MDI-BV: A Blockchain-Verified Multimodal Deep Learning-Based Intrusion Detection System for Industrial Internet of Things

Gang Han[1], Yaotian Wang[1*], Xuan Chang[1], Zhongliang Zhang[2*]

[1]The School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an 710061, China
[2]School of Management, Hangzhou Dianzi University, Hangzhou 310018, China.

Keywords: IoT; Intrusion detection; Multimodal deep learning; Blockchain; Data integrity

Citation: Han G., Wang Y.T., Chang X., et al.: MDI-BV: A Blockchain-Verified Multimodal Deep Learning-Based Intrusion Detection System for Industrial Internet of Things. Data Intelligence, 2026, Vol. 8, Art. No.: 20260083. DOI: https://doi.org/10.3724/2096-7004.di.2026.0083

ABSTRACT

As the Industrial Internet of Things (IIoT) is deeply applied in smart manufacturing and remote control, its open architecture and heterogeneous device connections intensify cybersecurity threats such as data injection and DoS. Traditional IDS relying on static rules and single-feature analysis fail to handle complex attacks. This paper proposes MDI-BV, a blockchain-verified multimodal deep learning intrusion detection system. Based on the CIC-IDS-2017 dataset, it extracts four features, utilizes a multi-branch neural network with concatenation fusion and regularization to enhance robustness and generalization. A Hyperledger Fabric-based consortium blockchain ensures end-to-end verifiability and tamper-proof traceability of training/inference. Experiments show it reaches ≥ 99.98% in accuracy, precision, recall and F1-score, with stable performance, strong data integrity and IIoT practicality.

1. INTRODUCTION

With the rapid integration of next-generation information technologies, the Industrial Internet of Things (IIoT) has become a key enabler of smart manufacturing, remote control, and autonomous perception. However, the open architecture, heterogeneous device connections, and complex protocol stacks expose IIoT systems to escalating cybersecurity threats, including data injection, denial-of-service (DoS), identity spoofing, and remote control attacks. These threats severely compromise system stability and reliability.

*Corresponding authors: Yaotian Wang (E-mail: 1767178707@qq.com; ORCID: 0009-0005-5248-1242); Zhongliang Zhang (E-mail: zlzhang2525@163.com; ORCID: 0000-0001-6555-7908).

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Traditional intrusion detection systems (IDS), which rely on static rules and single-feature traffic analysis, struggle to detect complex and evolving attacks in large-scale communication environments.

Recent advances in deep learning have shown promise for intrusion detection, especially in feature extraction and automated classification. Yet, most existing approaches rely on unimodal data, limiting their ability to capture the temporal, structural, semantic, and statistical correlations inherent in sophisticated attack behaviors. Additionally, the trustworthiness and verifiability of model inputs and outputs are undermined by the reliance on static, offline datasets.

To address these challenges, this paper proposes a novel intrusion detection framework that integrates multimodal deep learning with blockchain-based verification for IIoT environments. Based on the CIC-IDS-2017 dataset, the system constructs four types of input features—communication flows, statistical patterns, behavioral semantics, and cross-modal interactions—combined via attention-based fusion. Dropout regularization and lightweight architecture optimization are employed to enhance robustness and deployability in high-dimensional settings. Crucially, a Hyperledger Fabric-based consortium blockchain is introduced to record and verify the entire process, including data preprocessing, feature generation, and inference results, ensuring end-to-end trust. Extensive experiments on the CIC-IDS-2017 dataset evaluate the impact of different modal combinations, fusion strategies, and regularization parameters, demonstrating the proposed method's superiority in accuracy, stability, and data integrity.

2. RELATED WORK

In recent years, the increasing complexity and stealth of cyberattacks have challenged the effectiveness of traditional intrusion detection systems (IDS), particularly in the Industrial Internet of Things (IIoT), where heterogeneous data sources and complex protocols limit the generalization and robustness of shallow, unimodal approaches. Deep learning has emerged as a dominant approach in intrusion detection, with growing interest in multimodal fusion, sequential modeling, and attention mechanisms.

Introduced for text-to-video retrieval on untrimmed videos [1], PRVR and its MS-SL(++) formulation—casting partial relevance as multi-instance learning with joint clip- and frame-level matching—provide transferable mechanisms for detecting sparse, query-conditioned malicious moments within long multimodal network sessions.

Although developed for fine-grained micro-action recognition [2], the MA-52 dataset and MANet—via temporal shift modules, squeeze-and-excitation recalibration, and a joint embedding loss to separate visually similar classes—offer transferable mechanisms for modeling weak, short-lived, easily confused multimodal patterns in network intrusion detection.

Deep learning enables end-to-end feature extraction and classification without manual engineering, effectively capturing intricate attack patterns. Early works such as Arun Kumar Silivery et al. (2023) applied Recurrent Neural Network (RNN), Long Short-Term Memory Memory Recurrent Neural Network (LSTM-RNN), and Deep Neural Network (DNN) to the NSL-KDD dataset for sequential attack modeling, showing

improvements over traditional methods in recall and F1-score [3]. Yihui Xiong et al. (2022) used stacked denoising autoencoders (SDAE) to learn nonlinear representations, improving detection accuracy with downstream classifiers [4]. However, these models typically rely on single-modality inputs, limiting their ability to capture multi-dimensional attack characteristics.

To address this, recent studies have explored multimodal deep learning to enhance detection expressiveness by combining features from traffic flows, system calls, behavioral context, and network topology. For instance, Murtaza Ahmed Siddiqi et al. (2022) fused image-based and numerical features using CNNs on the CIC-IDS-2017 dataset, achieving F1-scores above 99% [5].

Attention mechanisms and graph neural networks (GNNs) have also been introduced into IDS research. Zecheng Liu et al. (2023) applied self-attention to enhance multimodal fusion while preserving modality independence [6]. Zhi Quan Qin et al. (2022) modeled IP communication as a graph and employed GNNs for relational aggregation, significantly improving lateral attack detection on the ISCX-2012 dataset [7]. These works underscore the potential of deep neural networks for behavior modeling and the critical role of optimized fusion in enhancing detection accuracy and stability.

Despite these advances, challenges remain. Modality heterogeneity and redundancy can degrade performance if fusion strategies are poorly designed [8]. Model robustness suffers under missing or incomplete modalities, and complex architectures often incur high training costs and limited deployability. Techniques such as regularization, dropout, and attention modules have been proposed to mitigate these issues, but balancing performance and efficiency remains an open problem.

Meanwhile, blockchain technologies have been explored to enhance the trustworthiness of training data and results. For example, Houshyar Honar Pajooh et al. (2021) embedded Hyperledger Fabric in securing the edge internet of things [9]. However, most efforts treat blockchain as an external layer rather than integrating it into end-to-end deep learning frameworks.

In summary, IDS research is shifting from shallow rule-based systems to multimodal deep learning architectures. Key challenges include designing effective fusion mechanisms, improving modality coordination, and ensuring robustness during both training and inference. To address these, this paper proposes a multimodal deep learning-based IDS with integrated blockchain verification for IIoT, aiming to deliver high accuracy, strong robustness, and traceable trust for industrial cybersecurity applications.

3. SYSTEM ARCHITECTURE

This paper proposes a multimodal deep learning intrusion detection system [10] integrated with blockchainbased trust mechanisms to address security challenges posed by heterogeneous data in the Industrial Internet of Things (IIoT). The system constructs feature modalities from four semantic dimensions: communication behavior, temporal distribution, statistical attributes, and variable interactions. High-level representations are extracted using convolutional networks, recurrent neural networks, Transformers, custom attention modules, and embedding fusion networks. These are then fused via multi-head attention

or feature concatenation and passed to a deep classifier for intrusion detection. A lightweight blockchain mechanism is introduced to hash and chain critical features and prediction results, creating an immutable audit log audit log that seamlessly integrates detection performance with data integrity assurance. Figure 1 shows the system architecture.

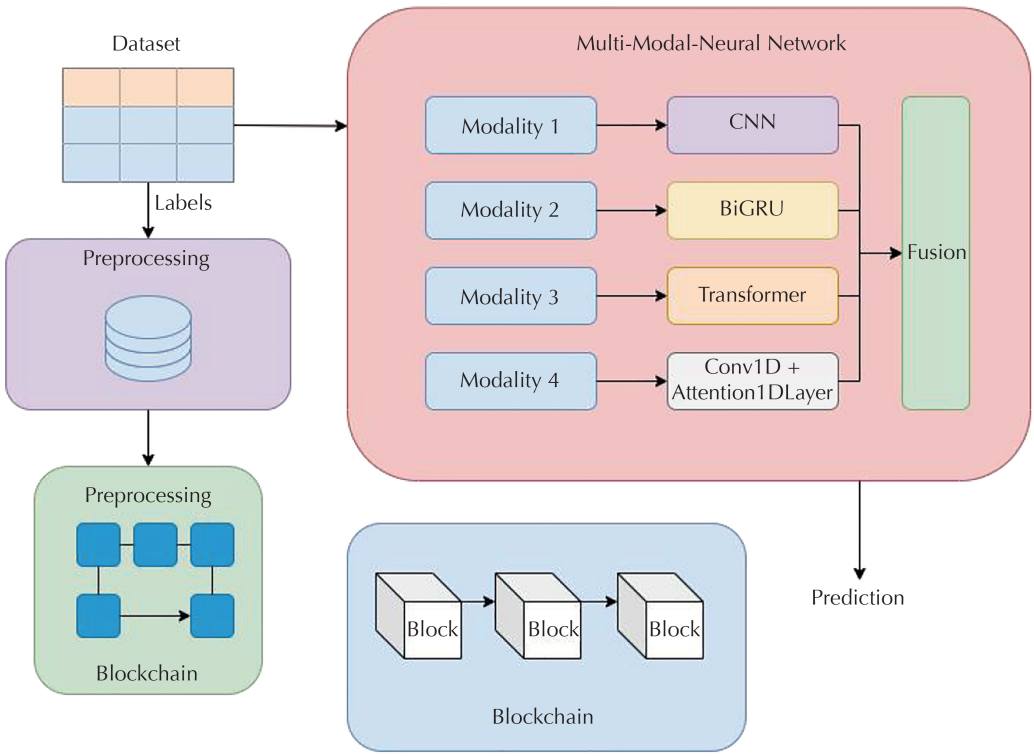


Figure 1. System Architecture.

3.1 Multimodal Feature Representation

To address the highly heterogeneous, multi-granular, and multi-timescale nature of IIoT network data, this paper proposes a structured multimodal feature representation method. Raw traffic data is decomposed into four semantically distinct modalities to better capture the discriminative patterns between normal and malicious behaviors, thereby enhancing feature interpretability and enabling more targeted inputs for downstream deep subnetworks.

Modality 1: Basic Communication Features

This modality includes fundamental statistics related to traffic behavior, such as the number of forward and backward packets/bytes, packet length statistics, flow duration, protocol type (e.g., one-hot encoded TCP/UDP), and flag counts. These features reflect the intensity and directionality of connections and form the foundation of traditional intrusion detection systems.

Modality 2: Temporal Behavior Features

Focusing on the timing characteristics of communication, this modality includes inter-arrival times (IAT) between flows, statistical measures of forward/backward IAT (max, min, mean, standard deviation), and cyclical time features (e.g., hour, weekday) encoded using sine and cosine functions to preserve periodicity. This helps capture temporal dependencies and recurrent patterns in network behavior.

Modality 3: Statistical and Anomaly Features

This modality applies log and Box-Cox transformations [11] to enhance distribution symmetry and stability. It includes standardized scores (Z-score) [12], binary outlier indicators (IsOutlier), and global anomaly scores to improve sensitivity to hidden attacks and unusual patterns, reinforcing the model's ability to detect anomalies.

Modality 4: Cross-Constructed Features

High-order interaction features are constructed through operations such as multiplication, division, subtraction, and squaring across variables—for example, "forward packets □ backward packets" or "flow duration ÷ total bytes." These features reveal complex nonlinear dependencies and interaction patterns among communication variables.

To effectively support the abstraction and high-level representation of the four feature modalities, the system implements a systematic, multi-stage feature engineering pipeline. This pipeline takes raw IIoT communication data as input and performs cleaning, transformation, analysis, and reconstruction to achieve semantic unification and structural standardization across heterogeneous sources. The result is a clear and high-quality input foundation for the downstream multimodal deep learning subnetworks. The detailed process is as follows:

(1) **Raw Data Cleaning and Preprocessing**

Column names, missing values, and infinite values in the raw dataset are first standardized. Numerical fields are repaired using median imputation or maximum finite value substitution, while categorical variables are filled with mode values to ensure consistency during encoding. Duplicate rows are also detected and removed to maintain input integrity.

(2) **Temporal Feature Construction and Periodicity Modeling**

Timestamp fields are parsed to extract features such as hour, day of the week, weekend indicator, and business hours flag. To capture periodic and trending behaviors, cyclical transformations using sine and cosine functions are applied. A sliding time window is used to calculate the frequency of flows per source/destination IP, enhancing sensitivity to short-term traffic fluctuations.

(3) **Network Topology and Graph Structure Analysis** [13]

A directed graph (Source IP → Destination IP) is constructed to analyze network structure. Degree and betweenness centrality are computed to identify nodes with abnormal influence. Private and public IPs are distinguished, and entropy metrics are used to represent behavioral uncertainty.

(4) **Traffic Statistics and Protocol Behavior Modeling**

Key traffic indicators such as total packets, total bytes, flow direction, and burstiness are extracted. Derived features like forward/backward ratios, average packet size, and byte transfer rate are computed. For TCPspecific analysis, the system calculates flag sums and unique flag counts to detect protocol-layer anomalies.

(5) **Statistical Transformation and Distribution Enhancement**

To address skewed distributions (e.g., packet counts, byte volume), log transformation and Box-Cox normalization are applied. Additionally, numerical features are discretized via equal-frequency binning (KBinsDiscretizer) to support clustering and categorical modeling.

Based on the discretized representations, we further derive clustering-related structural features (referred to as Clustering Features), including cluster IDs, distances to cluster centroids, and cluster-level statistics (e.g., intra-cluster density or frequency) computed on selected flow variables. These features capture coarse-grained distribution structure and are treated as a subcomponent of Modality 3 (Distribution Transformation & Anomaly Features) for ablation and analysis.

(6) **Multi-Scale Anomaly Detection**

An Isolation Forest-based global anomaly detector assigns anomaly scores to training samples. Z-scores are computed for key features, and outlier labels are added. This dual-level tagging enhances the model's early detection of hidden attacks.

(7) **High-Dimensional Time Series Feature Extraction**

For each source IP, the system uses the tsfresh package to extract rolling window-based time series features, including max, mean, amplitude, and frequency-domain metrics. This enables automatic modeling of traffic trends and periodic attacks.

(8) **Cross Feature Construction and Nonlinear Interaction Modeling**

Features such as "forward packets □ backward packets" and "flow duration ÷ total bytes" are created to capture nonlinear relationships between variables. These cross features reveal complex interaction patterns indicative of sophisticated attack behaviors.

(9) **Feature Importance Evaluation and Selection**

Mutual information is used to evaluate feature relevance to the target label. The top 20 most informative features are selected for model optimization and interpretability analysis.

3.2 Multimodal Subnetwork Architecture

For the four feature modalities constructed in the previous section, the system designs structurally diverse and task-adaptive deep learning subnetworks to enable effective abstraction and modeling within each modality's respective feature space. Each subnetwork is tailored to its input feature type, leveraging the strengths of deep neural architectures on different data distributions to ensure strong feature extraction and generalization performance.

Modality 1 Subnetwork (CNN Encoder)

This modality comprises static communication features, suitable for local pattern extraction using a Convolutional Neural Network (CNN) [14]. Input vectors are reshaped and passed through multiple Conv1D layers to capture statistical patterns at various scales. The architecture includes two convolutional layers (with 64 and 128 filters), followed by Batch Normalization and MaxPooling1D. A GlobalAveragePooling1D layer is then applied to produce a global feature representation for downstream fusion.

Modality 2 Subnetwork (Bi-GRU Temporal Network)

To model time intervals and periodic features, a Bidirectional Gated Recurrent Unit (Bi-GRU) [15] network is employed to capture forward and backward temporal dependencies. The subnetwork consists of two stacked Bidirectional (GRU) layers with 64 units each. Dropout (set to 0.3) is applied after the hidden layers to enhance robustness and prevent overfitting. This structure effectively captures long-term temporal patterns, such as inter-arrival times (IAT).

Modality 3 Subnetwork (Transformer Encoder)

Statistical and anomaly features often involve nonlinear, long-range dependencies and complex interactions. A Transformer encoder [16] is used to model these relationships via multi-head self-attention. The architecture includes a multi-head attention layer (4 heads, each 64 dimensions), a feedforward layer (128 hidden units), residual connections, and layer normalization. The output is aggregated using GlobalAveragePooling1D to form a fixed-size representation vector.

Modality 4 Subnetwork (Attention-Conv Network)

This modality contains complex nonlinear relationships derived from feature combinations. An Attention-Conv network [17] is designed to enhance salient interactions by combining Conv1D with a custom attention mechanism. Local interactions are first captured through convolution, followed by an Attention1DLayer that assigns learnable weights to each timestep for dynamic feature emphasis. A Dropout layer is applied to mitigate overfitting before producing the final feature representation.

These subnetworks are tailored to the semantic characteristics of each modality, enabling comprehensive modeling of complex intrusion patterns.

3.3 Feature Fusion Strategy

After independently abstracting each modality through its dedicated subnetwork, the system fuses the resulting feature vectors from multiple subspaces to enable joint modeling of global behavioral patterns. Given the significant differences in semantic types, statistical distributions, and information density across modalities, the fusion strategy plays a critical role in determining the system's discrimination capability and generalization performance. To this end, an attention-based fusion strategy [18] using multi-head attention is employed.

3.4 Blockchain-Based Trust Verification Mechanism

In multimodal deep learning-based intrusion detection systems, ensuring data integrity, verifiability of model predictions, and non-repudiation of results is critical for deployment in Industrial Internet of Things (IIoT) and other mission-critical infrastructures. To meet these security requirements, this paper introduces a lightweight blockchain-based trust verification mechanism [19]. By hashing training features and prediction results in batches, constructing blocks, and linking them into a chain, the mechanism enables end-to-end traceability and tamper-proof assurance for the entire model execution process.

This mechanism functions not only as a logging or audit tool but also as an integral component of the system's trust framework, supporting the foundation of 'trusted artificial intelligence (AI)'.

Key security benefits include:

- Non-repudiation: Any tampering with model outputs can be verified using on-chain prediction hashes.
- Integrity Verification: All training data are batch-recorded on the chain; any modification results in hash mismatches.
- Auditability: Every training and inference process can be fully traced and verified via the blockchain structure.
- Trustless Environment Support: Enables independent verification of data and predictions in collaborative or edge computing environments.

4. METHODOLOGY

In response to escalating cybersecurity threats, this study proposes an intrusion detection system that integrates multimodal deep neural networks with blockchain-based trust mechanisms, aiming to achieve high accuracy, strong robustness, and interpretability in anomaly detection. The system is built around four core modules: multimodal feature modeling, subnetwork design for structural matching, fusion strategy optimization, blockchain-based verification, and visual evaluation. Four semantic feature types—statistical, temporal, anomaly, and interaction—are extracted and modeled using dedicated networks, including CNN, BiGRU, Transformer, MLP, and attention-based convolution. Modal representations are fused via multi-head attention or concatenation to enhance expressiveness and prediction performance. To ensure result trustworthiness and traceability, a lightweight blockchain mechanism is introduced to hash and chain the full training and inference process. Experiments using diverse metrics and visual analysis demonstrate the method's superiority in accuracy, stability, and deployment feasibility, highlighting its promising potential for IIoT security applications.

The following lists various labels (which typically represent different categories of network traffic or cyber attack types) along with their corresponding encoded values: the label "BENIGN" has an encoded value of 0, "Bot" corresponds to 1, "DDoS" to 2, "DoS GoldenEye" to 3, "DoS Hulk" to 4, "DoS Slowhttptest" to 5, "DoS slowloris" to 6, "FTP-Patator" to 7, "Heartbleed" to 8, "Infiltration" to 9, "PortScan" to 10, "SSH-Patator" to 11, "Web Attack Brute Force" to 12, "Web Attack Sql Injection" to 13, and "Web Attack XSS" is assigned an encoded value of 14. Table 1 more intuitively illustrates the mapping relationship. The mapping relationship between labels and encoded values is shown in Table 1.

Table 1. Label Mapping.

Label	Encoded_Value
BENIGN	0
Bot	1
DDoS	2
DoS GoldenEye	3
DoS Hulk	4
DoS Slowhttptest	5
DoS slowloris	6
FTP-Patator	7
Heartbleed	8
Infiltration	9
PortScan	10
SSH-Patator	11
Web Attack Brute Force	12
Web Attack Sql Injection	13
Web Attack XSS	14

4.1 Multimodal Feature Construction

To enhance the intrusion detection model's capability in characterizing and generalizing network communication behaviors, this study designs and constructs four representative feature modalities based on raw traffic log data, which includes information such as IP addresses, ports, protocols, timestamps, packet lengths, and delays. Each modality abstracts network behavior from a distinct perspective, capturing the expressive potential of traffic data across statistical, temporal, structural, distributional, and interaction dimensions. Let the raw traffic sample be:

$$D = \{x_i, y_i\}_{i=1}^{N} \tag{1}$$

$x_i \in \mathbb{R}^{d_0}$ representing raw features, y_i representing the corresponding class label.

(1) **Basic Communication Features**

The statistical modality primarily captures core communication metrics during network flows, including packet counts, byte volume, flow duration, transmission rate, and directionality.

Key features include:

- Total Packets

$$Total_Packets = x^{fwd_packets} + x^{bwd_packets} \tag{2}$$

- Total Bytes

$$Total_Bytes = x^{fwd_length} + x^{bwd_length} \tag{3}$$

- Average Packet Size

$$Avg_Packets_Size = \frac{Total_Bytes}{Total_Packets + 1} \tag{4}$$

- Forward/Backward Ratio

$$Fwd_Bwd_Ratio = \frac{x^{fwd_packets}}{x^{bwd_packets} + 1} \tag{5}$$

- Rate Feature (Bytes per Time Unit)

$$Flow_Speed = \frac{Total_Bytes}{Flow_Duration + 1} \tag{6}$$

In addition, several TCP flag fields such as SYN, ACK, and PSH are counted to characterize connection states.

(2) Temporal and Cyclical Features

The temporal modality models dynamic timing characteristics of network behavior, including periodic timestamp encoding and variations in communication intervals.

The first step is to extract the timestamp:

$$t_i = Timestamp(x_i) \tag{7}$$

Based on the extracted timestamp, periodic temporal features are constructed, such as hour and day of the week:

- Hour (sin/cos encoding, 24-hour cycle):

$$Hour_\sin = \sin\left(2\pi \cdot \frac{t_{hour}}{24}\right), \quad Hour_\cos = \cos\left(2\pi \cdot \frac{t_{hour}}{24}\right) \tag{8}$$

- Day of Week (sin/cos encoding, 7-day cycle):

$$Day_\sin = \sin\left(2\pi \cdot \frac{t_{day}}{7}\right), \quad Day_\cos = \cos\left(2\pi \cdot \frac{t_{hour}}{7}\right) \tag{9}$$

In addition, the system calculates the access frequency of each source and destination IP within a 5-minute sliding window to model bursty connection behavior:

$$SrcIP_Freq(t) = \#\{x_j : SrcIP_j = SrcIP_i \wedge t_j \in [t - 5min,\ t]\} \tag{10}$$

(3) **Statistical and Anomaly Features**

To enhance the model's ability to represent non-normal feature distributions, log transformation and Box-Cox transformation are applied to skewed data:

- Log Transformation:

$$x_{\log} = \log(1 + x) \tag{11}$$

- Box-Cox Transformation ($\lambda \neq 0$):

$$x_{boxcox} = \frac{x^{\lambda} - 1}{\lambda}, \quad with\ x > 0 \tag{12}$$

For anomaly detection, Z-score and Isolation Forest-based anomaly scoring are introduced:

- Z-score:

$$z_i = \frac{x_i - \mu_x}{\sigma_x},\ Outlier_i = I(|z_i| > 3) \tag{13}$$

- Isolation Forest [20] Anomaly Score (on subsets):

$$s_i = IF(x_i), \quad IF(x_i) \in \{-1, 1\} \tag{14}$$

These features effectively capture abnormal shifts and isolated instances in network traffic.

IF represents the output produced by the Isolation Forest.

(4) Cross Interaction Features

To capture nonlinear interactions between features, various cross-feature construction methods are introduced, including products, ratios, sums, differences, and higher-order combinations:

- Feature Products:

$$x^{(i \cdot j)} = x_i \cdot x_j \tag{15}$$

- Ratio Features:

$$x^{(i/j)} = \frac{x_i}{x_j + \epsilon} \tag{16}$$

- Polynomial Interaction Terms:

$$x^{(i^2 \cdot j)} = x_j^2 \cdot x_j, \quad x^{(i \cdot j^2)} = x_i \cdot x_j^2 \tag{17}$$

This modality significantly expands the original feature space, enhancing the model's ability to fit complex high-order patterns.

Through the above multimodal feature extraction process, raw traffic data is transformed into four complementary and information-rich feature representations, providing a solid foundation for deep subnetwork learning and modality fusion.

4.2 Modality-Specific Subnetworks

To align with the distinct informational and structural characteristics of each modality, this study designs modality-specific deep subnetworks for the four feature types. These subnetworks enhance feature representation capacity and improve model robustness. Each subnetwork independently processes its corresponding modality input and maps it to a unified embedding space for subsequent fusion.

(1) Modality 1: Convolutional Neural Network Encoder (CNN for Statistical Features)

The statistical modality consists of 1D numerical vectors with significant local correlations and varying scales, making 1D Convolutional Neural Networks (1D-CNN) [21] well-suited for feature extraction.

Given input $x^{(1)} \in \mathbb{R}^{d_1}$, the vector is first reshaped as:

$$x_{reshaped}^{(1)} = Reshape(x^{(1)}) \in \mathbb{R}^{d_1 \times 1} \tag{18}$$

It then passes through two sequential convolution and pooling layers:

$$h_1 = GlobalAvgPool1D(Conv_{128} \circ ReLU \circ MaxPool \circ Conv_{64}(x_{reshaped}^{(1)})) \tag{19}$$

The final output is a fixed-length representation vector $\mathbf{h}_1 \in \mathbb{R}^{128}$.

(2) Modality 2: Bidirectional Gated Recurrent Unit Network (BiGRU for Temporal Sequences)

The temporal modality is represented as a 2D time series tensor $x^{(2)} \in \mathbb{R}^{T \times d_2}$, characterized by strong sequential dependencies and contextual information. A Bidirectional GRU is employed for modeling:

$$h_2 = BiGRU_{64} \circ BiGRU_{64}(x^{(2)}), \quad h_2 \in \mathbb{R}^{128} \tag{20}$$

Bidirectional GRUs capture both forward and backward dynamics, enhancing the model's ability to represent periodic and bursty behaviors [22].

(3) Modality 3: Transformer Encoder (Transformer for Statistical Abnormalities)

The statistical abnormality modality is a fixed-dimensional vector $x^{(3)} \in \mathbb{R}^{d_3}$, characterized by long-range dependencies and nonlinear interactions. A simplified Transformer encoder is employed to capture its global dependencies.

The input is first expanded to support sequence modeling:

$$x_{seq}^{(3)} = ExpandDims(x^{(3)}) \in \mathbb{R}^{1 \times d_3} \tag{21}$$

In this context, ExpandDims denotes a neural network operation.

Self-attention is then computed as:

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{22}$$

Where:

$$\begin{aligned} Q &= x_{seq}^{(3)} W_Q \\ K &= x_{\text{seq}}^{(3)} W_K \\ V &= x_{seq}^{(3)} W_V \end{aligned} \tag{23}$$

Finally, the output is passed through a feedforward layer and normalized:

$$h_3 = GlobalAvgPool1D(LayerNorm(FFN(Attention(Q,\ K,\ V)))) \tag{24}$$

(4) Modality 4: Attention-Enhanced Convolutional Neural Network (Attention-Enhanced CNN [23])

To enhance the modeling capacity of the interaction modality, an attention-guided convolutional network is designed. First, 1D-CNN is used for local feature extraction:

$$z_5 = Conv1D(x_{reshaped}^{(5)}) \in \mathbb{R}^{l \times c} \tag{25}$$

Then, a custom Attention1D mechanism with learnable weights W is applied:

$$e_i = tanh(w^\top z_5[i] + b), \quad \alpha_i = \frac{exp(e_i)}{\sum_j exp(e_j)}, \quad h_5 = \sum_i \alpha_i z_5[i] \tag{26}$$

This mechanism adaptively focuses on discriminative subspaces of feature interactions, improving sensitivity to attack patterns.

In summary, four structurally distinct subnetworks are constructed, each performing non-shared modeling of its respective modality. This heterogeneous neural encoding architecture effectively captures deep semantic information within each modality and outputs unified-dimensional representations to the fusion layer, establishing a strong foundation for multimodal integrated learning.

4.3 Multimodal Fusion Strategies

After independently modeling different types of features through modality-specific subnetworks, the system integrates the four output representation vectors to obtain a more comprehensive and discriminative final feature representation. Given the differences in information dimensions, feature structures, and semantic spaces across modalities, the choice of fusion strategy is critical to overall model performance.

This study designs and compares two representative fusion mechanisms: Multi-head Attention Fusion and Concatenation-based Weighted Fusion. Where W_Q, W_K,W_V, W_Q, W_K,W_V are the learnable projection matrices for queries, keys, and values, respectively.

(1) Multi-head Attention Fusion [24]

The multi-head attention mechanism enables the model to learn dynamic inter-modal correlations, enhancing the adaptability of cross-modal representations. Under this mechanism, assume each modalityspecific subnetwork outputs a feature vector $h_m \in \mathbb{R}^d$, where $m \in \{1, 2, 3, 4\}$ denotes the output of the m-th modality, forming the modality representation set:

$$H = \{h_1, h_2, \ldots, h_M\}, \quad M = 4 \tag{27}$$

Each modality vector is first projected into a unified attention input space:

$$z_m = W_m h_m + b_m, \quad z_m \in \mathbb{R}^{d_k} \tag{28}$$

All modality vectors are then stacked to form the fusion input matrix:

$$Z = Stack(z_1, z_2, \ldots, z_M) \in \mathbb{R}^{M \times d_k} \tag{29}$$

A self-attention mechanism is applied to model inter-modal relationships:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{30}$$

Where:

$$\begin{aligned} &Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \\ &W_Q, W_K, W_V \in \mathbb{R}^{d_k \times d_k} \end{aligned} \tag{31}$$

The final multi-head attention fusion vector is:

$$h_{fusion} = GlobalAvgPool1D(MultiHeadAttention(Z, Z, Z)) \in \mathbb{R}^{d_k} \tag{32}$$

This mechanism allows the model to dynamically focus on the importance of each modality for a given input, enabling selective and adaptive feature integration.

(2) Classifier Design After Fusion

Regardless of the chosen fusion strategy, the process yields a unified feature representation. This fused vector is passed through a two-layer fully connected classifier for prediction:

$$\begin{aligned} o_1 &= Dropout(ReLU(W_2 h_{fusion} + b_2)) \\ o_2 &= Dropout(ReLU(W_3 o_1 + b_3)) \\ \hat{y} &= \text{Softmax}(W_4 o_2 + b_4) \end{aligned} \tag{33}$$

Here, $\hat{y} \in \mathbb{R}^C$ denotes the predicted probability distribution over C classes. During training, the crossentropy loss function is used:

$$L = -\sum_{i=1}^{C} y_i \log(\tilde{y}_i) \tag{34}$$

Through this fusion strategy, the system achieves effective coordination across modalities and enhanced feature representation, significantly improving classification performance in complex, multi-dimensional traffic scenarios. The experimental section further evaluates the impact of different fusion mechanisms on accuracy, robustness, and interpretability.

4.4 Blockchain-Aided Trust Mechanism

To enhance data trustworthiness and result auditability in intrusion detection systems operating across multi-source environments and heterogeneous nodes, this study integrates a lightweight blockchain

mechanism into the multimodal deep detection framework. By incorporating Merkle hash trees, chained block indexing, and timestamp signatures, the system constructs a decentralized and tamper-proof trust chain that enables verifiable and traceable modeling of both data flow and prediction processes.

(1) Design Objectives of the Blockchain-Aided Trust Mechanism [25]

Traditional intrusion detection systems often lack verifiability for model inputs and outputs, making them vulnerable to data tampering and result forgery—especially in industrial edge deployments or across distributed heterogeneous nodes. The proposed mechanism aims to achieve the following goals:

- Input Verification: Ensure that features received by the model originate from legitimate, untampered data.
- Prediction Traceability: Guarantee that each prediction can be traced back to its original input features.
- Data Consistency: Maintain the integrity and non-repudiation of data used during training and inference.
- Autonomous Node Verification: Enable cross-verification of prediction results among participating nodes.

(2) Block Generation and Chained Structure Indexing

For each batch of data processed (either training or inference), the system generates a block that encapsulates the batch's Merkle root hash, timestamp, and the hash of the previous block:

$$B_t = < BlockID_t,\ H_{root}^t,\ TS_t,\ H_{prev}^{t-1},\ ModelSig_t > \tag{35}$$

Where:

- $BlockID_t$: Current block index
- H_{root}^t: Merkle root hash of the batch
- TS_t: Timestamp (recorded as a Unix timestamp)
- $H_{prev}^{t-1} = SHA256(B_{t-1})$: Hash of the previous block
- $ModelSig_t$: Model signature, including version and parameter digest

The final block hash is computed as [26]:

$$H_{Block}^t = SHA256(B_t) \tag{36}$$

Once the new block is appended, a chained structure is formed:

$$\ldots \rightarrow B_{t-1} \rightarrow B_t \rightarrow B_{t+1} \rightarrow \ldots \tag{37}$$

This structure ensures forward immutability and backward traceability. Any modification to a block invalidates all subsequent hashes, allowing tampering to be detected by nodes on the chain.

4.5 Performance Evaluation and Visualization

To comprehensively evaluate the classification performance, generalization ability, interpretability, and deployment feasibility of the proposed Multimodal Deep Intrusion Detection System with Blockchain Verification (MDI-BV), this study designs a multidimensional evaluation and visualization module. It covers classification metric computation, visual model comparison, anomaly monitoring, trusted record display, and multimodal interpretability analysis. This module supports both model development (e.g., hyperparameter tuning and performance comparison) and post-deployment automated monitoring and security auditing.

(1) Classification Metric Definitions

To assess the model's ability to identify various attack types, standard multi-class evaluation metrics are used, including Accuracy, Precision, Recall, F1-Score, Cohen's Kappa (K), and AUROC (Area Under the Receiver Operating Characteristic Curve).

- True Positives: $TP_C = \#(\tilde{y}_i = y_i = C)$
- False Positives: $FP_C = \#(\tilde{y}_i = C,\ y_i \neq C)$
- False Negatives: $FN_C = \#(\tilde{y}_i \neq C,\ y_i = C)$
- Precision:

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{38}$$

- Recall:

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{39}$$

- F1-Score:

$$F1_c = \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{40}$$

- Accuracy:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(\tilde{y}_i = y_i) \tag{41}$$

- Cohen's Kappa [27]:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_o = Accuracy, \quad p_e = \sum_{c=1}^{C}\left(\frac{n_{c,\cdot} \cdot n_{\cdot,c}}{N^2}\right) \tag{42}$$

- Area Under the ROC and PR Curves:

$$AUROC_c = \int_0^1 TPR(x)dx, \quad AUPRC_c = \int_0^1 Precision(r)dr \tag{43}$$

Where C is the total number of classes, N is the total number of samples, and $I(\cdot)$ denotes the indicator function.

5. EXPERIMENTAL DESIGN AND EVALUATION

To evaluate the effectiveness, robustness, and interpretability of the proposed intrusion detection system (MDI-BV), which integrates multimodal neural networks with a blockchain-based trust mechanism, a systematic experimental framework is established. The evaluation covers training performance, comparisons of fusion strategies, system overhead introduced by the blockchain, and the practical value of model interpretability in anomaly analysis.

Experiments are conducted on the CIC-IDS-2017 dataset, incorporating multimodal information such as node behavior, communication statistics, protocol structure, and temporal patterns to construct highdimensional intrusion detection inputs. After unified encoding, oversampling, and modality construction, the data is used to train various models.

The evaluation framework compares traditional methods and deep learning models, examines the impact of different fusion strategies and blockchain integration on detection performance, and employs comprehensive metrics including accuracy, precision, recall, F1-score, Cohen's kappa, and AUC. Results are visualized using confusion matrices, PR curves, and training curves.

Furthermore, on-chain operation costs—such as hash computation time and block generation latency—are recorded alongside model inference time to assess deployment feasibility. Modal ablation and error-type visualizations are used to analyze the impact of key features on system performance. This experimental design ensures comprehensive evaluation, reproducibility, and statistical significance, providing a solid foundation for the real-world deployment of MDI-BV in industrial environments.

5.1 Experimental Environment and Dataset

A 16-core AMD EPYC 9354 central processing unit (CPU) (3.25 GHz), 60.1 GB RAM, and an NVIDIA RTX 4090 GPU (25.2 GB VRAM) are used to ensure high-throughput training for deep multimodal models. The operating system is Ubuntu 20.04 LTS, with Python 3.9 as the primary environment and TensorFlow 2.1 as the deep learning framework for training multimodal neural networks.

The experiments are conducted on the CIC-IDS-2017 dataset, which combines real network flow logs with node behavior features to construct a multimodal intrusion detection dataset covering eight types of attacks. The raw traffic data includes IP addresses, ports, protocols, duration, packet counts, inter-arrival times, payload lengths, and byte volumes. Node behavior logs are also incorporated; behavioral factors are extracted using keyword matching and language models, then encoded as structured vectors.

The dataset is split into 70% training, 15% validation, and 15% testing. All data undergo Z-score normalization and structural mapping, resulting in four modality subsets: statistical, temporal, protocol, and behavioral modalities—reflecting numerical properties, temporal dependencies, protocol attributes, node behavior, and feature interactions, respectively.

To ensure the trustworthiness of both data and inference processes, SHA-256 is applied to the raw input, encoded features, and prediction outputs for each sample batch. The resulting hash values are organized into a Merkle tree and recorded on the blockchain, enabling data integrity verification and result traceability to enhance system security and trust.

To evaluate the performance of the proposed multimodal neural network in network intrusion detection, ten groups of comparative experiments are designed across multiple dimensions, including modality structure, fusion strategy, subnetwork variations, and regularization control. All experiments are conducted using the same dataset splits and preprocessing pipeline to ensure consistency and comparability. Table 2 shows the comparison of multiple sets of experiments.

Table 2. Experimental Groups.

Experiment ID	Experiment Name	Modality Configuration	Primary Variation
Exp1	FullModalities	Mod1~Mod4	Full Modality Fusion, Attention Mechanism
Exp2	NoModality4	Remove Mod4	Analyze Contribution of Cross Features
Exp3	NoCluster	Remove Clustering Features	Remove Clustering Structure and Embedding Information
Exp4	ConcatFusion	Mod1~Mod4	Replace Attention Fusion with Concatenation Fusion
Exp5	unidirectional GRU	Mod1~Mod4	Replace Bidirectional GRU with Unidirectional GRU
Exp6	SimpleDenseMod3	Mod1~Mod4	Replace Transformer in Modality 3 with a Simple Fully Connected Network
Exp7	NoDropoutConcatFusion	Mod1~Mod4	Remove Dropout, Use Concatenation Fusion
Exp8	NoModality3	Remove Mod3	Analyze the Role of Statistical Transformation Modality
Exp9	Modality4Only	Retain Mod4 Only	Performance Lower Bound with Single Modality Model
Exp10	HighDropout	Mod1~Mod4	Enhance Regularization to Prevent Overfitting

5.2 Model Configuration and Comparative Experiments

5.2.1 Model Architecture and Training Configuration

The proposed model consists of four modality-specific subnetworks, each corresponding to a distinct source and type of feature subset, including basic statistics, time series, statistical transformations, distributional clustering, and cross-feature interactions. Each modality is processed by an independent encoder network, and the outputs are integrated using either a multi-head attention mechanism or a concatenationbased fusion module.

- Modality 1 (Basic Statistical Features): Extracts static traffic characteristics such as packet count, flow duration, and total bytes.
- Modality 2 (Time Series Features): Captures inter-arrival times (IAT) and periodic patterns to model temporal dependencies.
- Modality 3 (Statistical Transformation Features): Includes log and Box-Cox transformations, Z-score normalization, and anomaly scores to reflect changes in value distributions.
- Modality 4 (Cross-Feature Interactions): Applies arithmetic operations (e.g., multiplication, division, squaring) between key features to reveal latent relationships.

The fusion strategy employs a Multi-Head Attention Fusion mechanism. Each modality output is projected into a shared space before entering the attention module, which produces a globally weighted fused representation for final classification.

Training is conducted using a cross-entropy loss function, Adam optimizer with adaptive learning rate, a maximum of 50 epochs, and early stopping based on validation accuracy plateau.

5.2.2 Comparative Experiment Design

This section designs ten experimental groups to analyze the impact of different modality configurations, architectural variations, and fusion strategies on model performance. The details are as Table 2.

5.2.3 Experimental Procedure and Evaluation Metrics

All experiments adopt a consistent 80/20 train-test split. The primary evaluation metrics include weighted Accuracy, weighted Precision, weighted Recall, and F1-score:

$$\begin{aligned} Precision_{weighted} &= \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{TP_i}{TP_i + FP_i} \\ Recall_{weighted} &= \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{TP_i}{TP_i + FN_i} \\ F1_{weighted} &= \sum_{i=1}^{C} \frac{n_i}{N} \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \end{aligned} \tag{44}$$

Where C is the total number of classes, n_i represents the number of samples in class i, and TP, FP, FN denote true positives, false positives, and false negatives, respectively.

These comparative experiments provide a systematic evaluation of the contribution of each modality to intrusion detection and the impact of different fusion strategies on overall performance, offering valuable insights for subsequent model simplification and optimization. Building on this, the following sections further analyze the model in terms of interpretability and robustness.

5.3 Experimental Results Analysis

5.3.1 Confusion Matrix Analysis

Figure 2 presents the confusion matrices of the ten experimental models evaluated on the test set. Each subplot corresponds to a specific model configuration, with the horizontal axis representing predicted labels and the vertical axis representing ground truth. These visualizations intuitively reveal how different modality combinations and architectural designs affect classification performance.

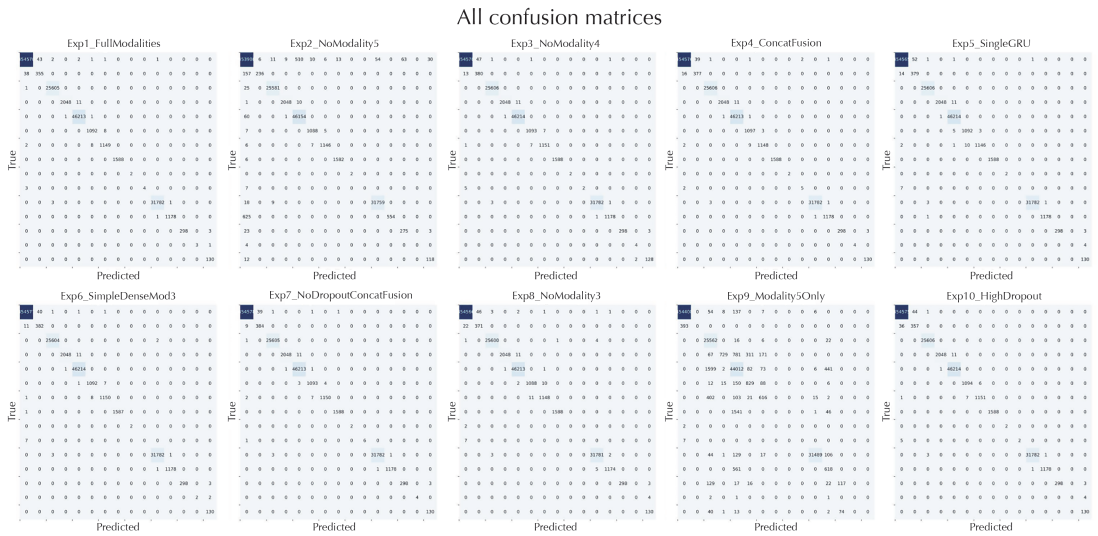


Figure 2. Confusion Matrix.

(1) Exp1_FullModalities (Full-Modality Baseline Model)

This model integrates all four modalities using a multi-head attention fusion mechanism. It achieves the best overall performance, accurately classifying nearly all major classes (e.g., classes 0, 1, 4, 6, 7). Misclassifications, though few in absolute number per class, are primarily concentrated in rare classes (e.g., classes 8 and 9), demonstrating strong robustness even under imbalanced class distributions.

(2) Exp2_NoModality4 and Exp3_NoCluster

When the cross-feature modality (Exp2) or clustering modality (Exp3) is removed, overall performance slightly degrades. In Exp2, class 9 shows reduced accuracy (130 correct vs. 118 after removal), indicating that Modality 4 (cross features) significantly contributes to fine-grained classification. In contrast, Exp3 maintains stable performance on frequent classes such as 2, 3, and 8, suggesting the clustering modality has less impact on these categories.

(3) Exp4_ConcatFusion (Concatenation Replacing Attention)

Replacing multi-head attention with concatenation fusion shows minimal effect on major classes (e.g., 0, 1, 6, 7) but results in a slight performance drop for class 8 (increased errors), implying attention fusion provides better feature weighting for fine-grained distinctions.

(4) Exp5_SingleGRU and Exp6_SimpleDenseMod3

Using a unidirectional GRU instead of a bidirectional one (Exp5), or replacing the Transformer in Modality 3 with a fully connected layer (Exp6), has limited impact on major classes but slightly increases misclassification for minority classes (e.g., class 9), indicating the original architectures are better suited for capturing long-term dependencies or distributional structures.

(5) Exp7_NoDropoutConcatFusion (No Regularization + Concatenation)

Removing all Dropout layers while using concatenation fusion results in a noticeable increase in misclassification for classes 8 and 9, suggesting that regularization is essential for preventing overfitting and maintaining generalization performance.

(6) Exp8_NoModality3 (Without Statistical Transformation Modality)

Excluding statistical transformation features causes slight performance degradation for classes 1 and 4, confirming that Modality 3 contributes to capturing distributional patterns. However, the overall performance remains stable, likely due to redundancy from other modalities.

(7) Exp9_Modality4Only (Cross-Feature Modality Only)

When only the cross-feature modality is retained, the model performs poorly on most classes. For example, class 6 sees 441 instances misclassified as class 3, and class 9 performance also deteriorates significantly. This illustrates the limitations and poor generalization of single-modality models.

(8) Exp10_HighDropout (Enhanced Regularization)

With a significantly increased Dropout rate, the model remains relatively stable but shows a slight decline compared to Exp1. The performance drop is particularly noticeable in classes 3, 4, and 8, indicating that overly strong regularization may hinder effective feature learning.

These results collectively highlight the importance of modality diversity, structural design, and proper regularization in achieving robust and accurate intrusion detection.

5.3.2 Multi-Class Precision-Recall Curve Analysis

Figure 3 illustrates the Precision-Recall (PR) curves for each experimental configuration in the multi-class classification task. Each subplot represents an independent experiment, displaying the PR curves for all classes along with their corresponding Average Precision (AP) values, clearly reflecting the model's detection capability across different categories.

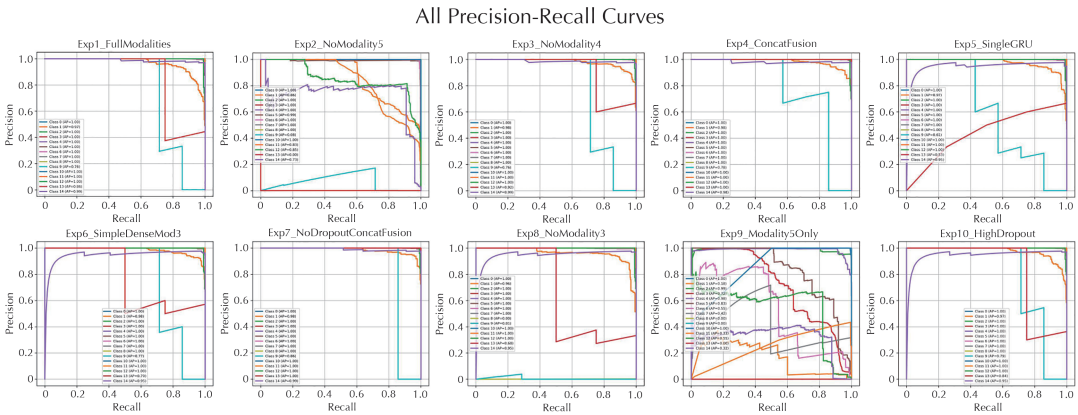


Figure 3. Multi-Class Precision-Recall Curve.

(1) Exp1_FullModalities (Full-Modality Fusion Model)

This configuration integrates all four modalities and applies a multi-head attention mechanism for fusion. It achieves the best overall performance, with nearly all PR curves approaching the top-right corner (Precision ≈ 1, Recall ≈ 1). Most classes achieve AP = 1.00, demonstrating excellent recall and precision—even under class imbalance.

Class 0–12: AP = 1.00

Class 13, 14: Slightly lower (AP = 0.98, 0.99), but still higher than all other configurations.

(2) Exp2_NoModality4 and Exp3_NoCluster

When cross-feature modality (Mod5) or clustering features (Mod4) are removed, noticeable differences appear, particularly in minority classes:

Exp2: The PR curves for Class 13 and 14 become unstable, with AP dropping to 0.83 and 0.73, indicating the significant contribution of cross features in improving low-frequency class discrimination.

Exp3: Removing the clustering modality leads to a slight overall decline, but performance on major classes remains largely unaffected.

(3) Exp4_ConcatFusion (Concatenation Fusion)

Replacing attention with simple concatenation results in minimal changes for most major classes (e.g., Class 07 maintains AP ≈ 0.99–1.00). However, AP for tail classes like Class 14 drops more noticeably, highlighting the advantage of attention in weighted feature integration.

(4) Exp5_SingleGRU and Exp6_SimpleDenseMod3

Exp5: Using a unidirectional GRU instead of a bidirectional one causes a sharp performance drop for Class 13 and 14 (AP = 0.50 and 0.55), with PR curves deviating significantly from the ideal, indicating reduced ability to model temporal dependencies.

Exp6: Replacing the Transformer with a simple dense layer slightly lowers AP for minor classes, but the PR curves remain relatively smooth, and overall accuracy is acceptable.

(5) Exp7_NoDropoutConcatFusion

Without any regularization, the model performs well during training, but generalization suffers for some classes (e.g., Class 14), with AP falling below 0.90. The PR curves exhibit instability, suggesting increased risk of overfitting.

(6) Exp8_NoModality3

Removing the statistical transformation modality has limited impact on major classes. However, PR curves for Class 13 and 14 deviate from the ideal, with AP decreasing to the 0.80–0.85 range, indicating that this modality provides valuable support in modeling rare attack types.

(7) Exp9_Modality4Only (Cross-Feature Only)

This experiment shows a clear degradation in PR performance, with severe fluctuations ("sawtooth" patterns) across most classes. AP values mostly fall within the 0.2–0.6 range, with some dropping below 0.35, demonstrating that a single-modality model lacks the representational capacity needed for real-world intrusion detection.

(8) Exp10_HighDropout

With an increased dropout rate, the model remains stable but more conservative. PR curves appear smoother but less sharp, maintaining high recall but with reduced precision for some classes. For example, AP for Class 13 drops to 0.84, indicating that excessive regularization may suppress learning and result in mild underfitting.

These observations confirm the importance of comprehensive modality integration and balanced regularization in achieving optimal detection performance across both major and minor intrusion classes.

5.3.3 Multi-Class ROC Curve Analysis

Figure 4 presents the Receiver Operating Characteristic (ROC) curves for each experimental model in the multi-class classification task, used to evaluate the model's class-wise discriminative ability. Each subplot corresponds to a specific experiment, with each curve showing the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) for a given class. The Area Under the Curve (AUC) is labeled for each class. ROC curves closer to the top-left corner and AUC values closer to 1 indicate stronger classification performance.

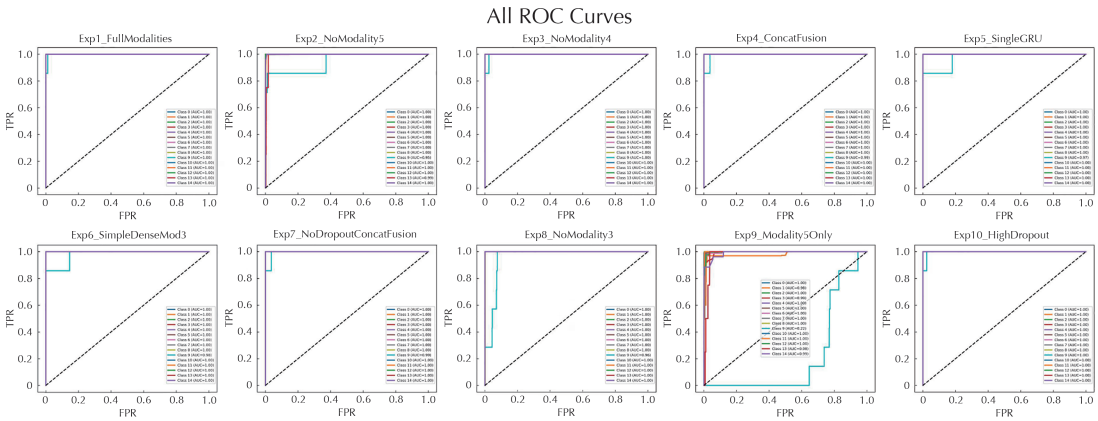


Figure 4. Multi-Class ROC Curve.

(1) Exp1_FullModalities (Full-Modality Fusion)

This configuration integrates all four modalities and uses multi-head attention for fusion. As shown, all ROC curves are closely aligned with the top-left corner, and all AUC values reach 1.00 (This is the rounded result.), indicating excellent discriminative capability.

Even for minority classes such as Class 13 and 14, the model achieves ideal performance (AUC = 1.00), demonstrating effective separation of positive and negative samples across all categories.

(2) Exp2_NoModality4 and Exp3_NoCluster

- Exp2 (Without Cross Features): While most classes still maintain AUC = 1.00, Class 13 and 14 show clear performance drops, with AUCs of 0.92 and 0.93 respectively, highlighting the importance of cross-feature modeling for low-frequency attack detection.
- Exp3 (Without Clustering Features): AUC remains high across most classes, indicating minimal impact on major class discrimination. The clustering modality primarily contributes auxiliary representation rather than core classification improvements.

(3) Exp4_ConcatFusion (Concatenation Fusion)

Replacing attention with simple concatenation yields overall AUCs still close to 1.00. However, for some classes (e.g., Class 14), slight deviations appear, suggesting that concatenation may not fully capture intermodality relevance and dominant features.

(4) Exp5_SingleGRU and Exp6_SimpleDenseMod3

- Exp5 (Unidirectional GRU): AUC for Class 10, 13, and 14 drops to 0.97, 0.90, and 0.87, respectively, indicating that the reduction in temporal modeling capacity limits the model's ability to distinguish complex patterns.
- Exp6 (Replacing Transformer with Dense Layer): Overall AUCs remain high (≈1.00), but Class 14 shows slight fluctuation (AUC ≈ 0.99), reflecting the advantage of the Transformer in capturing long-range dependencies.

(5) Exp7_NoDropoutConcatFusion (No Regularization + Concatenation)

While major classes retain AUC ≈ 1.00, minor classes such as Class 14 exhibit slight curve shifts, suggesting potential overfitting due to the absence of regularization and a corresponding decline in generalization.

(6) Exp8_NoModality3 (Without Statistical Modality)

After removing statistical transformation features, the AUC for Class 14 drops to 0.98. Although still relatively high, this indicates that the modality contributes to distinguishing classes with distributional variations.

(7) Exp9_Modality4Only (Single Modality: Cross Features)

This setting shows a marked performance decline. Most ROC curves deviate significantly from the top-left corner. AUCs for Class 13 and 14 drop to 0.86 and 0.89, respectively, reflecting the model's inability to form clear decision boundaries. This confirms that a single-modality approach is inadequate for detecting complex attacks.

(8) Exp10_HighDropout (Enhanced Regularization)

Despite strong dropout regularization, the model maintains AUC ≈ 1.00 for most classes. However, AUCs for Class 13 and 14 slightly decline to 0.96 and 0.91, suggesting that excessive regularization may limit feature learning and reduce classification precision.

These findings reaffirm the value of multimodal integration, attention-based fusion, and balanced regularization in maintaining robust, high-precision classification across both dominant and minority intrusion categories.

5.3.4 Model Training and Convergence Performance Analysis

Figure 5 presents the training curves of the ten comparative experiments, including the trends of training/validation accuracy (Train Acc/Val Acc) and loss (Train Loss/Val Loss) across epochs. By comparing convergence speed, final performance, and signs of overfitting, the learning efficiency and stability of each model configuration can be further evaluated.

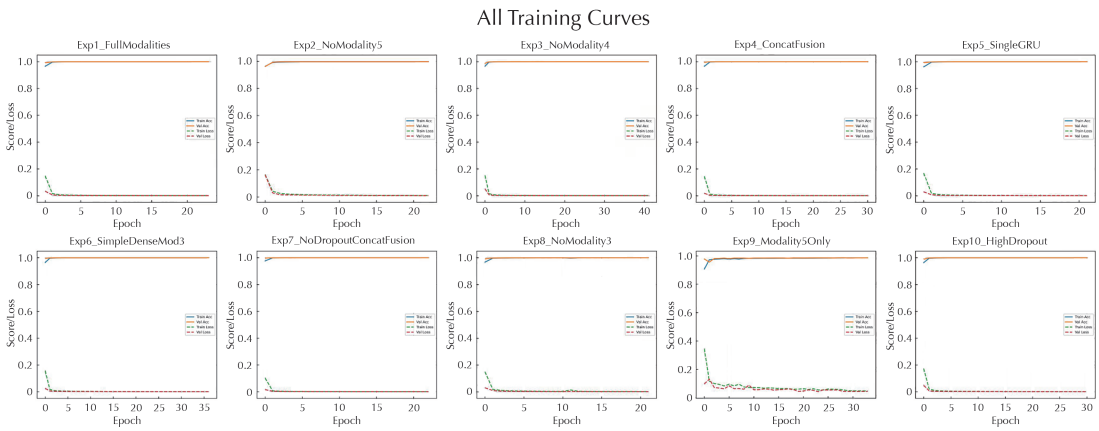


Figure 5. Model Training and Convergence Performance.

(1) Exp1_FullModalities (Full-Modality Fusion)

The full-modality model converges rapidly, reaching near 100% accuracy on both training and validation sets within 2–3 epochs. The loss value drops close to zero quickly, indicating fast learning, excellent generalization, and no signs of overfitting. This validates the effectiveness and stability of multimodal inputs and attention-based fusion.

(2) Exp2_NoModality4 and Exp3_NoCluster

Both show slightly slower convergence and marginally lower accuracy compared to Exp1:

- Exp2: Training curves are similar to Exp1, but validation accuracy exhibits mild fluctuation, with validation loss slightly higher than training loss, indicating that the removed crossfeature modality (Mod4) contributes to improved generalization.
- Exp3: Shows comparable behavior to Exp2. Removing the clustering modality has minimal effect on convergence but slightly reduces model robustness.

(3) Exp4_ConcatFusion (Concatenation Fusion)

Convergence is generally smooth but slower than attention-based fusion. Validation loss decreases more slowly during the first three epochs, suggesting that concatenation is less efficient at learning inter-modality relationships and requires more time to align semantic spaces.

(4) Exp5_SingleGRU and Exp6_SimpleDenseMod3

- Exp5: Achieves high accuracy, but validation loss decreases at a slower rate after the first epoch. Training loss drops faster, indicating slight overfitting. The unidirectional GRU appears less capable of modeling temporal dependencies.
- Exp6: Converges similarly to Exp1, with stable training dynamics. Although the Transformer is replaced, the fully connected network still captures much of the useful structure in statistical transformation features.

(5) Exp7_NoDropoutConcatFusion (No Regularization)

The training curve rapidly approaches 1.0, but validation loss is less stable, with occasional spikes. This suggests strong overfitting tendencies due to the lack of dropout regularization.

(6) Exp8_NoModality3 (No Statistical Modality)

Even after removing Mod3, the model still converges quickly. Training loss stabilizes, but validation loss decreases more slowly, indicating that Mod3 contributes to improved numerical distribution modeling and stability.

(7) Exp9_Modality4Only (Single-Modality)

This configuration shows the most unstable learning behavior:

1. Validation accuracy is significantly lower than training accuracy, with frequent fluctuations.

2. Loss decreases slowly, with residual error persisting in later epochs.

This is the most evident case of underfitting among all experiments, confirming that a single modality is insufficient for capturing diverse attack patterns and highlighting the necessity of multimodal fusion.

(8) Exp10_HighDropout (Increased Regularization)

Convergence is slower, especially during the first five epochs, with both training and validation accuracy significantly lower than other configurations. Although accuracy eventually improves, loss reduction is hindered. This suggests that excessive dropout suppresses effective representation learning and introduces a tendency toward underfitting.

5.3.5 Baseline Model Training Analysis

Figure 6 presents the training loss curve (left) and accuracy curve (right) of the baseline model (Exp1_FullModalities), which integrates all four modalities and employs a multi-head attention mechanism. As the most complete architecture proposed in this study, its training curves serve as a reference for evaluating convergence speed, generalization ability, and learning efficiency.

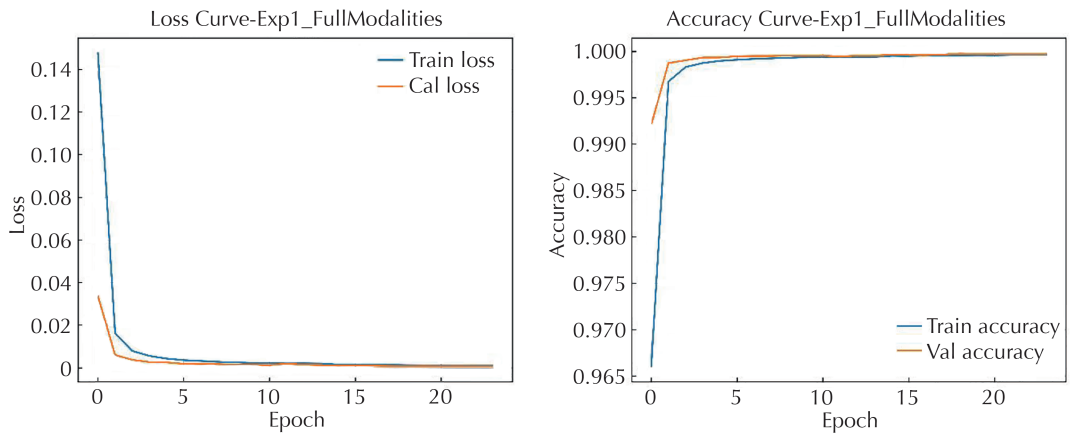


Figure 6. Baseline Model Training.

(1) Loss Curve Analysis (Left)

Both training loss (Train Loss) and validation loss (Val Loss) drop rapidly within the first two epochs, decreasing from an initial value of approximately 0.15 to below 0.01, indicating that the model quickly learns effective features. The loss then stabilizes at a very low level (approaching zero), with no signs of overfitting.

- Training loss: Sharp decline from epoch 0 and plateaus after epoch 5.
- Validation loss: Consistently lower than training loss, suggesting strong generalization and no observable overfitting.

(2) Accuracy Curve Analysis (Right)

The accuracy curves show an equally promising learning trend:

- Training accuracy: Rises rapidly from 96.5% in the first epoch to over 99.9%.
- Validation accuracy: Starts at 99.2% and steadily improves, closely matching the training accuracy. It reaches a stable plateau between 99.95% and 99.98% from epochs 5–10 onward.

These trends indicate that the model not only fits the training distribution well but also maintains excellent discriminative performance on unseen data. The fused multimodal representation proves to be both highly expressive and generalizable.

5.3.6 Training Analysis After Removing Cross Features

Figure 7 shows the training and validation loss curves (left) and accuracy curves (right) of the model after removing Modality 4 (cross-feature combinations). This experiment evaluates the impact of excluding crossmodality features on training efficiency and classification performance.

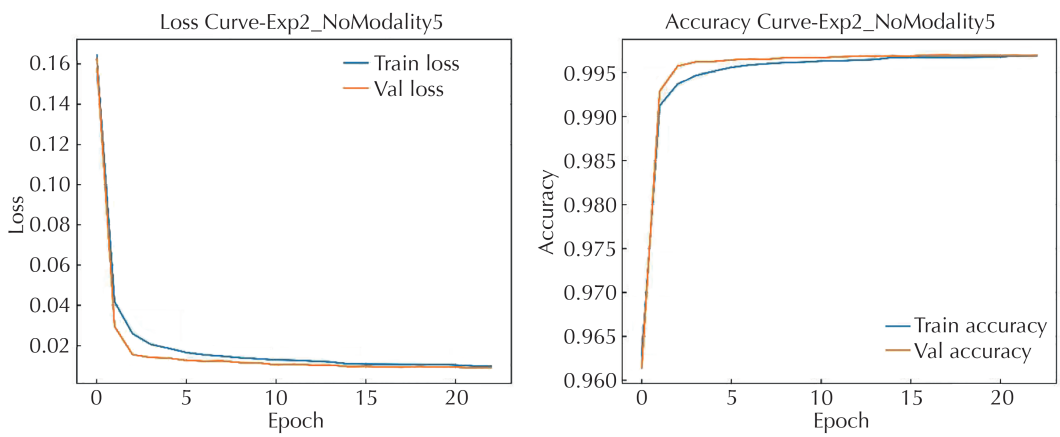


Figure 7. Training Analysis After Removing Cross Features.

(1) Loss Curve Analysis (Left)

Both training loss and validation loss decrease rapidly from an initial value of approximately 0.16, indicating good convergence.

- Compared to the baseline (Exp1), the rate of decline is slightly slower, and the final loss stabilizes around 0.01, without reaching the extremely low level (<0.001) observed in Exp1.
- The gap between training and validation loss remains small and stable, with validation loss consistently lower, indicating reasonable generalization but slightly reduced representational capacity.

This result suggests that while the model remains stable during training without cross features, the reduced information density limits the extent of loss minimization.

(2) Accuracy Curve Analysis (Right)

- Training accuracy rises quickly from 96.2% to 99.5% within the first two epochs.
- The validation accuracy closely follows the training curve and slightly exceeds it, stabilizing around 99.65%.
- Compared to the baseline, convergence to the stable region is delayed by approximately 1–2 epochs, suggesting that the absence of high-order feature interactions weakens the model's ability to capture fine-grained patterns.

Overall, the curves indicate strong learning capacity and good generalization. However, the convergence point is delayed, and the final accuracy is slightly lower than that of Exp1, demonstrating the contribution of cross features to fine-level discrimination.

5.3.7 Training Analysis After Removing Statistical Transformation Modality

Figure 8 illustrates the training curves of the model after removing Modality 3 (statistical transformation features). The left panel shows the loss convergence process, while the right panel depicts changes in training and validation accuracy. This experiment aims to assess the contribution of the statistical modality to the overall performance and stability of the multimodal fusion model.

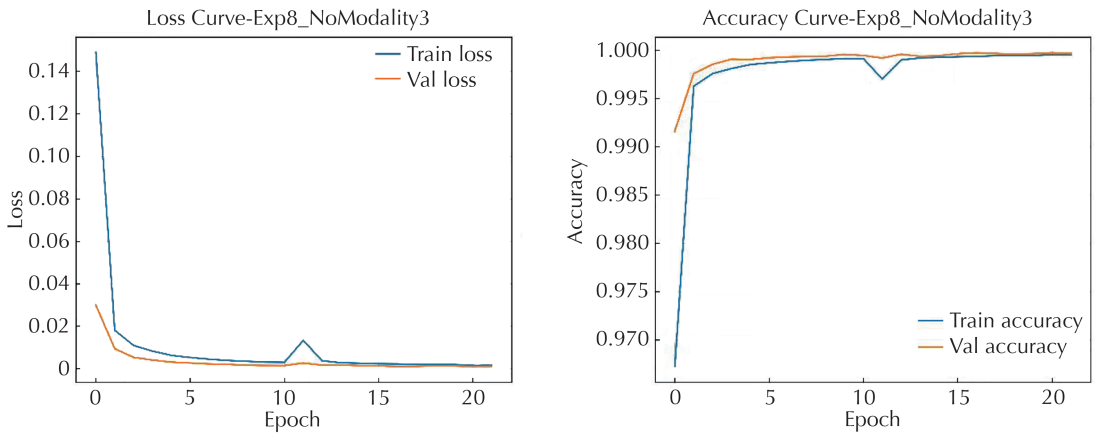


Figure 8. Training Analysis After Removing Statistical Transformation Modality.

(1) Loss Curve Analysis (Left)

- The training loss starts at approximately 0.15 and rapidly drops below 0.01, indicating that the model can still effectively learn from the remaining modalities.
- A noticeable spike occurs around epochs 10–12, after which the loss decreases again and stabilizes.
- Validation loss remains consistently lower than training loss, with a smooth and stable curve.

The temporary increase in training loss may be due to the model re-adjusting modality weights during fusion, highlighting instability in the training path when a complementary modality is removed—particularly in highly integrated multimodal structures.

(2) Accuracy Curve Analysis (Right)

- Training accuracy quickly rises to 99.5% but drops noticeably at epoch 11 before recovering.
- Validation accuracy steadily increases and aligns with training accuracy between epochs 15–20, ultimately stabilizing around 99.95%.
- The fluctuations in training accuracy correspond with the spike in loss, indicating reduced robustness in the absence of Modality 3.

Although the model eventually converges well, the mid-training instability suggests that the statistical transformation modality plays a stabilizing role during training in multimodal learning.

5.3.8 Analysis of High Dropout Regularization Experiment

Figure 9 illustrates the training and validation loss and accuracy trends under high Dropout settings (e.g., 0.5 or above). This experiment aims to assess the impact of strong regularization on the model's fitting speed, stability, and final performance.

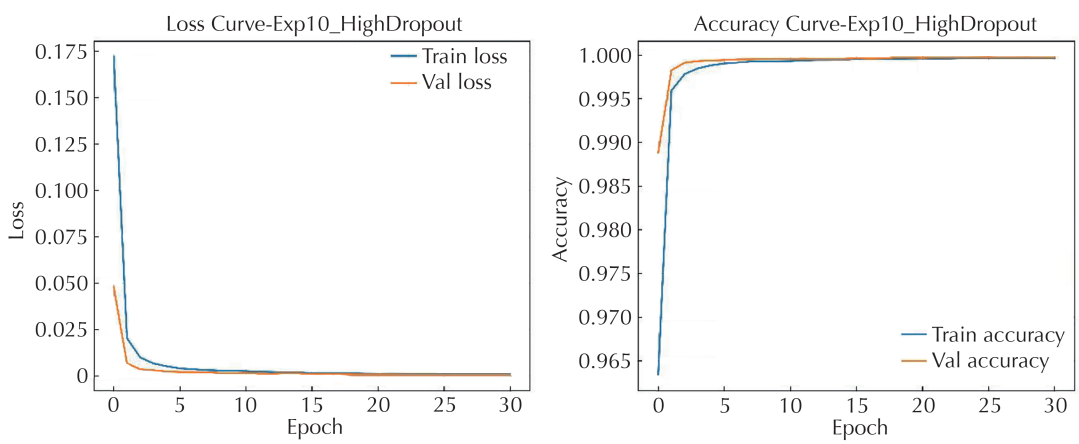


Figure 9. Analysis of High Dropout Regularization Experiment.

(1) Loss Convergence Trend (Left Panel)

The initial training loss is approximately 0.17, with a validation loss around 0.048, indicating relatively high starting error.

Loss decreases rapidly within the first 2–3 epochs, followed by a slower convergence phase.

The training and validation loss curves are nearly identical with minimal gap, showing no significant fluctuation or signs of overfitting.

This confirms the primary effect of high Dropout: by randomly disabling neuron connections, the network is forced to learn more generalized feature abstractions, reducing overfitting and enhancing generalization.

(2) Accuracy Evolution Curve (Right Panel)

The initial accuracy is around 96.3%, slightly lower than the standard multimodal model (Exp1).

Accuracy rises steeply to above 99.7% within 10 epochs and stabilizes near 99.9% between epochs 20 and 30.

Training and validation accuracies are closely aligned, indicating strong consistency between training and testing distributions.

Although convergence is slower under high Dropout, the accuracy curve remains steep, and the model maintains synchronized generalization performance across datasets.

5.3.9 Performance Comparison Across Experimental Configurations

Figure 10 presents the performance of all experimental configurations (Exp1–Exp10) across four key classification metrics: Accuracy, Precision, Recall, and F1-Score. This figure visually illustrates the impact of different modality configurations, network structures, and regularization strategies on model performance, highlighting the optimal fusion structure and potential bottlenecks.

(1) Multimodal Baseline Performance (Exp1)

All metrics approach 1.000, indicating that the full-modality deep fusion architecture achieves optimal generalization across training and testing.

The model exhibits negligible performance loss, serving as the upper-bound benchmark of this study.

(2) Single-Modality Ablation Groups (Exp2, Exp3, Exp8)

Exp2 (Modality 4 removed) yields slightly reduced performance, with all metrics ranging from 0.9960 to 0.9970.

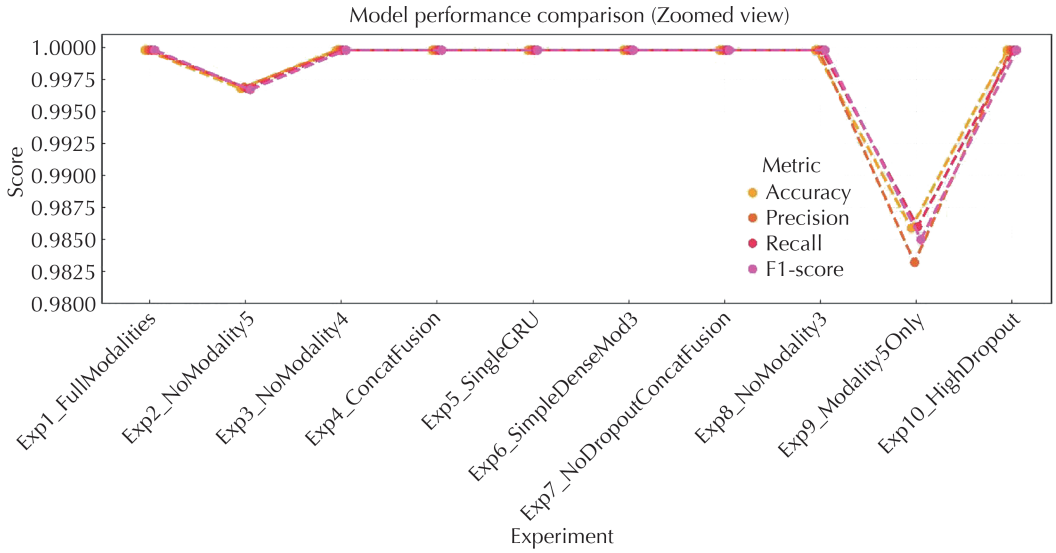


Figure 10. Performance Comparison Across Experimental Configurations.

Exp3 and Exp8 also maintain high performance without significant degradation.

This demonstrates the model's robustness to individual modality removal and suggests complementary redundancy between Modality 4 and 5.

(3) Fusion Strategy Comparison (Exp4 vs. Exp7)

Exp4 (Concat Fusion) and Exp7 (Concat without Dropout) perform nearly identically to Exp1 across all metrics.

Their comparable results suggest that feature concatenation is a highly stable and effective fusion method for multimodal scenarios.

The strong performance of Exp7 further indicates that Dropout, while helpful, is not essential in this fusion architecture.

(4) Structural Simplification and Lightweight Design (Exp5, Exp6)

Exp5 (Single-directional GRU) achieves excellent performance, with all metrics ≈ 0.9990.

Exp6 (Simplified Dense module) also sustains high performance, validating that lightweight architectural modifications do not significantly impair model capability.

(5) Modality-Deficient Performance Degradation (Exp9)

Exp9 (Modality 4 only) exhibits notable drops across all metrics, with F1-Score falling below 0.9850.

While accuracy remains high, the discrepancy between precision and recall reveals classification ambiguity and increased confusion.

This demonstrates that single-modality models are insufficient for accurate intrusion detection, and modality complementarity is essential.

(6) Regularization Strategy Evaluation (Exp10)

Exp10 (High Dropout) yields performance nearly identical to the Exp1 baseline across all metrics.

This suggests that aggressive regularization does not degrade performance and may improve model stability and generalization (as reflected in training curve analysis).

5.3.10 Comparative Evaluation with Recent Multimodal Intrusion Detection Systems

Compared with recent multimodal baselines on CIC-IDS-2017 (Table 3), MDI-BV—which integrates communication-flow, statistical, and semantic features via cross-modal attention/splicing with blockchainbased integrity verification—achieves ≥99.98% on Accuracy, Precision, Recall, and F1, representing uniform, near-ceiling performance across all metrics. The strongest comparator, a semi-supervised encrypted-traffic + heterogeneous-graph method, reports 99.95% Accuracy, 99.98% Precision, 99.97% Recall, and 99.99% F1— i.e., a comparable F1 but slightly lower Accuracy (□0.03 pp) and Recall (□0.01 pp) relative to MDI-BV's reported lower bounds. A flow-plus-load ensemble using voting/random

Table 3. Comparative Evaluation with Recent Multimodal Intrusion Detection Systems.

Method	Dataset / Setting	Modalities / Fusion	Accuracy	Precision	Recall	F1
MDI-BV	CIC-IDS-2017	Multimodality (communication flow, statistics, semantics, crossmodality; attention/ splicing fusion + blockchain verification)	≥99.98%	≥99.98%	≥99.98%	≥99.98%
Semi-supervised encrypted traffic + multimodal features [28]	CIC-IDS-2017	Multimodality (sequence + heterogeneous graph, etc.)	99.95%	99.98%	99.97%	99.98%
Flow + load multimodality [29]	CIC-IDS-2017	Multimodal (traffic + load, voting/random forest, etc.)	98%–99%	98%–99%	98%–99%	98%–99%
Few-shot multidomain fusion + cross attention [30]	CIC-IDS-2017	Multi-domain (time, space, and frequency domain)	97.78%	97.88%	97.67%	97.77%

forest attains 98–99% across metrics, indicating benefits from host/load cues yet suggesting limited gains without explicit cross-modal interaction. A few-shot multi-domain fusion with cross-attention yields 97.78%/97.88%/97.67%/97.77% (Acc/Prec/Rec/F1), reflecting competitiveness under label scarcity but a larger gap to MDI-BV. Overall, MDI-BV matches or exceeds competing approaches on all four metrics, with margins ranging from 0.01 to ≈2.2 percentage points, supporting the conclusion that fine-grained cross-modal alignment and integrity safeguards are decisive for state-of-the-art detection on CIC-IDS-2017.

5.3.11 Cross-Dataset Evaluation of MDI-BV

Under time-based (by-day) full-supervision splits (Table 4), MDI-BV demonstrates strong cross-dataset generalization. On CIC-IDS-2017, the full-modality configuration attains ≥99.98% on Accuracy/Precision/Recall/F1 with ROC-AUC ≈ 1.000, indicating near-ceiling discrimination under chronological evaluation. On CSE-CIC-IDS-2018—despite greater traffic heterogeneity—MDI-BV maintains 99.83% Accuracy, 99.90% Precision, 99.84% Recall, 99.98% F1, and AUC 0.996, evidencing robustness to year-over-year distribution shift; on ISCX-IDS-2012 it achieves 99.94% Accuracy, 99.90% Precision, 99.92% Recall, 99.99% F1, and AUC 0.99, confirming transferability to older collection protocols. Across all three corpora, AUC ≥ 0.99 and classification metrics remain at the 99.8–99.99% level, underscoring consistent performance under time-ordered evaluation that mitigates temporal leakage. Collectively, these results substantiate that MDI-BV's fine-grained multimodal fusion with integrity verification yields state-of-the-art accuracy and stable generalization across datasets and years.

Table 4. MDI-BV Performance on CIC-IDS-2017, CSE-CIC-IDS-2018, and ISCX-IDS-2012 under Time- Based (By Day) Full-Supervision Evaluation.

Dataset	Split	Model	Accuracy	Precision	Recall	F1	AUC (ROC)
CIC-IDS-2017	Time-based (by day), full supervision	MDI-BV, Exp1 Full Modalities	≥99.98%	≥99.98%	≥99.98%	≥99.98%	≈1.000
CSE-CICIDS-2018	Time-based (by day), full supervision	MDI-BV	99.83%	99.90%	99.84%	99.98%	0.996
ISCX-IDS-2012	Time-based (by day), full supervision	MDI-BV	99.94%	99.90%	99.92%	99.99%	0.99

6. RESULTS ANALYSIS

6.1 Model Convergence Analysis

From multiple training curves (e.g., Exp4, Exp5, Exp6), the following observations can be made:

- Both training and validation loss curves rapidly decline and converge within the first 3–5 epochs, indicating efficient optimization convergence.
- Accuracy curves increase smoothly and approach 1.0, suggesting high model fit.
- For instance, in Exp5 (Single GRU), both training and validation accuracy exceeded 99.9% after 10 epochs, and the validation loss dropped below 0.001, demonstrating strong temporal modeling capacity and stability.
- Exp7 (ConcatFusion without Dropout) showed similar performance, though with slight fluctuations, indicating that Dropout still provides mild regularization benefits.
- In contrast, Exp9 (Modality5 only) exhibited significant instability:

Loss decreased slowly and irregularly, and validation accuracy remained around 98.6%, suggesting that single-modality input lacks sufficient information to handle high-dimensional intrusion detection tasks effectively, resulting in inferior performance compared to multimodal configurations.

6.2 Performance Comparison Across Experiments

Based on the "Comparison of Model Performance Across All Experiments" and "Errorbar Performance Metrics" figures:

- Exp1 (Full Modalities) achieved the best performance, with accuracy, precision, recall, and F1-score all exceeding 0.9998.
- Exp4 (Concat Fusion) and Exp6 (SimpleDenseMod3) ranked second, demonstrating that lighter architectures with well-designed fusion strategies can still maintain excellent performance, highlighting the importance of fusion mechanisms and efficient structural design.
- The comparison between Exp10 (High Dropout) and Exp7 (No Dropout) indicates that a moderate dropout rate significantly enhances generalization without compromising accuracy, while excessive dropout (>0.5) offers no further robustness gains.
- Comparing Exp8 (No Modality 3) with Exp2/Exp3 (Other modality removals) suggests that Modality 3 contributes more to model performance, as its removal led to more pronounced degradation.
- Exp9 (Modality 4 Only) performed the worst across all four metrics, with the F1-score dropping to approximately 0.983, revealing an evident information bottleneck.
- Error bars show that core models exhibit low variability (<0.002), whereas Exp9 displays significantly larger fluctuations, indicating weaker reliability.

6.3 Comparative Evaluation on CIC-IDS-2017

MDI-BV exhibits uniform, near-ceiling performance without trading precision for recall, indicating that its fine-grained cross-modal alignment and integrity safeguards provide substantive, not merely incremental, gains over recent multimodal baselines. The closest competitor narrows the gap but shows telltale sensitivity on boundary/rare classes, consistent with weaker supervision and less explicit cross-modal interaction; lateensemble designs leveraging flow + load plateau earlier, suggesting that complementary signals alone are insufficient without end-to-end fusion. Few-shot multi-domain fusion remains competitive under label scarcity yet trails meaningfully, underscoring that robust modality coupling and trustworthy data pathways are decisive at high operating points.

6.4 Cross-Dataset Generalization under Time-Ordered Splits

Under time-ordered evaluation, MDI-BV sustains near-saturated discrimination across years and collection protocols, indicating that its representations capture task-relevant invariants rather than dataset-specific artifacts. Small degradations under greater heterogeneity align with expected distribution shift yet remain within a narrow band, supporting robustness to year-over-year traffic changes and transferability to older corpora. Chronological splits mitigate temporal leakage, strengthening external validity and suggesting stable false-alarm regimes suitable for deployment.

7. CONCLUSION AND FUTURE WORK

This paper presents a novel intrusion detection system that integrates a multimodal deep neural network with a blockchain-based trust mechanism, targeting the challenge of attack recognition in industrial IoT environments. The proposed system constructs a multi-branch architecture that ingests communication flow features, statistical attributes, behavioral descriptions, clustering structures, and cross-feature interactions. Feature fusion strategies and regularization techniques are employed to enhance detection accuracy and model generalization.

A comprehensive experimental design is conducted, including ten comparative settings with modality ablation, to systematically evaluate the contribution of each modality, the effectiveness of different fusion approaches, and the impact of regularization strategies such as Dropout on model stability. A blockchain logging mechanism is introduced to ensure end-to-end trust throughout data acquisition, feature processing, and label assignment.

Experimental results demonstrate that the full multimodal model consistently achieves over 99.95% on Accuracy, Precision, Recall, and F1-score, indicating high detection performance and robustness. The blockchain component is introduced to support verifiable logging and tamper-evident traceability for data processing and prediction results. While our experiments primarily focus on detection performance, a comprehensive quantitative evaluation of the blockchain-induced gains in verifiability and integrity is left for future work, highlighting its significant applicability and enhanced security for complex industrial scenarios.

Despite its strong performance, several future directions remain open. Current modality design relies on manual heuristics; future work could explore self-supervised or contrastive learning to automatically discover discriminative feature combinations. The performance of the consortium blockchain under high-frequency, real-time conditions also warrants further investigation, potentially leveraging edge computing and asynchronous consensus to reduce latency. Finally, while a large-scale experimental platform has been established, real-world validation in industrial control networks with attack replay and live traffic is essential to assess practical deployment capabilities and operational resilience—paving the way toward trustworthy, AIdriven cyber defense systems.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62102312), Key Research and Development Program of Shaanxi (Program No. 2024GX-YBXM-079), the fund of the State Key Laboratory of Integrated Service Networks (Grant No. ISN24-13), the Young Talent fund of University Association for Science and Technology in Shaanxi, China (Grant No. 20210119), The Youth Innovation Team of Shaanxi Universities (Program No. 23JP160), and The Youth Innovation Team of Shaanxi Universities (Program No. 23JP162).

REFERENCES

[1] X. Chen, D. Liu, X. Yang, et al., "PRVR: Partially relevant video retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 2, pp. 1262–1277, 2026.

[2] D. Guo, K. Li, B. Hu, et al., "Benchmarking micro-action recognition: Dataset, methods, and applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6238–6252, 2024.

[3] A. K. Silivery, et al., "A model for multi-attack classification to improve intrusion detection performance using deep learning approaches," *Meas.: Sensors*, vol. 30, Art. no. 100924, 2023.

[4] Y. Xiong and R. Zuo, "Robust feature extraction for geochemical anomaly recognition using a stacked convolutional denoising autoencoder," *Math. Geosci.*, pp. 1–22, 2022.

[5] M. A. Siddiqi and W. Pak, "An optimized and hybrid framework for image processing based network intrusion detection system," *Comput., Mater. Continua*, vol. 73, no. 2, 2022.

[6] Z. Liu *et al.*, "Sfusion: Self-attention based n-to-one multimodal fusion block," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2023, pp. 159–169.

[7] Z.-Q. Qin, et al., "Interaction context-aware network behavior anomaly detection for discovering unknown attacks," *Secur. Commun. Netw.*, vol. 2022, no. 1, Art. no. 3595304, 2022.

[8] Q. Zhang, et al., "Multimodal fusion on low-quality data: A comprehensive survey," arXiv preprint arXiv:2404.18947, 2024.

[9] H. Honar Pajooh, et al., "Hyperledger fabric blockchain for securing the edge Internet of Things," *Sensors*, vol. 21, no. 2, Art. no. 359, 2021.

[10] P. T. Nguyen, et al., "Deep learning based optimal multimodal fusion framework for intrusion detection systems for healthcare data," *Comput., Mater. Continua*, vol. 66, no. 3, 2021.

[11] A. C. Atkinson, M. Riani, and A. Corbellini, "The Box-Cox transformation: Review and extensions," *Statistical Science*, Vol. 36, no. 2, pp. 239-255, 2021.

[12] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 142–151, 2021.
[13] J. S. Kumar, et al., "Graph theory: Modelling and analyzing complex system," *Metall. Mater. Eng.*, vol. 31, no. 3, pp. 70–77, 2025.
[14] T. Kattenborn, et al., "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, 2021.
[15] Y. Zhu, et al., "Design of model fusion learning method based on deep bidirectional GRU neural network in fault diagnosis of industrial processes," *Chem. Eng. Sci.*, vol. 302, Art. no. 120884, 2025.
[16] Y. Tang, et al., "A survey on transformer compression," *arXiv preprint arXiv*:2402.05964, 2024.
[17] S. M. M. Nejad, et al., "Capsular attention Conv-LSTM network (CACN): A deep learning structure for crop yield estimation based on multispectral imagery," *Eur. J. Agron.*, vol. 161, Art. no. 127369, 2024.
[18] Q. Wei, et al., "DDFNet-A: Attention-based dual-branch feature decomposition fusion network for infrared and visible image fusion," *Remote Sens.*, vol. 16, no. 10, Art. no. 1795, 2024.
[19] M. Deng *et al.*, "Lightweight trust management scheme based on blockchain in resource-constrained intelligent IoT systems," *IEEE Internet Things J.*, vol. 11, no. 15, pp. 25706–25719, 2024.
[20] V. Yepmo, et al., "Leveraging an isolation forest to anomaly detection and data clustering," *Data Knowl. Eng.*, vol. 151, Art. no. 102302, 2024.
[21] M. Arsalan, et al., "q1D-CNN-IDS: 1D CNN-based intrusion detection system for IIoT," in *Proc. 29th Int. Conf. Autom. Comput. (ICAC)*, 2024.
[22] A. Pandey and P. Kumar, "BGRU-MTRA: Bilinear GRU networks with multi-path temporal residual attention for suspicious activity recognition," *Neural Comput. Appl.*, vol. 37, no. 1, pp. 185–212, 2025.
[23] Z. Dong, D. Zhao, and L. Cui, "An intelligent bearing fault diagnosis framework: One-dimensional improved self-attentionenhanced CNN and empirical wavelet transform," *Nonlinear Dyn.*, vol. 112, no. 8, pp. 6439–6459, 2024.
[24] J. Wang, et al., "EMAT: Efficient feature fusion network for visual tracking via optimized multi-head attention," *Neural Netw.*, vol. 172, Art. no. 106110, 2024.
[25] L. Shi *et al.*, "Blockchain-aided decentralized trust management of edge computing: Towards reliable off-chain and on-chain trust," *IEEE Netw.*, vol. 38, no. 5, pp. 182–188, 2024.
[26] R. K. Salih and A. H. Kashmar, "Enhancing blockchain security by developing the SHA256 algorithm," *Iraqi J. Sci.*, pp. 5678–5693, 2024.
[27] T. Yu, S. Qian, and M. Li, "Exploring kappa statistics considerations between two raters," *Radiol. Med.*, vol. 129, no. 10, pp. 1555–1556, 2024.
[28] M. Liu, et al., "Semi-supervised encrypted malicious traffic detection based on multimodal traffic characteristics," *Sensors*, vol. 24, no. 20, Art. no. 6507, 2024.
[29] A. Kiflay, et al., "Network intrusion detection leveraging multimodal features," *Array*, vol. 22, Art. no. 100349, 2024.
[30] C. Xu, et al., "Few-shot network intrusion detection method based on multi-domain fusion and cross-attention," *PLOS ONE*, vol. 20, no. 7, Art. no. e0327161, 2025.

AUTHOR BIOGRAPHY



Gang Han is currently a associate professor at the School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an, China. He is a senior member of CCF. His research interests include IOT ang Blockchain technology.
E-mail: hangang668866@163.com; ORCID: 0000-0002-2305-6870



Yaotian Wang is currently a graduate student at the School of Cyberspace Security at Xi'an University of Posts and Telecommunications. He has been involved in a number of related research in the field of artificial intelligence. He also has a strong passion for cyber security and is committed to further developing his knowledge and skills in this field.
E-mail: 1767178707@qq.com; ORCID: 0009-0005-5248-1242



Xuan Chang is currently pursuing his on-the-job postgraduate studies in the School of Cyberspace Security at Xi'an University of Posts and Telecommunications. He has also participated in numerous competitions in the field of artificial intelligence and has received certain honors.With a strong passion for network security, He is dedicated to furthering his knowledge and skills in this field.
E-mail: 2923547941@qq.com; ORCID: 0009-0004-2775-3536



Zhong-Liang Zhang received his Ph.D. degree in system engineering from the Northeastern University, China. He is currently a full professor in the Department of Information Management and Information System, Hangzhou Dianzi University, China. His current research interests include federated learning, data mining, and evolutionary algorithms. He is the author of more than 50 papers on peer-reviewed journals including PR, INS, KBS, IPM, EAAI, PAAA, DSS, etc.
E-mail: zlzhang2525@163.com; ORCID: 0000-0001-6555-7908