From Autoencoders to Transformers: Deep Learning for Intrusion Detection Systems

Marta Rares
Babeș-Bolyai University
28.10.2025

***Abstract*—Intrusion Detection Systems (IDS) are fundamental tools for securing modern networks. Traditional rule-based systems, while effective against known threats, fail to generalize to novel or evolving attack strategies. This report provides a theoretical overview of deep learning approaches for IDS. We introduce the problem context, review theoretical foundations, summarize prior approaches, and present a framework contrasting autoencoders for anomaly detection with transformer architectures for supervised classification. The report concludes with key insights and outlines directions for further exploration.**

I. Introduction

Cybersecurity remains one of the most pressing challenges in the digital era. As modern society depends on interconnected systems for communication, commerce, and critical infrastructure, the threat of malicious intrusions into computer networks has grown steadily. Traditional security mechanisms such as firewalls and signature-based Intrusion Detection Systems (IDS) are effective against known attacks, but they struggle when attackers employ novel strategies or modify their behavior to evade detection.

Among the most frequent and disruptive threats are:

- **Denial-of-Service (DoS) attacks:** overwhelming a target with illegitimate traffic, exhausting its resources, and rendering it unavailable to legitimate users.
- **PortScan attacks:** reconnaissance attempts where attackers systematically probe network ports to identify open services and potential vulnerabilities for later exploitation.

These attacks highlight the limits of static rules and motivate the need for adaptive learning-based systems. Deep learning provides such adaptability, learning directly from data to capture subtle, non-linear relationships in traffic features.

This report is organized as follows: Section II presents the theoretical background of IDS and machine learning principles. Section III reviews prior approaches in the field. Section IV introduces the theoretical framework of autoencoders and transformers for IDS. Section V concludes with insights and open challenges.

II. Theoretical Background

Intrusion Detection Systems are generally divided into two broad categories. *Signature-based IDS*, exemplified by tools such as Snort and Suricata, operate by scanning network traffic for predefined patterns of malicious activity [3]. These signatures can describe byte sequences, header anomalies, or protocol violations, and they are very effective at detecting known attacks. However, their static nature makes them blind to zero-day exploits or even small variants of existing attacks [2].

In contrast, *anomaly-based IDS* establish a model of normal network behavior and flag any significant deviation as suspicious [7]. This approach has the advantage of detecting previously unseen threats, but it also risks higher false positive rates, since unusual but legitimate traffic can be incorrectly flagged.

When viewed from a machine learning perspective, intrusion detection maps naturally to classification. In its simplest form, the task is binary: distinguish benign traffic from malicious flows. In practice, multi-class setups are common, with systems trained to recognize distinct categories of attacks such as DoS, PortScan, infiltration, or botnet activity. The input is usually tabular flow data, as provided by benchmark datasets like CICIDS2017 or UNSW-NB15, where each flow is described by features such as packet counts, byte rates, and TCP flag statistics [4]. The key challenges for learning-based IDS remain class imbalance (most traffic is benign), concept drift (attack strategies evolve over time), and the need for real-time operation in high-throughput environments.

III. Theoretical Framework

This section contrasts two paradigms in deep learning for IDS: autoencoders for anomaly detection and transformers for supervised classification.

A. Autoencoders for Anomaly Detection

Autoencoders were among the first deep models applied to anomaly detection [8]. They learn to compress input into a latent representation and reconstruct them as closely as possible. When trained on benign traffic, the network captures the manifold of normal behavior. At test time, malicious flows tend to yield

higher reconstruction errors, providing a natural anomaly score. This principle has proved useful for detecting novel attacks without the need for labeled malicious samples. Extensions such as denoising autoencoders and variational autoencoders further improve robustness and probabilistic interpretability [1].

A limitation of auto-encoders in practice is their reliance on high-quality training data. If malicious flows are accidentally included as 'normal', the model may learn to reconstruct them, undermining its effectiveness. Moreover, setting the anomaly threshold is a delicate balance between recall and false positives, especially in enterprise networks with diverse traffic.

B. Transformers for Supervised Intrusion Detection

Transformers represent a more recent paradigm shift. Originally designed for sequence modeling in natural language [6], they have been adapted to tabular data through architectures such as the TabTransformer. The core mechanism, self-attention, allows the model to dynamically weigh relationships between features, capturing global dependencies that older architectures often missed. Applied to IDS, transformers can integrate heterogeneous features — such as packet size distributions, protocol usage, and temporal statistics — into a unified representation. Empirical studies have reported that transformer-based detectors achieve state-of-the-art performance on modern benchmarks [5].

However, transformers come with their own challenges. They are computationally demanding and require large volumes of labeled data to generalize well. Even more importantly, they are not immune to adversarial examples: carefully perturbed network flows can sometimes fool a model into misclassifying malicious activity as benign [7]. This has sparked growing interest in combining transformers with explainability methods such as SHAP or gradient-based attribution to improve trustworthiness in real-world deployments.

C. Comparison

The contrast between the two paradigms illustrates the historical evolution of deep learning in IDS. Autoencoders, rooted in anomaly detection, excel at identifying unusual behaviors without requiring attack labels. Transformers, on the contrary, thrive when large, labeled datasets are available, achieving high precision by directly learning discriminative boundaries. Autoencoders are lightweight and adaptable, while transformers are powerful but resource intensive. Both approaches highlight the central trade-off in IDS research: balancing adaptability to novel attacks with accuracy and robustness in detecting known threats.

IV. Literature Review

The field of IDS has evolved through several stages:

- **Rule-based systems:** early IDS such as Snort relied entirely on hand-crafted signatures.
- **Classical machine learning:** Support Vector Machines (SVM), Decision Trees, and Random Forests were used with feature engineering, achieving moderate success.
- **Deep learning models:** MLPs introduced end-to-end learning of nonlinear boundaries; CNNs captured local patterns in feature vectors; RNNs (LSTMs, GRUs) modeled temporal dependencies in sequential flows.
- **Unsupervised methods:** autoencoders and variational autoencoders applied reconstruction error as anomaly score.
- **Modern approaches:** attention-based architectures (e.g., TabTransformer) capture global dependencies across heterogeneous tabular features.

This literature shows a historical progression from manual feature engineering to fully-trained representations, setting the stage for our framework discussion.

V. Theoretical Framework

This section contrasts two key paradigms in deep learning for IDS: autoencoders for anomaly detection and transformers for supervised classification.

A. Autoencoders for Anomaly Detection

An autoencoder is a neural network that learns to reproduce its input after passing it through a smaller internal representation. It consists of two parts: an encoder, which compresses the input into a latent vector, and a decoder, which tries to reconstruct the original data from this compressed form. The network is trained to minimize the difference between the input x and its reconstruction $\hat{x}$, usually with a squared error loss:

$$L(x, \hat{x}) = \|x - \hat{x}\|^2.$$

In intrusion detection, autoencoders are trained only on benign traffic. Because they learn the structure of normal flows, they can reconstruct them with little error. When the model encounters malicious traffic such as DoS or PortScan attempts, the reconstruction error tends to be much larger, since these patterns differ from what the model has seen. This error can therefore be used as an anomaly score: small errors indicate normal traffic, while large errors signal a possible attack.

The main strength of this approach is that it does not require explicit attack labels, making it useful when malicious examples are rare or unknown. It can also generalize to new types of intrusions by simply flagging them as deviations. However, autoencoders are sensitive to the quality of training data—if attacks are accidentally

included as "normal," the model may fail to recognize them. They also depend on choosing a good threshold for the anomaly score, and in practice their precision is often lower than that of supervised methods trained directly on labeled attacks.

B. Transformers for Supervised Intrusion Detection

Transformers are a modern neural network architecture originally developed for natural language processing. Their key idea is the mechanism of *self-attention*, which allows the model to evaluate how strongly different parts of the input relate to each other. Formally, given input representations, the attention mechanism computes

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

where Q, K, and V are learned projections of the input. This operation lets the model focus on the most relevant features when making a decision.

In the context of intrusion detection, network flows can be represented as tabular data with features such as packet counts, byte rates, or protocol flags. Transformers, especially in variants like the TabTransformer, embed these features and then apply layers of self-attention to capture both local and global relationships. Unlike older models that only detect short-range patterns, transformers can dynamically decide which features influence each other, making them well suited for complex traffic data.

The main advantage of transformers is their ability to capture dependencies across all features at once, often achieving state-of-the-art performance on IDS benchmarks. They can also provide interpretability by highlighting which features the model focused on. Their limitations, however, include higher computational cost and the need for large, high-quality datasets to train effectively. Despite these challenges, transformers represent the current frontier of deep learning approaches to IDS and illustrate how advances from other domains, such as language and vision, can be adapted to cybersecurity problems.

C. Comparison

- **Learning paradigm:** Autoencoders are unsupervised (reconstruction-based), transformers are supervised (label-based).
- **Generalization:** Autoencoders adapt to unseen threats, transformers excel on labeled attack classes.
- **Interpretability:** Autoencoders provide anomaly scores, transformers offer attention maps.
- **Scalability:** Autoencoders are lightweight, transformers require more computation.

This comparison illustrates the evolution from early anomaly-based deep learning to modern attention-driven classification.

VI. Conclusion

Deep learning has transformed intrusion detection, moving from static rules to adaptive models capable of capturing complex traffic patterns. Autoencoders provided an early unsupervised method to detect anomalies through reconstruction error. Transformers represent the modern frontier, leveraging attention to model global dependencies for supervised classification.

Both paradigms highlight complementary strengths: autoencoders for adaptability to novel attacks, transformers for high-accuracy detection with rich labeled data. Future research must balance these strengths with interpretability, computational efficiency, and robustness to adversarial manipulation.

References

[1] A. Caville et al. Anomal-e: A self-supervised network intrusion detection system based on graph neural networks. *arXiv preprint arXiv:2207.06819*, 2022.

[2] J. Kimanzi et al. Deep learning algorithms used in intrusion detection systems—a review. *arXiv preprint arXiv:2402.17020*, 2024.

[3] Martin Roesch. Snort - lightweight intrusion detection for networks. In *Proceedings of the 13th USENIX conference on System administration*, pages 229–238, 1999.

[4] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani. Cicids2017: A realistic cyber intrusion detection dataset. https://www.unb.ca/cic/datasets/ids-2017.html, 2018.

[5] Author Springer. Network intrusion detection using feature fusion with deep learning. *Journal of Big Data*, 2023.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[7] X. Zhang et al. Deep adversarial learning in intrusion detection: A data augmentation enhanced framework. *arXiv preprint arXiv:1901.07949*, 2019.

[8] J. Zhou et al. Gee: A gradient-based explainable variational autoencoder for network anomaly detection. *arXiv preprint arXiv:1903.06661*, 2019.