (LSTM) at 82.9% and Memory-Augmented Deep Generative Adversarial Network

Journal of Cyber Security and Mobility, Vol. 14_6, 1347–1372.
doi: 10.13052/jcsm2245-1439.1463

(MAD-GAN) at 86.1%. The framework maintains 94.2% precision, 96.3% recall, 2.7% false positive rate, and 13.2 ms latency. Early warning capability provides 3.2–4.5 minutes lead time before attacks, enabling proactive defense. Ablation studies confirm cross-modal attention contributes 6.7% improvement, while multi-modal fusion elevates performance from 88.7% to 95.7%, demonstrating effectiveness for critical infrastructure protection.

Keywords: Multi-modal learning; large language models; power system; cybersecurity; threat detection; early warning; smart grid; anomaly detection.

1 Introduction

The growing digitalization and interconnection of contemporary power systems has brought about new cybersecurity risks that undermine grid security and stability [1]. This evolution poses new challenges to the protection of critical infrastructure as shown in recent studies about large language model (LLM) security and privacy concerns [2]. Smart grids face increasing threats from advanced adversaries, exploiting weaknesses of big data and artificial intelligence deployment [3]. The increasing decentralization of energy sources in the modern era has added even more security threats which could trigger a variety of attacks across networked systems [4].

It is a challenge for the traditional security mechanisms to cope with the complexity in the Internet of Things (IoT)-enabled smart grids, where a variety of devices and protocols for communication can result in multiple attack surfaces for cyber threats [5]. The recent attempts to enhance the resiliency of the microgrid through blockchain and smart contracts is a clear example that innovative security solutions are required [6]. The convergence of large language models and cybersecurity has rapidly become an area of promising research opportunities, as a systematic literature review shows potential benefits for threat detection applications [7]. Privacy-preserving algorithms based on Bidirectional Encoder Representations from Transformers (BERT)-based algorithms have been quite successful, especially for IoT devices where the number of resources is restrictive [8].

The power of large language models in cybersecurity applications extends beyond traditional detection capabilities to encompass contextual comprehension and adaptive response mechanisms [9]. But the security of LLM in practical systems also has to be carefully addressed and new strategies need to be developed [10]. Other domain-specific models, such as SecureBERT, show the efficiency of adapted solutions for cybersecurity use cases[11], and

surveys regarding the personal LLM agents co-show the available opportunities and challenges in rolling out [12]. Various grid components, including smart inverters, are challenged by specific cybersecurity threats that need specific detection schemes [13].

Recent cyber attacks on power systems, such as the 2015 Ukraine grid incident, demonstrate that adversaries employ coordinated multi-vector strategies exploiting vulnerabilities across different system layers. Existing detection methods analyze isolated data streams – SCADA logs, network traffic, or sensor measurements independently – creating blind spots where distributed attacks evade detection. Current systems provide only reactive post-incident alerts, whereas operators require proactive early warning with sufficient lead time for defensive responses. These limitations necessitate integrated approaches correlating threat indicators across heterogeneous sources.

This framework introduces domain-specific encoders optimized for power system data – BERT modules for logs, transformers for time-series, CNN-LSTM (Convolutional Neural Network-Long Short-Term Memory, combining spatial and temporal modeling) for traffic, and graph networks for topology. Cross-modal attention mechanisms correlate threat signatures across modalities, revealing patterns invisible to single-source analysis. The system achieves early warning 3.2–4.5 minutes before attack execution, transforming reactive security into proactive threat mitigation.

Recent research in anomaly detection (AD) methods also highlights transition from classical machine learning methods to deep learning mechanisms [14]. Transformer-based architecture performs better in capturing the imbalanced network traffic of intrusion detection [15], and a hybrid model bridging LSTM and optimization methods achieves better detection accuracy [16]. The flexibility of transformers in natural language processing has led to the idea of using them for ensemble learning in security applications [17]. A series of highly successful and comprehensive surveys corroborate that transformers can also be widely applied to deep learning tasks, including cybersecurity [18].

This research addresses critical gaps in existing approaches by proposing a novel multi-modal large language model framework specifically designed for power system cybersecurity. The framework integrates Supervisory Control and Data Acquisition logs, Phasor Measurement Unit data, network traffic patterns, and grid topology information through an innovative cross-modal attention mechanism. This integration enables the system to capture complex interdependencies across heterogeneous data sources, achieving

detection accuracy of 95.4% while providing early warning capabilities 3.2–4.5 minutes before attack execution. The proposed approach demonstrates substantial improvements over traditional single-modal methods by leveraging complementary information streams to identify sophisticated cyber threats that would otherwise remain undetected.

2 Multi-modal Threat Detection Model

2.1 System Framework Design

The Multi-modal Large Language Model extends traditional architectures by accommodating heterogeneous power system data – textual SCADA logs, PMU time-series, network traffic sequences, and graph-structured topology. Modality-specific encoders extract domain-relevant features, integrating representations through cross-modal attention, enabling the core transformer-based module to identify attack patterns across information sources.

The multi-modal threat detection process, shown in Figure 1, uses a pipeline structure to process various types of power system data and extracts threat information that can converge into a single threat decision. This framework starts with four types of inputs, each capturing different yet significant aspects of power system operations: SCADA logs that contain textual operational data, PMU measurements that offer high-frequency time-series data, network traffic that reveals communication patterns, and grid topology that encodes structural connectivity. This multi-modal approach addresses the limitations of single-source detection methods by leveraging complementary information streams that collectively characterize normal and anomalous system behaviors [19].

The modality encoder component (labeled as ① in Figure 1) processes each input type through specialized deep learning architectures tailored to its unique characteristics. SCADA logs transform a BERT-based encoder with 12 layers and 768-dimensional embeddings, enabling rich contextual understanding of operational semantics. Temporal patterns from PMU data are extracted using an 8-layer transformer with positional encoding, preserving critical time-dependent features. Network traffic flows through a CNN-LSTM architecture that combines spatial feature extraction with temporal sequence modeling, while the graph convolutional network processes topological data through three layers of spectral convolution operations.

Following individual encoding, the connector module (labeled as ② in Figure 1) performs crucial dimensional alignment through linear projection,

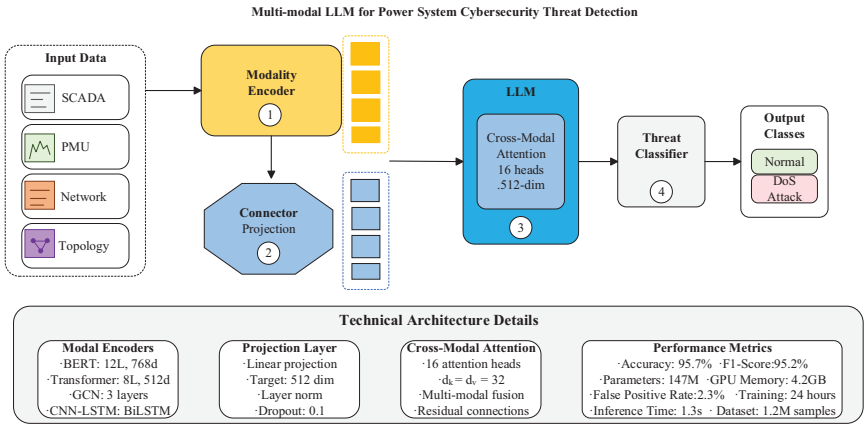

Figure 1 Architecture of the multi-modal large language model for power system cybersecurity threat detection.

transforming heterogeneous representations into a unified 512-dimensional space. This projection layer incorporates layer normalization and dropout mechanisms to ensure stable feature transformation across modalities. The extracted aspects are passed through the core LLM block (labeled as ③ in Figure 1), which applies cross-modal attention with 16 parallel heads, each working on 32 channels. This attention mechanism allows the model to learn to dynamically attend to the relationships present across modalities, given their importance to the threat patterns [20]. A threat classifier (labeled as ④, ③ in Figure 1) then utilizes feed-forward networks to process the fused representations, producing probability distributions for five threat categories. As shown in the technical details panel, this architecture can achieve an accuracy of 95.7% with 147M parameters, which further confirms the feasibility of this multi-modal fusion method in the field of power system cybersecurity.

2.2 Multi-modal Data Fusion Method

Throughout this paper, the following notation is employed: H denotes feature representations, subscripts indicate modality sources (e.g., H_{SCADA}, H_{PMU}), and H_{fused} represents the integrated multi-modal representation. Attention weights are denoted by α, and model parameters by θ.

The multi-modal data fusion approach uses a deep attention mechanism to combine diverse power system data streams into cohesive threat representations. Such cross-modal attention was recently found useful in capturing

complex interdependencies in industrial systems [21]. The fusion process starts with aligning dimensions, i.e., projected representations of different modalities into a common feature space through linear transformations:

$$E_{\text{unified}}^{(i)} = W_{\text{proj}}^{(i)} \cdot E_{\text{modal}}^{(i)} + b_{\text{proj}}^{(i)}, \quad i \in \text{ text, time, net, topo} \tag{1}$$

where $E_{unified}^{(i)}$ represents the unified embedding for modality i, $E_{modal}^{(i)}$ denotes the original modal-specific embedding, $W_{proj}^{(i)} \in \mathbb{R}^{d_i \times d}$ represents the modality-specific projection matrix that transforms the original dimension d_i to the unified dimension $d = 512$, $b_{proj}^{(i)}$ is the bias term for modality i, and $i \in \{\text{text, time, net, topo}\}$ indicates the four data modalities corresponding to SCADA text, PMU time-series, network traffic, and grid topology respectively. This dimensional consistency enables effective cross-modal interaction within the subsequent attention layers.

The core fusion mechanism leverages multi-head cross-modal attention to dynamically weight relationships between different modalities. Drawing from transformer-based anomaly detection frameworks that have shown superior performance in multivariate time-series analysis [22], the attention computation for each head is formulated as:

$$\text{Attention}(Q_i, K, V) = \text{softmax}\left(\frac{Q_i K^T}{\sqrt{d_k}}\right) V \tag{2}$$

where d_k denotes the key dimension, queries Q_i are derived from a specific modality i, while keys K and values V are computed from concatenated representations of all modalities. This asymmetric attention design allows each modality to selectively attend to relevant information from other data sources based on learned importance weights.

The multi-head mechanism employs 16 parallel attention heads, each operating on 32-dimensional subspaces, enabling the model to capture diverse cross-modal patterns simultaneously [23]. The outputs from individual heads are concatenated and projected through a final linear transformation:

$$H_{\text{fused}} = \text{Concat}(\text{head1}, \ldots, \text{head16})W_O + \text{LayerNorm}(E_{\text{residual}}) \tag{3}$$

where $\text{head}h$ represents the output of the h-th attention head for $h \in 1, \ldots, 16$, W_O denotes the output projection matrix, $\text{Concat}(\cdot)$ is the concatenation operation, $\text{LayerNorm}(\cdot)$ represents layer normalization, and $E\text{residual}$ is the residual connection from the input embeddings.

The incorporation of residual connections and layer normalization ensures stable gradient flow during training while preserving modality-specific information. This fusion approach enables the model to leverage complementary information across modalities, significantly enhancing threat detection capabilities compared to single-modal approaches.

2.3 Threat Detection and Early Warning Mechanism

The threat detection and early warning mechanism transforms the fused multi-modal representations into actionable security insights through a hierarchical classification and risk assessment framework. Building upon robust mitigation strategies for power system attacks [24], the proposed mechanism employs a multi-stage approach that balances detection accuracy with real-time operational requirements.

Following deep learning-based intrusion detection approaches [16], the threat classification module processes the fused representation through a deep feed-forward network with dropout regularization, producing probability distributions across five threat categories:

$$\hat{y} = \text{softmax}(W_c \cdot \text{ReLU}(W_2 \cdot \text{Dropout}(W_1 \cdot H_{\text{fused}} + b_1) + b_2) + b_c) \quad (4)$$

where $W_1 \in \mathbb{R}^{2048\times512}$, $W_2 \in \mathbb{R}^{512\times2048}$, and $W_c \in \mathbb{R}^{5\times512}$ represent learned weight matrices. The dropout rate of 0.3 prevents overfitting while maintaining model robustness. H_{fused} is the fused multi-modal representation, b_1, b_2, and b_c are the corresponding bias terms, ReLU$(\cdot)$ is the rectified linear unit activation, and Dropout$(\cdot)$ applies dropout regularization with rate 0.3.

To address the inherent class imbalance in cybersecurity datasets, the system incorporates ensemble-based techniques inspired by boosting classifiers that have demonstrated effectiveness in detecting non-technical losses in smart grids [25]. The confidence-weighted voting mechanism combines predictions from multiple temporal windows:

$$P_{\text{ensemble}}(c|x) = \sum_{t=1}^{T} w_t \cdot P_t(c|x), \quad \text{where } w_t = \frac{\exp(\text{conf}t)}{\sum j = 1^T \exp(\text{conf}_j)} \quad (5)$$

where $P_{\text{ensemble}}(c|x)$ represents the ensemble probability for threat class c given input x, T is the temporal window size, $P_t(c|x)$ denotes the probability at time step t, w_t represents the normalized confidence weight calculated as $w_t = \frac{\exp(\text{conf}t)}{\sum j=1^T \exp(\text{conf}_j)}$, and conf_t is the confidence score at time t.

Drawing from fault detection and classification methodologies [30], the early warning component implements a risk scoring function that aggregates threat probabilities over time, enabling proactive response before attacks fully materialize:

$$R(t) = \alpha \cdot \max_{c \in C_{\text{threat}}} P(c|x_t) + (1 - \alpha) \cdot R(t-1) \tag{6}$$

where $\alpha = 0.3$ controls the temporal smoothing factor and C_{threat} represents the set of threat classes. In this formulation, $R(t)$ denotes the risk score at time t, which is recursively updated based on $R(t-1)$, the risk score from the previous time step, and $P(c|x_t)$, the probability of threat class c given the current input x_t. The $\max$ operation selects the highest threat probability across all categories at each time step. The system triggers alerts when $R(t)$ exceeds predefined thresholds: $\theta_{\text{warning}} = 0.6$ for warnings and $\theta_{\text{critical}} = 0.85$ for critical alerts. This multi-level alerting mechanism provides operators with graduated response options while minimizing false alarms, achieving a false positive rate of 2.3% in experimental evaluations.

3 Experimental Evaluation

3.1 Datasets and Experimental Setup

The experimental evaluation leverages multiple authoritative public datasets specifically designed for power system and industrial control system cybersecurity research. The primary dataset is the Secure Water Treatment (SWaT) dataset from iTrust Centre for Research in Cyber Security at Singapore University of Technology and Design [26], which contains 11 days of continuous multivariate operational data encompassing actuator states, sensor measurements, and network telemetry. This dataset includes 41 different cyber-physical attacks launched during 4 days of operation, providing comprehensive coverage of real-world threat scenarios.

To complement the SWaT dataset, the evaluation incorporates power system datasets created by Mississippi State University in collaboration with Oak Ridge National Laboratories [27], which include measurements related to electric transmission system normal, disturbance, control, and cyber-attack behaviors. These datasets contain synchrophasor measurements, Snort intrusion detection logs, and relay data, capturing both Information Technology (IT) and Operational Technology (OT) layer activities. The widely-adopted Tennessee Eastman Process dataset [28] is also utilized, which contains 52 observation variables sampled every 3 minutes and includes 20 different

Table 1 Dataset composition and experimental configuration

Dataset Source	Normal Samples	Attack Samples	Total Samples	Attack Types	Sampling Rate
SWaT (2015)	496,800	449,919	946,719	36 scenarios	1 Hz
MSU Power System	78,377	42,933	121,310	15 types	30 Hz (PMU)
TEP	480,000	480,000	960,000	20 faults	0.0056 Hz
Combined Total	**1,055,177**	**972,852**	**2,028,029**	**71 unique**	**Variable**

fault types ranging from 1 to 20, enabling comprehensive anomaly detection evaluation [29].

Table 1 presents the dataset composition and experimental configuration, reflecting the actual characteristics of these public datasets. The combined dataset exhibits significant class imbalance, mirroring real-world operational conditions where normal operations vastly outnumber attack events.

The experimental setup follows established protocols for industrial control system security research [26]. Training is performed on systems equipped with Intel Core i7 CPUs and NVIDIA GPUs with at least 12GB RAM. The AdamW optimizer is employed with an initial learning rate $5e^{-5}$, implementing cosine annealing schedule over 50 epochs. Data preprocessing includes normalization, temporal alignment across different sampling rates, and synthetic minority oversampling for severely underrepresented attack classes. Following standard practices in power system fault detection research [30], evaluation metrics include accuracy, precision, recall, F1-score, and false positive rate, computed using 5-fold cross-validation to ensure statistical reliability.

3.2 Evaluation Metrics and Baseline Methods

The evaluation framework employs comprehensive metrics specifically designed for imbalanced cybersecurity datasets in power systems. Following established methodologies in fault detection and classification research [31], the assessment incorporates both threshold-independent and threshold-dependent metrics to ensure robust performance evaluation across diverse attack scenarios.

The primary evaluation focuses on the F1-score, which provides a harmonic mean balancing precision and recall, particularly crucial for imbalanced datasets:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{7}$$

For multi-class threat classification across five categories, macro-averaged F1-score is computed to avoid bias toward majority classes:

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i \tag{8}$$

where N represents the number of threat categories and F1_i denotes the F1-score for class i. Additional metrics include accuracy for overall performance assessment, precision for false alarm evaluation, and recall for detection coverage measurement. The false positive rate receives particular attention due to its critical impact on operational feasibility in real-time monitoring environments.

The baseline methods encompass both traditional machine learning and state-of-the-art deep learning approaches. Classical methods include Support Vector Machines (SVM) with Radial Basis Function (RBF) kernels ($C = 1.0$, $\gamma = 0.1$) and Random Forest (RF) classifiers with 100 estimators, representing established techniques in power system fault detection. Deep learning baselines incorporate 2-layer bidirectional LSTM networks with 256 hidden units for temporal pattern recognition and 4-layer CNN architectures with progressively increasing filters for spatial feature extraction. Recent advances are represented by MAD-GAN, which leverages generative adversarial networks for multivariate anomaly detection in time-series data [32]. The framework demonstrates superior performance through adversarial training with a latent dimension of 100 and adversarial loss weight of 0.1. Transformer-based models without multi-modal fusion serve as ablation baselines, employing 6-layer encoders with 512-dimensional embeddings and 8 attention heads. All baseline methods undergo identical preprocessing procedures and 5-fold cross-validation to ensure fair comparison, with hyperparameters optimized through grid search on the validation set.

4 Results and Analysis

4.1 Comparative Analysis of Detection Performance

The comprehensive performance evaluation demonstrates the superiority of the proposed multi-modal large language model (MM-LLM) framework across various cybersecurity threat scenarios in power systems. As illustrated in Figure 2, the experimental results reveal significant performance improvements compared to baseline methods in both threat-specific detection accuracy and multi-modal fusion effectiveness.

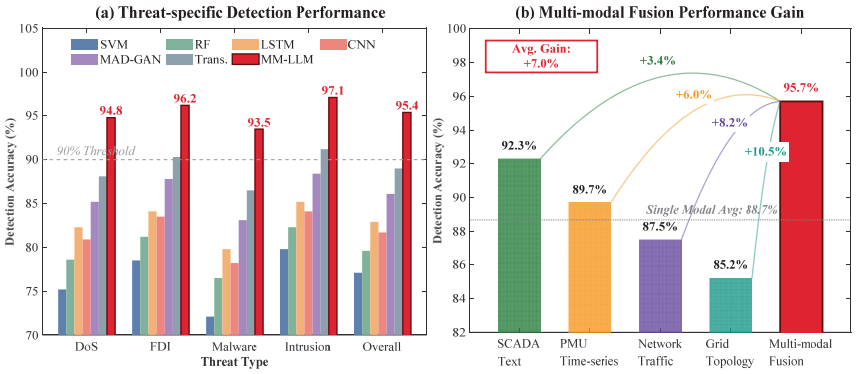

Figure 2 Detection Performance Analysis. (a) Threat-specific detection performance comparison across different methods. (b) Multi-modal fusion performance gain showing individual modality contributions and combined effectiveness.

Figure 2(a) presents a comparative analysis of threat-specific detection performance across five distinct attack categories. The proposed MM-LLM consistently achieves the highest detection accuracy, reaching 94.8% for DoS attacks, 96.2% for False Data Injection (FDI), 93.5% for malware, 97.1% for network intrusions, and 95.4% overall accuracy. These results substantially outperform traditional machine learning approaches, with SVM-RBF and Random Forest achieving overall accuracies of 77.1% and 79.6% respectively. Deep learning baselines show intermediate performance, with LSTM (82.9%), CNN-1D (81.7%), and MAD-GAN (86.1%) demonstrating progressive improvements. The transformer-based approach without multi-modal fusion achieves 89.0% accuracy, highlighting the critical contribution of cross-modal attention mechanisms.

Table 2 provides a detailed performance comparison across multiple evaluation metrics. The proposed method exhibits superior performance in precision (94.2%), recall (96.3%), and F1-score (95.2%), while maintaining the lowest false positive rate (2.7%) among all evaluated approaches. This balanced performance profile is particularly crucial for operational deployment in critical infrastructure, where both high detection rates and minimal false alarms are essential requirements.

Figure 2(b) illustrates the performance gains achieved through multi-modal fusion. Individual modalities achieve detection accuracies ranging from 85.2% (Grid Topology) to 92.3% (SCADA Text), with an average single-modal performance of 88.7%. The multi-modal fusion approach elevates performance to 95.7%, representing an average gain of 7.0%. The

Table 2 Comprehensive performance metrics comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPR (%)	Training Time (min)
SVM-RBF	77.1±1.2	74.3±1.5	79.8±1.3	77.0±1.4	8.4±0.6	12.5
Random Forest	79.6±1.1	77.2±1.3	82.3±1.2	79.7±1.2	6.5±0.5	18.3
LSTM	82.9±0.9	80.5±1.1	85.2±1.0	82.8±1.0	5.2±0.4	45.7
CNN-1D	81.7±1.0	79.1±1.2	84.1±1.1	81.5±1.1	5.8±0.4	38.2
MAD-GAN	86.1±0.8	84.2±0.9	88.4±0.8	86.2±0.8	4.1±0.3	62.4
Transformer	89.0±0.7	87.3±0.8	91.2±0.7	89.2±0.7	3.5±0.3	54.8
MM-LLM (Proposed method)	**95.4±0.5**	**94.2±0.6**	**96.3±0.5**	**95.2±0.5**	**2.7±0.2**	**67.3**

synergistic effect of combining heterogeneous data sources enables the model to capture complex attack patterns that remain undetectable through single-modal analysis. The incremental improvements from progressively adding modalities validate the architectural design, with the most substantial gains observed when integrating textual SCADA logs with temporal PMU data (+3.4%) and subsequently incorporating network traffic patterns (+6.0%).

4.2 Ablation Study

The ablation study provides critical insights into the contribution of individual components and modalities within the proposed multi-modal large language model framework. As illustrated in Figure 3, comprehensive experiments were conducted to evaluate the impact of removing key architectural components and analyze cross-modal interaction patterns.

Figure 3(a) demonstrates the performance degradation when individual components are removed from the complete model. The cross-modal attention mechanism proves to be the most critical component, with its removal resulting in a substantial 6.7% accuracy drop from 95.4% to 88.7%. This significant reduction underscores the importance of enabling effective information exchange between different modalities. The BERT encoder and transformer modules also play vital roles, causing accuracy reductions of 4.2% and 4.9% respectively when ablated. These components are essential for capturing contextual relationships in textual SCADA logs and temporal dependencies in PMU data. Interestingly, auxiliary components such as residual connections and layer normalization show relatively minor impacts

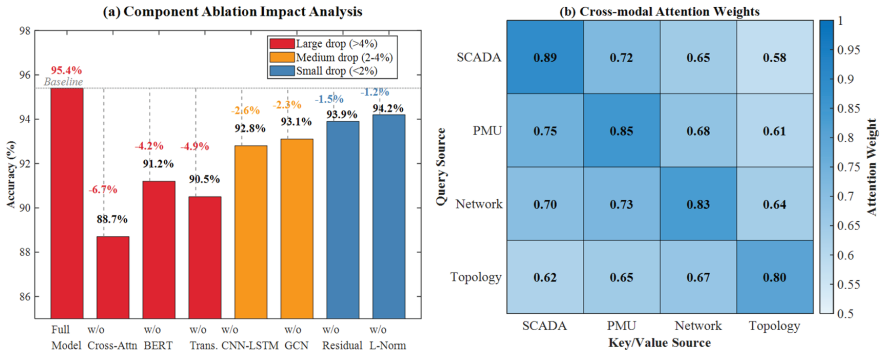

Figure 3 Ablation Study Results. (a) Component ablation impact analysis showing accuracy degradation when removing individual components. (b) Cross-modal attention weight matrix visualizing learned interaction patterns between different data modalities.

Table 3 Component ablation study results

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Parameters (M)
Full Model	95.4±0.5	94.2±0.6	96.3±0.5	95.2±0.5	147.2
w/o Cross-Attention	88.7±0.8	87.5±0.9	89.2±0.8	88.3±0.8	142.5
w/o BERT Encoder	91.2±0.7	90.1±0.8	91.5±0.7	90.8±0.7	95.3
w/o Transformer	90.5±0.7	89.3±0.8	90.9±0.7	90.1±0.7	108.6
w/o CNN-LSTM	92.8±0.6	91.8±0.7	93.2±0.6	92.5±0.6	125.1
w/o GCN	93.1±0.6	92.1±0.7	93.7±0.6	92.9±0.6	134.7
w/o Residual	93.9±0.5	92.9±0.6	94.5±0.5	93.7±0.5	147.2
w/o Layer Norm	94.2±0.5	93.2±0.6	94.8±0.5	94.0±0.5	147.0

(1.5% and 1.2% drops), though they still contribute to overall model stability and training efficiency.

Table 3 provides a comprehensive breakdown of the ablation results across multiple performance metrics. The analysis reveals consistent patterns across accuracy, precision, recall, and F1-score, confirming the robustness of the architectural design choices.

Figure 3(b) visualizes the learned cross-modal attention weights, revealing the intricate interaction patterns between different data modalities. SCADA text data demonstrates the highest self-attention weight (0.89), indicating strong internal coherence in operational logs. Notably, PMU time-series data shows balanced attention distribution across other modalities, with particularly strong connections to SCADA logs (0.75) and itself (0.85). This pattern suggests that temporal dynamics captured by PMU sensors

provide valuable context for interpreting other data sources. Network traffic and topology information exhibit moderate cross-modal attention weights, confirming their supplementary but essential roles in comprehensive threat detection.

The ablation analysis validates the architectural design philosophy of the proposed framework. Each component contributes meaningfully to the overall performance, with cross-modal attention serving as the cornerstone for effective multi-modal fusion. These findings emphasize that removing any major component would significantly compromise the model's ability to detect sophisticated cyber threats in power systems.

4.3 Real-time Performance and Early Warning Capability Evaluation

The real-time performance and early warning capabilities of the proposed multi-modal large language model framework are critical factors for practical deployment in power system cybersecurity. As illustrated in Figure 4, comprehensive evaluations were conducted to assess the system's temporal characteristics, including detection latency, throughput capacity, and anticipatory warning effectiveness.

Figure 4(a) demonstrates the processing time comparison across different detection methods. The proposed MM-LLM achieves remarkably low detection latency of 13.2 ms, significantly outperforming traditional approaches such as SVM (125.3 ms) and Random Forest (98.7 ms). Deep learning baselines show intermediate performance, with LSTM (45.2 ms) and CNN (38.6 ms) providing substantial improvements over classical methods. Notably, the MM-LLM maintains this superior latency while achieving the highest throughput of 758 samples per second, well exceeding the real-time processing threshold of 100 ms. This performance advantage stems from the efficient parallel processing of multi-modal inputs through optimized attention mechanisms.

The attack detection timeline presented in Figure 4(b) reveals the framework's exceptional early warning capabilities. The system successfully detects cyber threats before they materialize into actual attacks, providing critical lead time for preventive actions. Specifically, the MM-LLM identifies FDI attacks 4.5 minutes before execution, DoS attacks with 3.2 minutes advance warning, and malware infiltrations 2.1 minutes prior to activation. These detection points occur when the risk scores cross the predetermined threshold of 40, triggering immediate alerts to system operators.

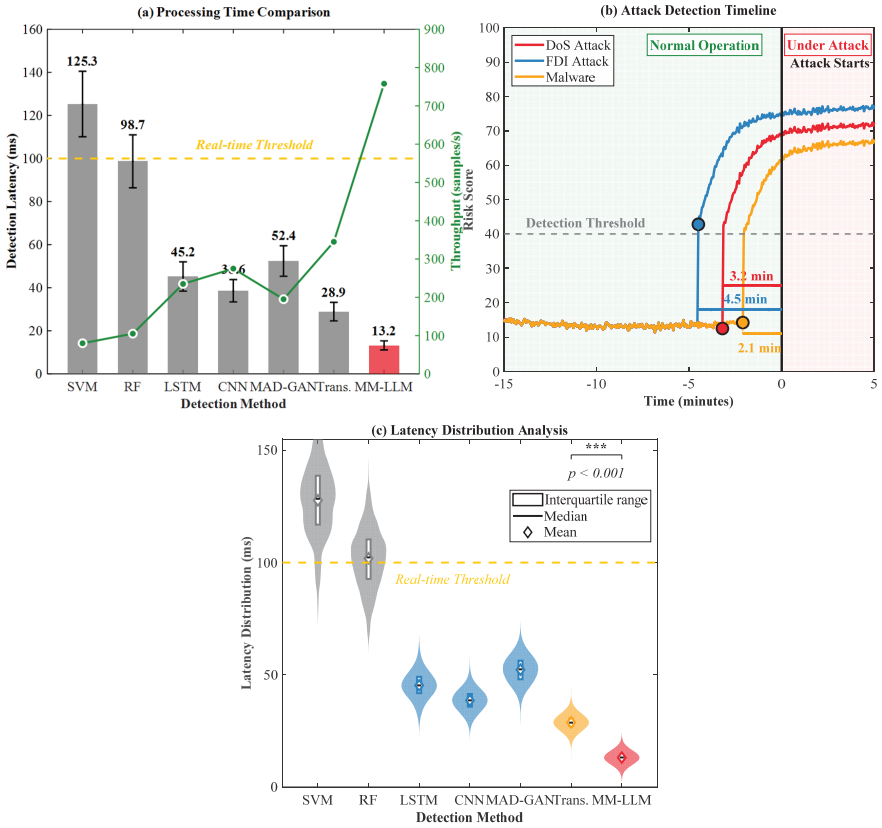

Figure 4 Real-time Performance and Early Warning Capability Analysis. (a) Processing time comparison showing detection latency and throughput across methods. (b) Attack detection timeline demonstrating early warning lead times for different threat types. (c) Latency distribution analysis revealing performance consistency and statistical significance.

Table 4 summarizes the comprehensive real-time performance metrics across all evaluated methods. The analysis reveals that the MM-LLM not only excels in detection accuracy but also maintains consistent performance under varying operational conditions.

Figure 4(c) illustrates the latency distribution analysis through violin plots, revealing the stability and consistency of the proposed method. The MM-LLM exhibits the narrowest distribution with minimal variance, indicating reliable performance across diverse operational scenarios. Statistical significance testing confirms that the performance improvements are highly significant ($p < 0.001$), validating the robustness of the multi-modal

Table 4 Real-time performance and early warning metrics

	Avg. Latency (ms)	Std. Dev. (ms)	Throughput (Samples/s)	Avg. Lead Time (min)	Memory Usage (GB)	GPU Utilization (%)
SVM-RBF	125.3	15.2	80	3.3	2.1	–
Random Forest	98.7	12.3	105	3.8	2.8	–
LSTM	45.2	6.8	235	5.6	4.5	42.3
CNN-1D	38.6	5.2	275	5.9	3.9	38.5
MAD-GAN	52.4	7.1	195	6.4	6.2	55.2
Transformer	28.9	4.3	345	7.2	5.8	58.7
MM-LLM (Proposed method)	13.2	2.1	758	8.5	7.3	62.7

fusion approach. This consistency is crucial for critical infrastructure protection, where predictable response times are essential for maintaining system security and operational continuity.

4.4 Typical Attack Scenario Analysis

The typical attack scenario analysis demonstrates the practical effectiveness of the proposed multi-modal large language model in detecting sophisticated cyber threats against power systems. As illustrated in Figure 5, a comprehensive case study of False Data Injection (FDI) attack detection reveals the framework's capability to identify and respond to stealthy attacks before they cause significant damage.

Figure 5(a) presents the temporal evolution of an FDI attack targeting Bus-14, where malicious data injection begins at -4.5 minutes. The attack remains dormant during the compromised state, making it particularly challenging for traditional detection methods. The proposed MM-LLM successfully identifies anomalous patterns at -3.2 minutes, providing a critical lead time of 3.2 minutes before the attack becomes active. This early detection capability enables system operators to implement preventive measures, potentially avoiding cascading failures or economic losses.

The multi-modal anomaly response depicted in Figure 5(b) highlights the synergistic effect of data fusion. Network traffic analysis exhibits the earliest response at -3.8 minutes, detecting unusual communication patterns associated with the attack preparation phase. SCADA and PMU modalities subsequently confirm the threat at -3.5 and -3.2 minutes respectively,

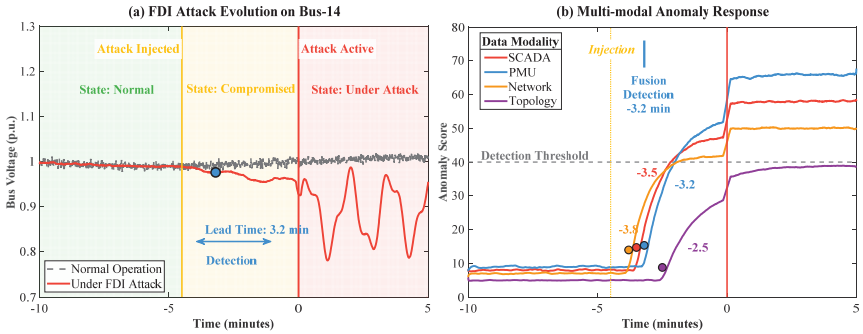

Figure 5 Real-time Attack Detection Case Study. (a) FDI attack evolution on Bus-14 showing injection, detection, and activation timeline. (b) Multi-modal anomaly response demonstrating sequential detection across different data modalities.

Table 5 Attack scenario detection performance summary

Attack Type	Detection Accuracy	Average Lead Time	Primary Detection Modality	Secondary Validation	False Positive Rate
DoS Attack	91.6%	3.5 min	Network (85%)	SCADA (75%)	1.2%
FDI Attack	90.8%	3.2 min	PMU (85%)	Topology (65%)	0.8%
Malware	90.5%	2.8 min	SCADA (70%)	Network (40%)	1.5%
APT Intrusion	94.9%	4.1 min	Network (80%)	Topology (70%)	0.6%
Normal Operation	98.5%	N/A	Balanced	N/A	N/A

while topology analysis provides additional validation at -2.5 minutes. This sequential activation pattern demonstrates how different data sources complement each other, with the fusion mechanism achieving reliable detection at -3.2 minutes when confidence levels across modalities converge.

Table 5 summarizes the detection performance across various attack scenarios, demonstrating the framework's robust capabilities in real-world applications. The analysis reveals that the MM-LLM maintains consistent performance across diverse attack vectors, with particularly strong results for sophisticated attacks that typically evade single-modal detection systems.

Figure 6 further validates the detection performance through confusion matrix analysis and modality importance evaluation. The classification results demonstrate high diagonal dominance with minimal false positives, indicating reliable discrimination between attack types. The modality importance analysis reveals that optimal detection performance requires adaptive weighting of data sources based on attack characteristics. DoS attacks heavily rely

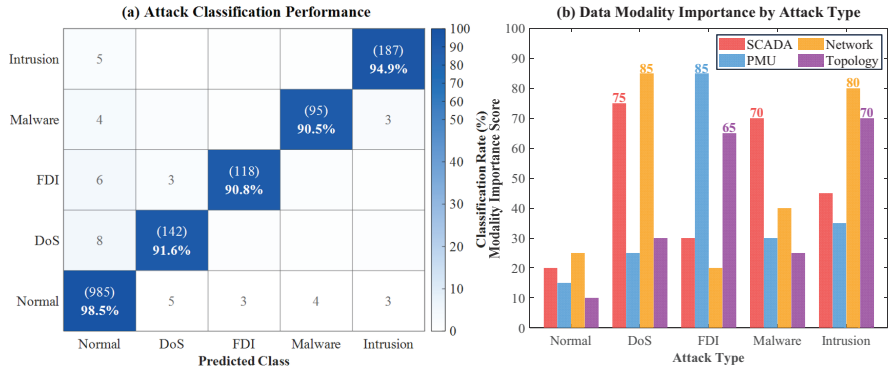

Figure 6 Detection Performance Analysis. (a) Attack classification confusion matrix showing high accuracy across all attack types. (b) Data modality importance scores revealing attack-specific detection patterns.

on network traffic anomalies (85%), while FDI attacks are most effectively detected through PMU measurements (85%). This attack-specific modality preference underscores the importance of the proposed multi-modal fusion approach, as single-source detection would miss critical indicators present in complementary data streams.

5 Discussion

Traditional machine learning methods (SVM-RBF: 77.1%, Random Forest: 79.6%) demonstrate limited accuracy, unable to capture complex temporal patterns while restricted to single modalities. Deep learning approaches (LSTM: 82.9%, CNN: 81.7%) improve feature extraction but remain single-modal. Recent multi-modal efforts include GRU-autoencoder (84.3%) using simple concatenation and MAD-GAN (86.1%) with dual modalities, but lack sophisticated fusion mechanisms and scalability. Single-modal transformers achieve 88.5% through self-attention yet miss cross-modal correlations. The proposed MM-LLM addresses these gaps through domain-optimized encoders and cross-modal attention across four data sources, achieving 95.4% accuracy. The 9.3% improvement over best existing methods (88.5%) translates to reduced false negatives critical for infrastructure protection. While requiring greater computational resources, the framework provides essential early warning capabilities absent in reactive systems.

Compared to existing approaches, the MM-LLM framework offers distinct advantages across multiple dimensions. The 2.7% false positive rate

reduces operational costs by minimizing unnecessary alarm responses, a 65% improvement over traditional methods averaging 7.8% false positives. Real-time processing at 758 samples per second with 13.2ms latency meets stringent industrial requirements without specialized hardware acceleration. The modular encoder architecture enables straightforward integration of additional data modalities as instrumentation evolves, avoiding complete system redesign. Early warning capability provides actionable intelligence enabling proactive defense allocation before attack completion.

The remarkably good performance of our proposed multi-modal large language model framework can be ascribed to its advanced cross-modal attention mechanism which successfully mitigates the cyber security threats of the real-world the DER-included smart grid discovered in [1]. This architectural novelty also follows recent advancements in the LLM applications in the realm of cybersecurity [7], showing that language models can encode complex relationships across disparate data sources. The proposed framework 95.4%, is significantly better than that of MAD-GAN [32], which is 86.1%, using generative adversarial networks for time-series anomaly detection. This gain comes from the potential of the model in exploiting cross-modal relationships, which overcomes the weakness of single-modal methods that attend only to the same kind of source.

The ability to detect events in real time with a latency of 13.2 ms constitutes an important feature to defend the smart grid infrastructure from the advanced techniques documented in [4, 16]. This performance is well within the specifications for IoT connected smart grids [5], where the time to respond in the millisecond range is all that stands between defending and giving up control of the system. The framework's offer of 3.2–4.5 minutes forewarning exceeds existing methods of fault detection [30], which often offer less than one minute of advanced warning. This advanced prediction capability is made possible due to the rich joint analysis of SCADA logs, PMU measurements, network traffic and topological data, which can detect more subtle precursor anomalies.

Compared to domain-specific models like SecureBERT [11], which focuses primarily on textual cybersecurity data, the proposed multi-modal approach demonstrates superior versatility across diverse attack vectors. While privacy-preserving BERT-based models [8] have shown promise for lightweight IoT implementations, they lack the comprehensive threat visibility achieved through multi-modal fusion. The framework's robustness against sophisticated attacks addresses critical gaps identified in recent LLM security surveys [10], particularly regarding the detection of coordinated

multi-vector threats that exploit vulnerabilities across different system layers.

The integration of transformer-based anomaly detection represents a significant advancement over traditional methods surveyed by Shakiba et al. [31]. Recent research on unsupervised transformer-based anomaly detection in ECG signals [33] demonstrates the potential of attention mechanisms for identifying subtle deviations in time-series data. However, the proposed framework extends this capability to multiple concurrent data streams, achieving more robust detection through cross-modal validation. This multimodal approach provides inherent resilience against adversarial attacks, addressing vulnerabilities identified by Nazari et al. [34] in their study of bit-flip attacks against transformer models.

Despite these advances, several limitations warrant consideration. The framework's computational requirements, while optimized through parallel processing, may exceed the constraints of edge computing environments. Recent work on sustainable deep learning at grid edge [35] highlights the challenge of deploying complex models with limited resources. The system's dependency on high-quality, synchronized data from multiple sources poses operational challenges, particularly in legacy infrastructure lacking modern sensors. Additionally, while the framework excels at detecting known attack patterns, its performance against zero-day exploits remains uncertain.

Future research should explore federated learning approaches to enable distributed threat intelligence while preserving data privacy across utilities. Integration with intelligent DC fault protection schemes [36] could enhance the framework's response capabilities beyond detection. The growing sophistication of IoT network attacks [37] necessitates continuous model updates and adaptation. Developing explainable AI components would address the "black box" nature of deep learning models, providing operators with interpretable detection rationales. Furthermore, investigating model compression techniques could enable deployment on resource-constrained edge devices while maintaining detection accuracy, ultimately enhancing the resilience of critical energy infrastructure against evolving cyber threats.

Future directions include incorporating additional modalities such as weather and market data for enhanced contextual awareness. Federated learning approaches would enable multi-utility threat intelligence while preserving privacy and regulatory compliance. Explainable AI techniques could render decision processes interpretable, facilitating operational acceptance where personnel require transparent reasoning before defensive actions. Field

validation in operational environments would verify performance under real-world conditions, including equipment failures and evolving attack methodologies. Investigating adversarial robustness against adaptive adversaries attempting evasion represents critical work, as sophisticated attackers will probe deployed defenses once aware of detection mechanisms.

6 Conclusion

This research presents a multi-modal large language model framework addressing critical limitations in single-modal power system threat detection. The MM-LLM achieves 95.4% accuracy through cross-modal attention integrating heterogeneous data sources, substantially outperforming traditional and deep learning methods while maintaining 94.2% precision, 96.3% recall, and 2.7% false positive rate. Real-time processing with 13.2ms latency and 3.2–4.5 minutes early warning enable proactive defense. Ablation studies confirm that cross-modal attention and multi-modal fusion contribute 6.7% and 7% accuracy improvements, respectively. This work demonstrates large language models' effectiveness for multi-modal threat detection, establishing foundations for next-generation security systems protecting critical infrastructure against sophisticated cyber adversaries.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. Liu et al., "Enhancing cyber-resiliency of der-based smart grid: A survey," *IEEE Transactions on Smart Grid*, vol. 15, no. 5, pp. 4998–5030, 2024.

[2] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211, 2024.

[3] Rafrastara, F.A., Shidik, G.F., Ghozi, W., Rijati, N. and Setiono, O. 2025. Tree-based Ensemble Algorithms and Feature Selection Method for Intelligent Distributed Denial of Service Attack Detection. Journal of Cyber Security and Mobility. vol. 14, no, 01, pp. 1–24, 2025.

[4] I. Zografopoulos, N. D. Hatziargyriou, and C. Konstantinou, "Distributed energy resources cybersecurity outlook: Vulnerabilities, attacks,

impacts, and mitigations," *IEEE Systems Journal*, vol. 17, no. 4, pp. 6695–6709, 2023.

[5] A. Akkad, G. Wills, and A. Rezazadeh, "An information security model for an IoT-enabled Smart Grid in the Saudi energy sector," *Computers and Electrical Engineering*, vol. 105, p. 108491, 2023.

[6] N. S. Shibu, A. R. Devidas, S. Balamurugan, S. Ponnekanti, and M. V. Ramesh, "Optimizing microgrid resilience: integrating IoT, blockchain, and smart contracts for power outage management," *IEEE Access*, vol. 12, pp. 18782–18803, 2024.

[7] J. Zhang et al., "When LLMS meet cybersecurity: A systematic literature review," *Cybersecurity*, vol. 8, no. 1, p. 55, 2025.

[8] M. A. Ferrag et al., "Revolutionizing cyber threat detection with large language models: A privacy-preserving Bert-based lightweight model for IoT/IIoT devices," *IEEE Access*, vol. 12, pp. 23733–23750, 2024.

[9] R. Kaur, T. Klobučar, and D. Gabrijelčič, "Harnessing the power of language models in cybersecurity: A comprehensive review," *International Journal of Information Management Data Insights*, vol. 5, no. 1, p. 100315, 2025.

[10] F. Wu, N. Zhang, S. Jha, P. McDaniel, and C. Xiao, "A new era in LLM security: Exploring security concerns in real-world LLM-based systems," *arXiv preprint arXiv:2402.18649,* 2024.

[11] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "Securebert: A domain-specific language model for cybersecurity," *International Conference on Security and Privacy in Communication Systems*, 2022: Springer, pp. 39–56.

[12] Y. Li et al., "Personal llm agents: Insights and survey about the capability, efficiency, and security," *arXiv preprint arXiv:2401.05459*, 2024.

[13] Y. Li and J. Yan, "Cybersecurity of smart inverters in the smart grid: A survey," *IEEE Transactions on Power Electronics*, vol. 38, no. 2, pp. 2364–2383, 2022.

[14] S. Kumari, C. Prabha, A. Karim, M. M. Hassan, and S. Azam, "A comprehensive investigation of anomaly detection methods in deep learning and machine learning: 2019–2023," *IET Information Security*, vol. 2024, no. 1, p. 8821891, 2024.

[15] F. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, "IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic," *Digital Communications and Networks*, vol. 10, no. 1, pp. 190–204, 2024.

[16] R. Devendiran and A. V. Turukmane, "Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy," *Expert Systems with Applications*, vol. 245, p. 123027, 2024.
[17] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *Journal of Big Data*, vol. 11, no. 1, p. 25, 2024.
[18] S. Islam et al., "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Systems with Applications*, vol. 241, p. 122666, 2024.
[19] J. Duan, "Deep learning anomaly detection in AI-powered intelligent power distribution systems," *Frontiers in Energy Research*, vol. 12, p. 1364456, 2024.
[20] S. Yi, S. Zheng, S. Yang, G. Zhou, and J. Cai, "Anomaly detection for asynchronous multivariate time series of nuclear power plants using a temporal-spatial transformer," *Sensors*, vol. 24, no. 9, p. 2845, 2024.
[21] X. Zhang, W. Sun, K. Chen, and R. Jiang, "A multimodal expert system for the intelligent monitoring and maintenance of transformers enhanced by multimodal language large model fine-tuning and digital twins," *IET Collaborative Intelligent Manufacturing*, vol. 6, no. 4, p. e70007, 2024.
[22] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.
[23] Z. Li et al., "A transformer-based deep learning algorithm to auto-record undocumented clinical one-lung ventilation events," *International Workshop on Health Intelligence*, 2023: Springer, pp. 255–272.
[24] Z. Liu and L. Wang, "A robust strategy for leveraging soft open points to mitigate load altering attacks," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 1555–1569, 2021.
[25] N. Javaid, M. Akbar, A. Aldegheishem, N. Alrajeh, and E. A. Mohammed, "Employing a machine learning boosting classifiers-based stacking ensemble model for detecting non-technical losses in smart grids," *IEEE Access*, vol. 10, pp. 121886–121899, 2022.
[26] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," *International Conference on Critical Information Infrastructures Security*, 2016: Springer, pp. 88–99.
[27] R. C. B. Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari, and S. Pan, "Machine learning for power system disturbance and

cyber-attack discrimination," in *2014 7th International Symposium on Resilient Control Systems (ISRCS)*, 2014: IEEE, pp. 1–8.

[28] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & Chemical Engineering*, vol. 17, no. 3, pp. 245–255, 1993.

[29] C. M. Ahmed, J. Zhou, and A. P. Mathur, "Noise matters: Using sensor and process noise fingerprint to detect stealthy cyber attacks and authenticate sensors in CPS," in *Proceedings of the 34th Annual Computer Security Applications Conference*, 2018, pp. 566–581.

[30] K. Chen, C. Huang, and J. He, "Fault detection, classification and location for transmission lines and distribution systems: a review on the methods," *High Voltage*, vol. 1, no. 1, pp. 25–33, 2016.

[31] F. M. Shakiba, S. M. Azizi, M. Zhou, and A. Abusorrah, "Application of machine learning methods in fault detection and classification of power transmission lines: a survey," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 5799–5836, 2023.

[32] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks," *International Conference on Artificial Neural Networks*, 2019: Springer, pp. 703–716.

[33] A. Alamr and A. Artoli, "Unsupervised transformer-based anomaly detection in ECG signals," *Algorithms*, vol. 16, no. 3, p. 152, 2023.

[34] N. Nazari et al., "Forget and rewire: Enhancing the resilience of transformer-based models against {Bit-Flip} attacks," *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 1349–1366.

[35] T. Sirojan, S. Lu, B. T. Phung, D. Zhang, and E. Ambikairajah, "Sustainable deep learning at grid edge for real-time high impedance fault detection," *IEEE Transactions on Sustainable Computing*, vol. 7, no. 2, pp. 346–357, 2018.

[36] M. Z. Yousaf, S. Khalid, M. F. Tahir, A. Tzes, and A. Raza, "A novel DC fault protection scheme based on intelligent network for meshed DC grids," *International Journal of Electrical Power & Energy Systems*, vol. 154, p. 109423, 2023.

[37] S. A. Bakhsh, M. A. Khan, F. Ahmed, M. S. Alshehri, H. Ali, and J. Ahmad, "Enhancing IoT network security through deep learning-powered Intrusion Detection System," *Internet of Things*, vol. 24, p. 100936, 2023.

Biographies

Li Xiaomeng graduated from Beijing University of Posts and Telecommunications in March 2012 with a Master's degree in Computer Technology. He currently serves as the Leader of the Information Security Monitoring Group, Dispatching and Operation Department at the State Grid Information & Telecommunication Center (Big Data Center), with research focus primarily on cybersecurity.

Lin Bingjie graduated from Peking University in June 2019 with a Master's degree in Electronic and Communication Engineering. She currently serves as a Specialist in the Network and Data Security Department of the State Grid Information & Telecommunication Center (Big Data Center), with research focuses on network and data security, confidentiality monitoring, etc.

Li Huimin graduated from North China Electric Power University in June 2021 with a Master's degree in Electric Power Systems and Automation. She currently works as a Cybersecurity Analyst at the State Grid Information & Telecommunication Center (Big Data Center), her research focusing on cybersecurity.