Literature Review: Deep Learning-Based Autonomous Anomaly Detection in SDN-IoT Networks

Maxime BOSSANT

May 2025

Introduction

The integration of Software-Defined Networking (SDN) and the Internet of Things (IoT) has revolutionized network management by enabling centralized control and scalability. However, this convergence introduces significant security challenges, including a new attack surface (SDN controller) and diverse threat vectors such as Distributed Denial-of-Service (DDoS) and malware attacks. Traditional anomaly detection systems struggle to adapt to the dynamic nature of SDN-IoT networks, necessitating Autonomous Anomaly Detection (AAD) systems powered by Deep Learning (DL). These systems leverage models like CNNs, LSTMs, and autoencoders to detect anomalies in real time. However, DL models are vulnerable to adversarial attacks, particularly data poisoning, which manipulate training data to degrade model performance. This review synthesizes current research on adversarial threats to DL-based AAD systems, detection strategies, and future directions.

1 Threats in SDN-IoT Networks

1.1 Adversarial Attacks

Khazane et al. [1] present a detailed taxonomy of adversarial attacks on IoT networks, distinguishing them by attacker knowledge, goal, capability, and strategy. They argue that previous taxonomies were too narrow, ignoring IoT-specific constraints like non-differentiable features or limited observability. They describe several well-known adversarial methods applied in network environments, such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Zeroth-Order Optimization (ZOO). These techniques illustrate how adversaries can craft input perturbations to bypass detection models even in black-box settings, which are common in real-world IoT deployments.

1.2 Data Poisoning Attacks

In SDN-IoT systems, continual learning enables anomaly detection models to adapt to evolving traffic patterns. However, this adaptability exposes them to data poisoning attacks, where adversaries inject malicious or mislabeled samples into training data to degrade model performance. Yasarathna and Le-Khac [2] highlight how such poisoning can corrupt DL models by introducing fake or mislabeled data in the learned decision boundaries during retraining phases.

Label flipping and outlier injection are commonly observed techniques. These attacks aim to confuse the model without causing significant anomalies in the input distribution, making them difficult to detect in real time [3].

1.3 Other Attack Vectors

In addition to poisoning, SDN-IoT systems face Distributed Denial of Service (DDoS) attacks, spoofing, brute-force, Mirai and evasion-based threats (Neto et al. [4]). Novaes et al. [5] discuss how the SDN controller's centralized nature makes it a DDoS target, especially when flow tables are saturated or controller communication is disrupted.

1.4 Data-Free Attacks

Data-free model extraction attacks represent a growing threat in scenarios where adversaries lack access to the original training data but can still query the deployed model. As shown by Truong et al. [6], attackers can use synthetic data and gradient-based inversion techniques to reconstruct the structure and parameters of a target model. This is especially relevant in SDN-IoT contexts where anomaly detection models may be deployed using APIs.

Lin et al. [7] extend this approach to query-limited environments, showing that even with restricted access, an attacker can extract a high-fidelity replica of a model. These techniques can lead to further evasion or poisoning attacks, by using the extracted model to test adversarial inputs offline. These techniques, though mainly demonstrated in computer vision, do not rely on specific model architectures, making them applicable to DL-based NIDS in SDN-IoT systems.

2 Defense Mechanisms and Detection Strategies

2.1 Adversarial Attacks in NIDS

Adversarial attacks target the weakness of DL-based NIDS by creating light input perturbations that lead to misclassification. Khazane et al. [1] propose a comprehensive taxonomy of adversarial threats in IoT systems, and show that different attacker strategies—such as transfer-based, query-based, or logic corruption attacks—can bypass detection by exploiting limitations in both the model and the training data. To address these threats, recent defense strategies include adversarial training, gradient masking, feature squeezing, and input reconstruction. However, as highlighted by the authors, these methods often suffer from high computational costs, limited generalization, and reduced accuracy on clean data, making them difficult to deploy in real-time SDN-IoT environments. Popular models like LSTM and CNN have been shown to be vulnerable when exposed to carefully crafted adversarial samples. Several studies also emphasize that DL models trained on static datasets often fail to generalize under adversarial pressure.

2.2 Autonomous and Real-Time Anomaly Detection (AAD)

Autonomous Anomaly Detection (AAD) refers to intrusion detection systems that operate without requiring constant human supervision, retraining, or manual labeling. These systems are particularly valuable in SDN-IoT contexts, where traffic patterns evolve rapidly and labeled data is rare making these networks very vulnerable to various cyber threats.

Yasarathna and Le-Khac [2] show that AAD systems are designed to dynamically adapt to evolving threats, which is crucial in SDN-IoT networks where traditional static detection approaches are insufficient. They argue that such autonomous systems can improve resilience as they can detect novel attacks in real time and adjust to changes in traffic behavior (i.e., concept drift). They also highlight the importance of lightweight and explainable models to ensure deployment feasibility in constrained environments like IoT devices. However, DL-based AAD systems are susceptible to adversarial attacks, particularly in continual learning settings, where models must adapt to evolving threats and changing network conditions, that's why some strategies must be implemented.

While traditional ML algorithms such as Random Forest and SVM have shown good performance in SDN detection tasks ([8]), they often lack the capacity to autonomously adapt to evolving patterns or detect unknown attacks. In contrast, DL models like CNN and LSTM are more suitable for building scalable and adaptive systems.

Chaganti et al. [9] propose an LSTM-based intrusion detection system for SDN-enabled IoT environments. Their architecture uses four stacked LSTM layers to classify network flows as normal or one of five attack types (e.g., DoS, DDoS, port scanning). The model, trained on traffic from simulated IoT devices via OpenFlow, achieves 97.1% accuracy. Compared to traditional models like SVM and CNN, their LSTM design offers superior accuracy without requiring manual feature selection. The authors also validate the model on a second dataset, confirming its generalizability across different SDN-IoT settings. Although training time and memory use are significant, the approach shows strong promise for scalable and autonomous anomaly detection.

2.3 Mitigating Attacks

Given the vulnerabilities described above, several countermeasures have been proposed to enhance the robustness of DL-based NIDS in SDN-IoT environments. To address the threat of poisoning attacks, Yasarathna and Le-Khac [2] propose an enhanced cross-validation strategy that tracks variations in model accuracy across different validation windows. If the model's accuracy changes too much during training, it might mean the data was poisoned. This method compares different validation sets over time and looks for suspicious shifts. However, this system could be less efficient against slow attacks or incremental poisoning attacks that replicate the distribution of the original data.

Novaes et al. [5] propose a defense system for SDN networks using adversarial training based on a GAN . Their approach generates artificial attack samples to train the model in a more robust way, making it less sensitive to evasion attempts. While the method is effective, using GANs adds extra layers, which may increase training time and memory usage in practice.

In addition to DL-based techniques, non-DL statistical methods such as entropy-based detection combined with Z-Test, as implemented in the SDNTruth system ([10]), demonstrate good performance without increasing time consumption.

3 Deep Learning Models for Anomaly Detection

3.1 Core Models: CNN, LSTM, AE-LSTM

CNNs, LSTMs, and hybrid AE-LSTM models have demonstrated strong results in SDN-IoT anomaly detection tasks. Ruffo et al. [11] provide an extensive taxonomy showing CNNs are well-suited for identifying features in traffic data, particularly packet flows and port activity matrices.

LSTM models, by contrast, are designed for temporal analysis and have shown higher efficiency in detecting sequential or long-term threats [2]. AE-LSTM models combine the autoencoder's reconstruction ability with LSTM's temporal memory, allowing them to detect deviations by learning low-dimensional representations of normal traffic patterns and identifying reconstruction failures as anomalies.

3.2 Signature vs Anomaly-Based NIDS

Traditional NIDS (Network Intrusion Detection System) often rely on signature-based methods, which are efficient against known threats with very few False Positive (good precision) but inadequate for detecting unknown attacks. Ruffo et al. [11] and M. S. Elsayed and Jurcut [12] emphasize that anomaly-based detection, particularly with DL, enables systems to flag deviations from expected behavior without predefined rules.

Anomaly-based systems try to learn what "normal" traffic looks like and raise alerts when they see something different. However, it can be prone to false positives, especially in IoT networks where traffic behavior may change rapidly. Novaes et al. [5] describe both signature-based and anomaly-based NIDS approaches. They explain that signature-based detection is effective for known threats, while anomaly-based detection is better suited for identifying unknown or zero-day attacks. Combining both strategies could offer a more balanced solution.

3.3 Dataset and Generalization Challenges

One of the main problems with DL-based anomaly detection is the lack of realistic datasets. Models trained on static datasets like NSL-KDD or CICIDS2017 often struggle to generalize to real-world IoT environments. As noted by Ruffo et al. [11] and M. S. Elsayed and Jurcut

[12], most publicly available datasets, such as CICIDS2017 or NSL-KDD, often lack up-to-date attack types and do not fully reflect the complexity of real SDN-IoT environments.

New datasets are being proposed to bridge this gap. One example is the InSDN dataset presented by Elsayed, Le-Khac, and Jurcut [13], which is designed specifically for SDN environments. It includes modern attacks such as botnet, port scanning, and various flooding techniques, and reflects realistic SDN traffic patterns. By incorporating a range of benign and malicious flows in an OpenFlow-enabled testbed, InSDN improves diversity and can help deep learning models generalize better to real-world SDN threats.

Moreover, supervised models require labeled data, which is expensive to obtain. Ruffo et al. [11] point out that in their survey of 105 research works on deep learning-based intrusion detection for SDN, 96 relied on labeled datasets. This heavy dependence on supervised learning shows an important limitation in real-world situations, where labeled data is often rare and expensive to obtain.

4 Conclusion and Research Opportunities

Deep learning has become a key tool for detecting anomalies in SDN-IoT networks, especially with models like CNN, LSTM, and AE-LSTM. These models can detect many types of attacks with good accuracy, and they adapt better than traditional rule-based systems. However, they are not perfect, DL-based systems are vulnerable to adversarial threats like data poisoning, and they require large amounts of labeled data, which is often hard to get.

To improve current systems, researchers are exploring new directions. For example, Ruffo et al. [11] point out that unsupervised learning is still underused. However, it has strong potential to create autonomous detection systems, especially when large amounts of unlabeled data are available. This approach could reduce dependence on labeled datasets and make anomaly detection systems more flexible in real-world conditions.

Another promising area is reinforcement learning, which can help networks learn how to react to new or evolving threats. It could be used to dynamically adapt security rules dynamically, without needing manual updates.

In short, future work should focus on making AAD systems more autonomous, scalable, and adaptable, while reducing their reliance on labeled datasets.

References

[1] H. Khazane et al. “Holistic Review of Machine Learning Adversarial Attacks in IoT Networks”. In: *Future Internet* 16.1 (2024), p. 32. DOI: 10.3390/fi16010032.

[2] T.L. Yasarathna and Le-Khac. “Advancing Security in SDN-IoT Networks: DL-Based Autonomous Anomaly Detection with Enhanced Cross-Validation”. In: *Springer LNNS* (2024). DOI: 10.1007/978-3-031-74127-2_41.

[3] T.L. Yasarathna et al. “Cross-Validation for Detecting Label Poisoning Attacks: A Study on Random Forest Algorithm”. In: *IFIP SEC* (2024). DOI: 10.1007/978-3-031-65175-5_32.

[4] E. C. P. Neto et al. “CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment”. In: *Sensors* 23.13 (2023), p. 5941. DOI: 10.3390/s23135941.

[5] M.P. Novaes et al. “Adversarial Deep Learning Approach Detection and Defense Against DDoS Attacks in SDN Environments”. In: *Future Generation Computer Systems* (2021). DOI: 10.1016/j.future.2021.06.047.

[6] Jean-Baptiste Truong et al. “Data-Free Model Extraction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4771–4780. DOI: 10.48550/arXiv.2011.14779.

[7] Zijun Lin et al. “QUDA: Query-Limited Data-Free Model Extraction”. In: *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security (ASIA CCS ’23)*. ACM, 2023, pp. 913–924. DOI: 10.1145/3579856.3590336.

[8] R. Santos et al. “Machine Learning Algorithms to Detect DDoS Attacks in SDN”. In: *Concurrency and Computation: Practice and Experience* (2019). DOI: 10.1002/cpe.5402.

[9] R. Chaganti et al. “Deep Learning Approach for SDN-Enabled Intrusion Detection System in IoT Networks”. In: *Information* 14.1 (2023), p. 41. DOI: 10.3390/info14010041.

[10] T. Linhares et al. “SDNTruth: Innovative DDoS Detection Scheme for Software-Defined Networks (SDN)”. In: *Journal of Network and Systems Management* 31 (2023). DOI: 10.1007/s10922-023-09741-4.

[11] V. G. da S. Ruffo et al. “Anomaly and Intrusion Detection Using Deep Learning for Software-Defined Networks: A Survey”. In: *Expert Systems with Applications* (2024). DOI: 10.1016/j.eswa.2024.124982.

[12] S. Dev M. S. Elsayed N-A. Le-Khac and A. D. Jurcut. “Machine-Learning Techniques for Detecting Attacks in SDN”. In: *IEEE ICCSNT* (2019). DOI: 10.1109/ICCSNT47585.2019.8962519.

[13] Mahmoud Said Elsayed, Nhien-An Le-Khac, and Anca D. Jurcut. “InSDN: A Novel SDN Intrusion Dataset”. In: *IEEE Access* 8 (2020), pp. 165263–165284. DOI: 10.1109/ACCESS.2020.3022633.