An Intrusion Detection Method Based on Transformer and Transfer Learning

Kunpeng Wang*, Xiaoling Bai, Bohai Tang, Yunsong Ge

Harbin Institute of Information Technology, Harbin,150431, P.R. China.

***Corresponding Author**

DOI: https://dx.doi.org/10.47772/IJRISS.2025.903SEDU0449

Received: 24 July 2025; Accepted: 29 July 2025; Published: 30 August 2025

ABSTRACT

To overcome the limitations of existing intrusion detection systems, particularly in the areas of encrypted traffic analysis, cross-domain adaptation, and small-sample learning scenarios, this study proposes the TTL-IDS model, which integrates the Transformer architecture with transfer learning techniques. The model incorporates a multi-head self-attention mechanism with position encoding to effectively capture long-range dependencies in network traffic, a critical capability for identifying subtle and complex attack patterns. Furthermore, a hierarchical feature transfer framework is introduced, leveraging domain adversarial training to facilitate robust knowledge transfer from the source domain to the target domain. Experimental results validate the effectiveness of TTL-IDS in enhancing detection accuracy and domain generalization. This research not only demonstrates the model's practical advantages but also offers novel insights and methodologies for strengthening security in dynamic and heterogeneous network environments.

Keywords: Intrusion detection; Transformer; Transfer learning; Domain adaptation; Network security

INTRODUCTION

The rapid proliferation of digital technology has triggered an unprecedented surge in data generation and connectivity, fundamentally transforming how individuals and organizations operate (Aslam, 2024). This acceleration, further intensified by global disruptions such as the COVID-19 pandemic, has also reshaped the landscape of criminal activity. Traditional crimes are increasingly migrating into cyberspace, fueling a new wave of cyber threats (Malick et al., 2024). Consequently, the cybersecurity environment has become more complex and perilous, with organizations facing a growing frequency and sophistication of attacks.

Traditional security measures, often reliant on static rules and signature-based detection are struggling to keep pace. Such approaches tend to be reactive and prove ineffective against rapidly evolving threats (Anderson, 2022). According to the 2025 *Verizon Data Breach Investigation Report*, global enterprises experience network attacks in 44% of cases on average, while the occurrence of zero-day exploits has risen by 34% compared to the previous year (Verizon, 2025).

As adversarial tactics evolve, signature-based intrusion detection systems (IDS) such as Snort and Suricata, which depend heavily on predefined rules and signature libraries, are losing effectiveness (Naayini, 2025). The widespread adoption of encrypted communication protocols, particularly TLS 1.3 has made over 80% of network traffic inaccessible to deep packet inspection (DPI), severely constraining traditional detection capabilities (Sharma & Lashkari, 2025). Moreover, advanced persistent threats (APTs) often exhibit prolonged latency periods; for example, the SolarWinds breach went undetected for up to 178 days, rendering conventional statistical models that rely on short-term patterns far less effective (Cimpanu, 2020).

The challenge is further compounded by the dynamic nature of modern networks and the heterogeneity of connected devices, especially within Internet of Things (IoT) ecosystems. These factors contribute to feature distribution drift, which degrades detection performance. For instance, the UNSW-NB15 dataset reports a 31.7%

drop in cross-device detection accuracy under such conditions (Moustafa & Slay, 2015).

In response to these limitations, researchers have turned to deep learning-based approaches as alternatives to traditional IDS. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown potential in extracting spatial and temporal features, respectively. However, these models are not without shortcomings. Bi-directional LSTM (BiLSTM) networks, for example, suffer from gradient vanishing issues when modeling long sequences, resulting in a reported 22% reduction in F1 scores for sessions exceeding 1,000 time steps (Laghrissi et al., 2021). Furthermore, their generalization capability remains limited in low-resource scenarios; when labelled data in the target domain is scarce, fewer than 1,000 annotated samples, the recall rate of the ResNet-1D model drops to 68% (Althiyabi et al., 2024).

LITERATURE REVIEW

Recent advancements in intrusion detection research have primarily progressed along two promising directions aimed at overcoming the limitations of traditional systems: Transformer-based architectures and cross-domain transfer learning.

The first direction leverages Transformers, originally introduced by Vaswani et al. (2017) which employ a self-attention mechanism to effectively capture long-range dependencies in sequential data, thereby eliminating the fixed-length constraints of earlier models. This architecture has demonstrated strong potential in cybersecurity applications; for example, Transformer-based approaches have achieved an accuracy of 95.3% on the VirusTotal dataset for malware detection. However, many such models overlook the semantic information embedded within network protocols, which can be crucial for precise threat identification. Addressing this limitation, Rahali and Akhloufi (2021) proposed a Transformer-based malware detection framework built on Bidirectional Encoder Representations from Transformers (BERT). Their approach conducted static analysis on Android application source code and used preprocessed features to classify malware samples into known categories, underscoring the value of pretrained language models in security contexts.

The second major direction focuses on cross-domain transfer learning, which seeks to improve model generalization across diverse network environments. Xue et al. (2022), for instance, utilized adversarial generative networks (GANs) to facilitate feature transfer across IoT devices; however, their model exhibited a high false positive rate when detecting DDoS attacks on the CIC-IDS2017 dataset. More recently, Rezaabad et al. (2022) demonstrated that combining domain adaptation with contrastive learning can enhance detection accuracy by 19.8% in low-resource scenarios. Despite these gains, the approach significantly increased computational costs, tripling training complexity and limiting its practicality for real-time applications.

Nonetheless, a critical challenge remains unresolved: effectively balancing a model's ability to capture long-range dependencies with the need for efficient and stable generalization across domains. Current methods face notable constraints. For example, the use of absolute positional encoding in flow-based models introduces high phase sensitivity, leading to a 12.4% increase in prediction error. Likewise, gradient conflicts in domain adversarial training often result in unstable convergence, with observed performance fluctuations of up to ±8.7%.

To address these challenges, this study proposes Transformer-based Transfer Learning Intrusion Detection System (TTL-IDS), a novel framework designed to enhance both scalability and cross-domain robustness. TTL-IDS incorporates a sparse attention mechanism, reducing computational complexity from $O(n^2)$ to $O(n \log n)$ and making it more suitable for large-scale traffic analysis. In addition, the model employs a dual-path domain adversarial neural network (Dual-Path DANN) that disentangles domain-invariant and domain-specific components within traffic features. To improve training stability, the framework integrates the Dynamic Weight Averaging (DWA) algorithm, which dynamically balances loss contributions between source and target domains, thereby enhancing performance in heterogeneous environments.

Beyond transfer learning and attention mechanisms, recent advances in deep learning-based intrusion detection have increasingly focused on automatic feature extraction and sequence modeling. For example, Qazi et al. (2022) proposed a 1D-CNN model that achieved 89.2% accuracy on the KDD99 dataset by scanning network traffic byte sequences using convolutional kernels. However, the fixed-size nature of these kernels limits the

model's ability to capture global dependencies, particularly in variable-length sessions.

Parallel developments in transfer learning have sought to mitigate data scarcity in target domains. Xu et al. (2020) introduced a meta-learning-based framework that treats the comparison of network traffic pairs (normal vs. malicious) as a core learning task. Trained and tested across multiple datasets, their model achieved an average detection rate of 99.62%, demonstrating strong generalization to previously unseen attacks in few-shot learning scenarios.

Despite these advances, there remains a critical need for a unified framework that combines efficient sequence modeling, semantically rich feature extraction, and robust cross-domain generalization. TTL-IDS addresses this gap by integrating the strengths of Transformer architectures and transfer learning into a cohesive, scalable intrusion detection solution tailored for dynamic, real-world network environments.

Research Method

The TTL-IDS model comprises four key components: a multi-modal embedding layer that integrates protocol fields, payload bytes, and timing features; a position-aware Transformer encoder designed to capture long-range dependencies; a dual-channel domain adaptation network that separates domain-invariant and domain-specific features; and a dynamic classifier that enhances prediction confidence in low-sample scenarios.

Multimodal embedding layer

The multimodal embedding layer is designed to integrate heterogeneous network traffic features, namely protocol-specific information, payload content, and temporal characteristics, into a unified representation space suitable for downstream intrusion detection tasks. This layer comprises three key components.

First, protocol field embedding transforms categorical protocol types (e.g., TCP, UDP, HTTP) into fixed-dimensional vector representations. Pre-trained word embedding models such as Word2Vec or BERT can be employed to capture semantic relationships among different protocol types, enabling the model to leverage contextual similarities in network communications.

Second, payload byte embedding processes the raw payload data contained in network packets. CNNs are used to extract meaningful patterns from byte-level sequences, encoding both structural and content-based characteristics into dense vector formats.

Third, temporal feature encoding captures time-dependent patterns by transforming attributes such as inter-arrival times and session durations into numerical representations suitable for sequence modeling.

The resulting feature vectors, representing protocol semantics, payload content, and temporal dynamics are concatenated to produce a comprehensive multimodal embedding vector, as defined in Equation (1):

$$E = [E_{protocol}; E_{payload}; E_{temporal}] \quad (1)$$

In this expression, $E_{protocol}$, $E_{payload}$ and $E_{temporal}$ denote the embedding vectors derived from protocol fields, payload bytes, and temporal features, respectively.

Position aware Transformer encoder

The position-aware Transformer encoder is designed to effectively model long-range dependencies in network traffic sequences by combining a learnable position encoding scheme with a multi-head self-attention mechanism. This design allows the model to better capture the sequential nature of traffic flows, which is essential for detecting subtle patterns in both benign and malicious activities.

To incorporate positional information, a learnable relative position encoding matrix P is used. This matrix has dimensions $d_{\text{model}} \times d_{\text{model}}$, where d_{model} denotes the dimensionality of the model's hidden layers. Unlike

fixed or sinusoidal encodings, learnable relative encodings allow the model to adaptively learn the significance of relative positions between tokens, enhancing its generalization across sessions of varying lengths.

The core of the encoder is the multi-head self-attention mechanism, which enables the model to capture dependencies across different positions in the sequence by attending to multiple subspaces in parallel. Each attention head computes its own set of queries (Q), key (K), and value (V) vectors, and the attention output is computed using the standard scaled dot-product attention formula:

$$Attention\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{R}) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k is the dimensionality of the key vectors. This allows the model to learn complex interactions and contextual relationships across the sequence, which are crucial for accurately detecting anomalies or attacks.

To address the high computational cost typically associated with full self-attention, a sparse attention mechanism is incorporated. This technique reduces the computational complexity from $O(n^2)$ to $O(n \log n)$, where n is the sequence length, by limiting attention calculations to a subset of relevant positions. This makes the encoder scalable to large-scale network traffic data.

Dual channel domain countermeasure network

The dual-channel domain adversarial network is designed to enhance the model's ability to generalize across domains by explicitly separating domain-invariant and domain-specific components within the learned traffic feature representations. This is achieved through a domain adversarial training strategy, which encourages the shared encoder to learn features that are discriminative for classification while remaining indistinguishable across domains. By minimizing the discrepancy between the source and target domain feature distributions, this approach improves the cross-domain adaptability of the IDS.

In this framework, a domain classifier is introduced with the objective of identifying whether a given feature representation originates from the source or target domain. The input to the domain classifier is the high-level feature representation H, produced by the Transformer encoder. The classifier outputs a domain label y_{domain}, indicating the predicted domain (i.e., source or target). During training, this classifier is updated to improve its ability to distinguish between domains, while the encoder is trained to fool the domain classifier, thereby promoting the extraction of domain-invariant features.

To achieve this adversarial dynamic, the model employs a domain adversarial loss function that guides the training process. Specifically, the loss is defined as:

$$l_{domain} = -E_{H\sim D_{source}}\, log\, y_{domain}(H) - E_{H\sim D_{target}}\, log\big(1 - y_{domain}(H)\big) \tag{3}$$

In this formulation, D_{source} and D_{target} represent the data distributions of the source and target domains, respectively, and H denotes the extracted features from the Transformer encoder. This loss encourages the encoder to produce representations that are indistinguishable with respect to domain, thereby reducing domain shift and improving performance in cross-domain intrusion detection tasks.

Dynamic classifier

The dynamic classifier is designed to improve prediction performance in cross-domain scenarios, particularly when the target domain suffers from data scarcity. To achieve this, the model employs the DWA algorithm, which dynamically balances the loss contributions from the source and target domains. By adaptively adjusting the loss function weights during training, the classifier can optimize its learning process and maintain robust prediction confidence, even in small-sample settings.

The DWA algorithm computes domain-specific loss weights based on the exponential moving average of recent loss values for the source and target domains. This mechanism ensures that more attention is given to the domain

with higher training difficulty (i.e., larger loss), allowing the model to adaptively shift focus as needed. The weights for the source and target domains at time *t* are calculated as:

$$\alpha_t = \frac{exp\big(\lambda \cdot l_{source}(t)\big)}{exp\big(\lambda \cdot l_{source}(t)\big) + exp\left(\lambda \cdot l_{target}(t)\right)} \tag{4}$$

$$\beta_t = \frac{exp\left(\lambda \cdot l_{target}(t)\right)}{exp\big(\lambda \cdot l_{source}(t)\big) + exp\left(\lambda \cdot l_{target}(t)\right)} \tag{5}$$

Here, α_t and β_t represent the dynamically computed weights for the source and target domain losses, respectively, λ is a temperature parameter that controls the sensitivity of the weighting mechanism and, $l_{source}(t)$ and $l_{target}(t)$ denote the classification losses at time *t* for the source and target domains.

The classifier loss function then integrates these weights to form a composite loss that guides model optimization. It is defined as:

$$l_{classifier} = \alpha_t \cdot l_{source} + \beta_t \cdot l_{target} \tag{6}$$

This dynamic loss formulation allows the model to adaptively balance the influence of both domains during training, which is particularly beneficial in few-shot or imbalanced domain settings.

Experiments and analysis

Experimental setup

To evaluate the effectiveness of the proposed intrusion detection framework, experiments were conducted using the CIC-IDS2017 dataset (Ring et al., 2019), a widely used benchmark in network security research. Developed by the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC), this dataset closely resembles real-world traffic data captured in PCAP format. It includes a diverse set of seven attack types, such as DDoS, Botnet, and Brute Force, among others. The dataset contains approximately 2.8 million samples and features 78 dimensions, covering protocol headers, statistical traffic patterns, and temporal attributes. This comprehensive feature set enables the training and evaluation of models in detecting anomalous behaviors and identifying potential security threats. For the purpose of experimentation, the dataset was partitioned into 60% for training, 20% for validation, and 20% for testing, ensuring balanced performance assessment and model tuning.

To benchmark performance, a selection of representative algorithms was used for comparison. These include Random Forest (RF) as a traditional machine learning baseline; 1D-CNN and BiLSTM as foundational deep learning models; BERT-Flow as a Transformer-based variant; and DANN as a domain-adaptive model employing transfer learning. These models were selected to assess the proposed framework against both conventional and state-of-the-art approaches across multiple paradigms.

Model performance was assessed using a set of both accuracy-focused and efficiency-focused evaluation metrics. The main performance indicators include Accuracy, Weighted F1 Score (F1), and ROC-AUC. To evaluate computational efficiency, the training time in hours was recorded. Additionally, to assess the model's ability to generalize across domains, a Target Adaptation Rate (TAR) was used, calculated as the ratio of the F1 score in the target domain to that in the source domain ($TAR = F1_{target}/F1_{source}$).

All experiments were conducted using a NVIDIA A100 GPU with 40GB of memory. The models were trained using the AdamW optimizer, with an initial learning rate of 3e-4 and a weight decay factor of 0.01. The maximum number of training epochs was set to 100, with an early stopping policy applied when validation performance failed to improve over 10 consecutive epochs. This setup ensures a fair and efficient evaluation of the proposed method against strong baselines.

RESULTS AND ANALYSIS

Performance comparison experiment

According to the selected typical model algorithms, performance comparison experiments were conducted on the CIC-ISDS2017 dataset. The experimental results are shown in Table 1.

Table 1. Performance comparison of models' algorithm on the CIC-IDS2017 dataset (%)

Model	Accuracy	F1	ROC-AUC	Training time (h)
Random Forest	89.2	88.3	0.912	0.5
1D-CNN	92.7	91.8	0.934	3.2
BiLSTM	93.1	92.5	0.941	5.8
BERT-Flow	94.5	93.9	0.962	8.1
DANN	95.2	94.3	0.971	9.5
TTL-IDS (our)	**96.7**	**96.2**	**0.983**	**6.3**

Table 1 presents a comparative analysis of the performance of algorithms on the CIC- IDS2017 dataset. The conclusions obtained can be summarized as follows:

The F1 score of TTL-IDS surpasses the optimal baseline (DANN) by 1.9%, validating the synergistic benefits of combining Transformer with transfer learning. The inference latency stands at 8.2ms (in contrast to 15.7ms for BERT-Flow), fulfilling the requirements for real-time detection. Although the training time is 8.6% longer compared to BiLSTM, the AUC has seen a 4.2% improvement, indicating a superior balance between efficiency and accuracy.

Cross domain detection experiment

Source domain: CIC-ISDS2017 (fully marked); Target field: UNSW-NB15 (only 1% labeled data), evaluation index: target field F1 value TAR. The performance comparison of cross domain detection is shown in Table 2.

Table 2. Cross domain detection performance comparison (%)

Model	F1 (Target Domain)	TAR
DANN	76.5	81.1
Meta-IDS	78.2	83.0
TCN-LSTM	72.8	77.3
TTL-IDS(our)	**85.1**	**90.2**

From the experimental data results in Table 2, the following conclusions can be drawn:

The experimental results demonstrate that the proposed dual-channel domain adversarial network significantly improves cross-domain intrusion detection performance. Specifically, it achieves an 8.6% increase in F1 score on the target domain compared to the baseline DANN model, while attaining a TAR of 90.2%. These results indicate that the model is capable of effectively transferring knowledge from the source domain to the target domain, even under conditions of data scarcity or distributional shift. Furthermore, analysis of the learned feature representations reveals that the shared feature space in TTL-IDS is more discriminative and compact, with within-class distances reduced by 37%, suggesting improved clustering of semantically similar traffic samples across domains.

Overall, the proposed TTL-IDS framework outperforms all benchmark models across multiple evaluation metrics, underscoring its effectiveness in both accuracy and generalization. These findings validate the robustness of the algorithm and highlight its potential as a practical solution for intrusion detection in dynamic and heterogeneous network environments. Moreover, the integration of Transformer-based modeling, domain adversarial learning, and dynamic loss balancing introduces a novel architectural paradigm for cybersecurity applications, offering both theoretical contributions and real-world applicability in modern threat detection systems.

CONCLUSION

The proposed TTL-IDS model effectively addresses several key limitations of existing IDS, particularly in handling encrypted traffic, achieving cross-domain adaptability, and maintaining reliable performance in small-sample scenarios. By integrating Transformer-based architectures with transfer learning techniques, TTL-IDS learns rich, transferable representations that generalize effectively across heterogeneous network environments. Its architecture combines multimodal embedding layers for fusing diverse traffic features, position-aware Transformer encoders for capturing long-range dependencies, dual-channel domain adversarial networks for disentangling domain-invariant and domain-specific features, and dynamic classifiers that adaptively balance domain losses to improve generalization under data scarcity.

Experimental results confirm that TTL-IDS delivers strong detection accuracy and robust cross-domain performance, outperforming both traditional and deep learning-based baselines. Moreover, the model achieves a favourable efficiency–accuracy trade-off, maintaining low inference latency and reasonable training time, thus meeting the practical requirements of real-time intrusion detection.

From a real-world integration perspective, TTL-IDS is designed with compatibility in mind. Its modular architecture allows seamless deployment alongside existing Security Information and Event Management (SIEM) platforms, enabling security analysts to incorporate its outputs into centralized monitoring workflows. Furthermore, its lightweight inference mode supports containerization and cloud-based deployment, facilitating scalability across distributed or hybrid infrastructures. These characteristics make TTL-IDS well-suited for both enterprise networks and resource-constrained environments, such as IoT or edge devices.

Nonetheless, certain limitations remain. While the sparse attention mechanism improves scalability, deploying TTL-IDS in high-throughput, real-time environments may still pose computational challenges, particularly when processing large-scale encrypted traffic. In addition, the dual-path adversarial learning framework, though effective, increases training complexity and may require hardware acceleration for large-scale deployments.

Future research will focus on further optimizing the model for live production environments, including reducing computational overhead, enhancing scalability for distributed systems, and developing lightweight variants for edge-based detection. Another promising direction involves integrating continual learning strategies to adapt to evolving attack patterns without extensive retraining, ensuring sustained effectiveness in dynamic and rapidly changing network contexts.

ACKNOWLEDGMENTS

This work was supported by the School Fund of the Natural Science Foundation of Heilongjiang Province, China.

REFERENCES

1. Althiyabi, T., Ahmad, I., & Alassafi, M. O. (2024). Enhancing IoT security: A few-shot learning approach for intrusion detection. Mathematics, 12(7), 1055.
2. Anderson, R. (2022). Why traditional cybersecurity is failing. Journal of Network Defense Strategies, 12, 34–47.
3. Aslam, M. (2024). Ai and cybersecurity: an ever-evolving landscape. International Journal of Advanced Engineering Technologies and Innovations, 1.
4. Cimpanu, C. (2020). US charges five hackers from Chinese state-sponsored group APT41. Zdnet.

Retrieved July 25, 2025, from https://www.zdnet.com/article/us-charges-five-hackers-part-of-chinese-state-sponsored-group-apt41

5. Laghrissi, F., Douzi, S., Douzi, K., & Hssina, B. (2021). Intrusion detection systems using long short-term memory (LSTM). Journal of Big Data, 8(1), 65.
6. Mallick, M. A. I., & Nath, R. (2024). Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments. World Scientific News, 190(1), 1-69.
7. Moustafa, N., & Slay, J. (2015). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In 2015 military communications and information systems conference (MilCIS) (pp. 1-6). IEEE.
8. Naayini, P., Myakala, P. K., & Bura, C. (2025). How ai is reshaping the cybersecurity landscape. Available at SSRN 5138207.
9. Qazi, E. U. H., Almorjan, A., & Zia, T. (2022). A one-dimensional convolutional neural network (1D-CNN) based deep learning system for network intrusion detection. Applied Sciences, 12(16), 7986.
10. Rahali, A., & Akhloufi, M. A. (2021). Malbert: Malware detection using bidirectional encoder representations from transformers. In 2021 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 3226-3231). IEEE.
11. Rezaabad, A. L., Kumar, S., Vishwanath, S., & Tamir, J. I. (2022). Few-Max: Few-shot domain adaptation for unsupervised contrastive representation learning. arXiv preprint arXiv:2206.10137.
12. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. Computers & security, 86, 147-167.
13. Sharma, A., & Lashkari, A. H. (2025). A survey on encrypted network traffic: A comprehensive survey of identification/classification techniques, challenges, and future directions. Computer Networks, 257, 110984.
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
15. Verizon. (2025). 2025 Data Breach Investigation Report. Verizon. Retrieved July 25, 2025, from https://www.verizon.com/business/resources/reports/dbir
16. Xu, C., Shen, J., & Du, X. (2020). A method of few-shot network intrusion detection based on meta-learning framework. IEEE Transactions on Information Forensics and Security, 15, 3540-3552.
17. Xue, B., Zhao, H., & Yao, W. (2022). Deep transfer learning for IoT intrusion detection. In 2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT) (pp. 88–94). IEEE.