When Explanations Help Attackers: XAI-Guided Adversarial Attacks to Network Intrusion Detection Systems

Gabriele Mangiacapre, Alfredo Nascita, Francesco Cerasuolo, Antonio Montieri, Antonio Pescapè
University of Napoli Federico II
ga.mangiacapre@studenti.unina.it, {alfredo.nascita, francesco.cerasuolo, antonio.montieri, pescape}@unina.it

***Abstract*—Network attacks have grown in recent years, raising significant security concerns in network infrastructures. Deep Learning-based Network Intrusion Detection Systems (NIDSs) are widely adopted to monitor network traffic and identify malicious activities, thanks to their high detection performance. However, their decision processes are often opaque, limiting trustworthiness and interpretability. To overcome this limitation, eXplainable Artificial Intelligence (XAI) techniques are increasingly used to provide insights into NIDS behavior. Although XAI enables analysts to better understand and validate security decisions, it also introduces new risks: if explanations become accessible to attackers, they could be exploited to devise targeted evasion strategies. In this work, we investigate whether XAI explanations can be leveraged to evade detection in NIDSs. Using SHAP to interpret our model, we identify high-impact features to inform adversarial modifications of malicious flows. Specifically, we apply feature-level perturbations derived from SHAP attributions to assess the NIDS' vulnerability to explanation-driven attacks. Experimental results show that XAI-guided perturbations achieve a 94% evasion rate by modifying a single high-impact feature. In contrast, random perturbations reach only 54%, highlighting the effectiveness of explanation-driven attacks.**

***Index Terms*—Adversarial Attacks, eXplainable AI, Network Intrusion Detection Systems, Software-Defined Networking, Network Attack Traffic**

I. Introduction

In recent years, cyber-attacks have surged dramatically, with roughly 600 million daily incidents [1]. This alarming trend highlights the urgent need for service providers to proactively defend the integrity, confidentiality, and availability of digital infrastructures. In this context, Network Intrusion Detection Systems (NIDSs) play a pivotal role in protecting modern networks. Concurrently, the widespread adoption of Software Defined Networking (SDN) has transformed network architectures by introducing centralized management and advanced programmability [2]. While highly beneficial for operational flexibility, these SDN characteristics also amplify the necessity for robust security mechanisms capable of detecting and mitigating attacks targeting network infrastructures.

In the last decades, researchers have consistently used Deep Learning (DL) to design NIDSs [3, 4]. However, the inherent black-box nature of these models limits their adoption in critical domains where transparency, accountability, and actionable insights are vital [5]. To address this limitation, eXplainable Artificial Intelligence (XAI) techniques are increasingly utilized to clarify the decision-making process of DL-based NIDSs, for example, by quantifying the contribution of individual input features to the model's predictions [6]. Although such knowledge enhances the interpretability of NIDSs, it may simultaneously enable adversaries to craft targeted attacks against them. In particular, by identifying the features that most strongly influence attack detection, XAI may reveal the input dimensions most vulnerable to manipulation, enabling attackers to strategically modify network traffic to bypass the NIDSs more effectively than through unguided, random perturbations.

In this work, we leverage XAI techniques to analyze a NIDS model, identifying the features that most heavily influence its decisions. Then, we analyze to what extent these insights could be exploited by an adversary to improve evasion effectiveness. Notably, we evaluate adversarial perturbations targeting the NIDS's most critical features to determine if such XAI-guided manipulations significantly increase the evasion success rate.

The remainder of this paper is organized as follows. Section II reviews related work and positions our contribution. Section III presents the proposed methodology. Section IV describes the dataset and the experimental setup, while Section V presents the evaluation results. Finally, Section VI summarizes the main findings and discusses future research directions.

II. Related Works

The adoption of DL for NIDSs has led to significant performance improvements, but has also introduced a critical lack of interpretability. To mitigate this limitation, XAI has emerged as a solution to provide actionable insights into model decisions. This interpretability requirement is particularly crucial in SDN environments, which are characterized by dynamic traffic patterns and programmable infrastructures, for understanding and securing model behavior against evolving attack conditions. To provide a comprehensive overview of the current landscape, Table I summarizes the most relevant studies dealing with XAI and adversarial attacks in the context of NIDSs, highlighting their key characteristics.

In the current literature, several works exploit XAI to uncover the decision-making process of Machine Learning

ISBN 978-3-903176-82-9 © 2026 IFIP

TABLE I: Related studies dealing with XAI and/or adversarial attacks for NIDS. The last row summarizes the present work.

Paper	Year	Dataset	DL	Architecture	Input	XAI Method	Adv. Attacks	Gen.
Marino et al. [7]	2018	NSL-KDD	✓	MLP	S	Adversarial	◐	✗
Tcydenova et al. [8]	2021	NSL-KDD	✗	SVM	S	LIME	◐	✗
Barnard et al. [9]	2022	NSL-KDD	✗	XGBoost	S	SHAP	○	✗
Tserenkhuu et al. [10]	2025	InSDN, X-IIoTID	✓	CNN, MLP, LSTM	S	SHAP, LIME	○	✗
Cherian [11]	2025	SDN dataset	✓	1D-CNN, LSTM	S	SHAP	○	✗
Okada et al. [12]	2025	CIC-IDS2017, TON_IoT	✓	MLP	S	SHAP	●	✗
Pham-Thai et al. [13]	2026	InSDN,	✓	CNN	S	SHAP, LIME	○	✗
Chang Chung and Han [14]	2026	X-IIoTID	✗	MLP	S	SHAP	●	✗
Our work	2026	InSDN	✓	2D-CNN	HF	SHAP	●	✓

Input: Statistical (S), Header Fields (HF); **Adversarial Attacks:** ○ (absent), ◐ (used for explanations), ● (XAI-guided); **Generalization (Gen.)**.

(ML) models in SDN environments. For instance, Tserenkhuu et al. [10] propose an SDN-based NIDS framework for IoT networks that integrates DL with XAI-driven feature selection. Their approach uses domain-constrained features identified via SHapley Additive exPlanations (SHAP) [15] and statistical feature importance, thereby enhancing interpretability while preserving high detection rates and reducing computational overhead. Similarly, Cherian [11] introduce a NIDS for SDN that combines convolutional and recurrent neural networks with XAI. Their framework employs data preprocessing to boost model transparency and ensure effective threat detection. Moreover, Pham-Thai et al. [13] leverage both classical ML and DL classifiers alongside XAI to reduce feature set dimensions and computational costs, enabling more efficient deployment. Beyond the SDN domain, Barnard et al. [9] present a two-stage NIDS where SHAP explanations derived from an XGBoost model are used to train an autoencoder capable of distinguishing known from unknown attacks, proving that explainability can support both model transparency and robustness.

Alongside XAI, prior work has shown that adversarial ML can be leveraged to enhance the understanding of NIDSs. In this context, Marino et al. [7] propose an adversarial-based explanation framework for intrusion detection systems, generating minimal perturbations of input features to correct misclassified samples and highlight the most influential variables behind incorrect decisions. Furthermore, Tcydenova et al. [8] present a detection framework for adversarial attacks against ML-based NIDSs. Their XAI-assisted approach uses Local Interpretable Model-agnostic Explanations (LIME) to profile normal data during an initialization phase, and then analyzes new inputs during the detection phase to identify adversarial manipulations.

Similar to the present research, Okada et al. [12] and Chang Chung and Han [14] combine XAI and adversarial attacks against NIDSs from an attacker's perspective. The former exploits explanations of DL-based models to identify critical features, crafting realistic perturbations in actual network traffic spaces without compromising the inherent maliciousness of the attacks. The latter combines explainability with gradient-based optimization, ranking features through SHAP and applying a masked projected gradient descent procedure to perturb the most influential ones.

Positioning. As summarized in Table I, unlike most prior works, our study does not employ XAI merely to interpret model behavior [9–11, 13], but rather exploits it to actively shape and direct the attack strategy. Furthermore, we consider an *early-detection* strategy, where malicious traffic is identified at the initial stages of the communication. In particular, we use packet header fields as inputs rather than *post-mortem* statistics computed over the entire traffic flow, enabling earlier detection while also reducing computational overhead. Differently from existing studies on adversarial attacks that employ perturbations to identify and correct model misclassifications [7] or to detect adversarial manipulations during inference [8], our approach leverages XAI explanations to guide targeted packet-level perturbations aimed at evading the NIDS. Although Okada et al. [12] and Chang Chung and Han [14] adopt a similar offensive perspective, the former is restricted to a limited set of attack strategies, whereas the latter employs per-sample explanations to perturb individual instances using gradient-based techniques that often yield unrealistic perturbations and do not account for generalization across distinct samples. Conversely, our work broadens this scope by evaluating a wider range of attacks within the highly dynamic SDN environment and by considering realistic perturbations also on a disjoint set of samples, distinct from those used to compute explanations, thereby explicitly assessing generalization.

III. Methodology

In this section, we detail the proposed methodology for executing XAI-guided adversarial attacks. As illustrated in Fig. 1, the overall procedure consists of two main phases: an *explanation phase*, which interprets the NIDS' decisions to identify the most critical input features, and a subsequent *attack phase*, where these insights are actively leveraged to execute informed adversarial perturbations. Below, we provide a brief background on adversarial attacks, followed by a detailed description of both phases and a final overview of the comparison baseline utilized.

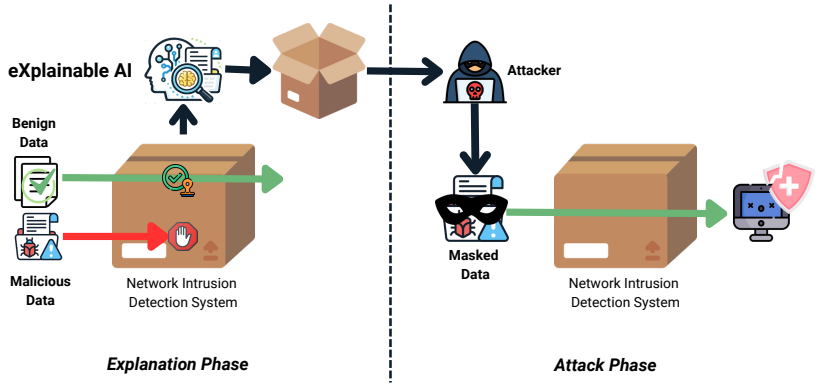

Fig. 1: Overall procedure of XAI-guided attacks. The **explanation phase** identifies the most important features for the NIDS, which are then targeted in the **attack phase** to generate adversarial perturbations based on the SHAP explanations.

Adversarial Attacks. Adversarial evasion attacks against intrusion detection systems aim to identify minimal and carefully constrained perturbations that cause a classifier to alter its prediction while preserving the semantic plausibility of the input. Formally, given an input sample $x \in \mathbb{R}^M$ correctly classified by a model $f(\cdot)$, an evasion attack searches for a perturbed sample $x^* = x + \delta$ that flips the predicted class, while the perturbation magnitude is minimized and domain constraints are satisfied. This process can be formulated as a constrained optimization problem:

$$\min_{\delta} \|\delta\| \quad \text{s.t.} \quad f(x+\delta) \neq f(x), \;\; x+\delta \in \mathcal{C}, \qquad (1)$$

where δ represents the perturbation vector, $\|\delta\|$ is a norm quantifying the perturbation's magnitude, and $\mathcal{C}$ denotes the set of admissible inputs defined by feature validity and protocol-consistency constraints [16]. For network traffic data, such constraints include value ranges, integer requirements, and structural coherence of packet-level fields.

In an evasion scenario, attackers aim to force the classifier to misclassify malicious traffic as benign. Consequently, while the attack is targeted in its desired outcome, the adversary's modifications are strictly bounded: any altered feature must result in a semantically valid and operational network packet. By identifying the minimal perturbation required for evasion, we can directly measure the vulnerability of the model's decision boundaries.

Explainable Phase. To improve the transparency of the proposed DL-based NIDS, we complement predictive performance with feature-attribution explanations, obtained with the SHAP framework. To analyze feature contributions to the model's predictions, we rely on `DeepSHAP`, a SHAP variant tailored to DL. `DeepSHAP` assigns an attribution score to each input feature, quantifying its positive or negative contribution to the output. Let ϕ_i denote the SHAP value for the i-th feature. To facilitate comparison across samples and classes, we normalize the SHAP values by dividing each ϕ_i by the sum of all feature attributions. This process yields a relative *importance* measure that is robust to scale differences in the output values. Lastly, to move from per-sample explanations to a global interpretation of the model behavior, we aggregate the normalized SHAP values across multiple samples. Specifically, we compute the median of the normalized attributions $\tilde{\phi}_i^{(k)}$ for each feature i over all the considered samples $k = 1, \ldots, N$, thus obtaining a robust estimate of each feature's contribution.

XAI-Guided Adversarial Attacks. The attack phase targets the features that the XAI analysis has identified as the most influential. The goal is to determine the minimal perturbations required to flip a correct malicious classification into a benign one by modifying only a small subset of high-impact features, while preserving the overall plausibility. This inversion serves a twofold purpose: (i) assessing the robustness of the model's decision boundaries, and (ii) validating the actionable impact of the most influential features. Operationally, minimal perturbations are sequentially applied to the top-3 features of each sample by systematically decreasing or increasing their values in integer steps up to toward their respective minimum or maximum admissible bounds, ensuring that the resulting modifications remain protocol-consistent. For each feature, the smallest modification that results in a change of the NIDS's prediction is selected. Once a successful perturbation is found for a sample, the process is halted, so that only one feature is modified per sample (or none, if no single-feature perturbation is effective). This procedure isolates the effect of individual features while maintaining semantic plausibility. By avoiding combinatorial feature interactions, our approach focuses exclusively on minimal, realistic single-feature manipulations that alter the model's decision.

Random Attacks Baseline. As a counterpart to our XAI-guided adversarial procedure, we conduct a parallel experiment where features are selected and modified *randomly*. This unguided approach serves as a baseline, allowing us to evaluate the effectiveness of random perturbations and directly compare them against our targeted strategy. To mirror the latter, up to three randomly chosen features are perturbed independently, and each can be modified at most once per sample. As a result, each sample undergoes up to three perturbation attempts—one for each selected feature—yielding a maximum of $N_{biflows} \times 3$ total attempts across the experiment, where $N_{biflows}$ denotes

TABLE II: Attack categories included in the InSDN traffic dataset.

Attack Type	Description
Web Attack (WA)	Attacks targeting web applications, such as Cross-Site Scripting (XSS) and SQL Injection, aiming to compromise data integrity or confidentiality.
DDoS	Distributed attacks designed to exhaust network or server resources through coordinated TCP, UDP, or ICMP floods originating from multiple hosts.
DoS	Single-source attacks generating excessive traffic to specific services to disrupt availability.
Brute Force Attack (BFA)	Brute-force or dictionary attacks attempting to obtain user credentials via repeated, systematic login attempts.
Probe	Network reconnaissance activities, including port scanning, host discovery, and service fingerprinting, to identify exploitable vulnerabilities.

the total number of evaluated biflows (i.e., samples). Consistent with the XAI-guided procedure, we manipulate only one feature at a time, stopping the process as soon as a modification successfully changes the NIDS's prediction.

IV. Experimental Setup

This section describes the experimental setup. We introduce the dataset, which comprises diverse network attacks targeting SDN, followed by an overview of the adopted DL-model architecture and training strategy. Lastly, we outline the evaluation settings and metrics applied to both the XAI-guided adversarial attacks and the random perturbation baseline.

InSDN Intrusion Dataset. We leverage the publicly available *InSDN* dataset [17], which contains network traffic captured in a controlled SDN testbed deployed in a virtualized environment. The testbed includes a combination of emulated and virtualized devices: Mininet hosts emulate end-user machines, Open vSwitch instances act as software switches, and real SDN controllers (RYU, POX) manage the network. The raw traffic is segmented into biflows[1]. For each packet, we extract the following features: Packet Length (`PL`); Inter-Arrival Time (`IAT`), defined as the time difference with respect to the previous packet; packet direction (`DIR`); TCP Window Size (`WIN`), set to 0 for UDP packets; Time-To-Live (`TTL`), representing the remaining hop count before packet discard; and TCP Flags (`FLG`), encoded as an integer value. To enable *early network attack detection*, we extract these features from the first 10 packets of each biflow, applying zero-padding to shorter biflows to ensure fixed-length input sequences.

The attack classes present in the InSDN dataset are summarized in Tab. II. In our experiments, all attack categories are grouped into a single *malicious* traffic class to perform binary classification.

Model Architecture and Training Setup. To design our data-driven NIDS, we employ a 2D-CNN architecture [18]. In recent literature, this architecture has consistently achieved strong performance on intrusion detection tasks [19]. The 2D-CNN consists of two convolutional layers, each followed by pooling and normalization operations, which extract hierarchical features from the input. Then, the output is flattened and fed into a fully connected layer that produces a probability distribution over the target classes. Overall, the model includes $\approx$ 540k trainable parameters. The input is structured as an $N_{packets} \times N_{features}$ matrix, with $N_{packets} = 10$ per biflow and $N_{features} = 6$ per packet.

[1]A biflow (i.e., a *bidirectional* flow) consists of packets identified by a quintuple (*source IP*, *source port*, *destination IP*, *destination port*, and *transport protocol*), where the roles of *source* and *destination* are interchangeable.

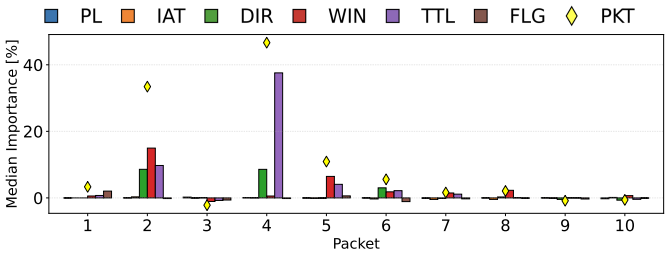

Fig. 2: Median Feature Importance for the first 100 malicious samples.

For our experiments, the 2D-CNN is trained for a binary classification task (viz. benign vs. malicious traffic). The training process spans a maximum of 200 epochs, incorporating an early stopping mechanism with a patience of 20 epochs to prevent overfitting. Optimization starts with a learning rate of 0.1, which is progressively reduced by a factor of 3, down to a minimum threshold of 10^{-7}. Model evaluation follows a standard hold-out validation strategy, splitting the dataset into 70% for training and 30% for testing.

Evaluation Settings and Metrics. To evaluate the success rate of the XAI-guided evasion procedure, we assess the attack under two structured settings: (i) a *self-consistent setting*, where the 100 malicious samples targeted for perturbation are the same as those used for the SHAP value computation; and (ii) a *generalization setting*, which exploits a disjoint set of 100 malicious samples not used during the XAI analysis.

Classification performance is measured using the per-class *F1-score*, defined as the harmonic mean of precision and recall. Additionally, we assess the model's vulnerability to adversarial feature perturbations using the *Evasion Rate*, defined as $N_{success}/N_{samples}$. Here, $N_{success}$ represents the number of samples successfully misclassified after the perturbation, and $N_{samples}$ is the total number of evaluated samples. This metric quantifies the proportion of samples for which modifying one or more input features results in a change in the predicted class, indicating a successful evasion.

V. Experimental Results

In this section, we present the experimental results. First, we discuss the NIDS performance and analyze its decision-making process using SHAP-based feature attribution (Sec. V-A). Then, we evaluate the effectiveness of our XAI-guided evasion strategy (Sec. V-B) and compare it against the random attacks baseline (Sec. V-C).

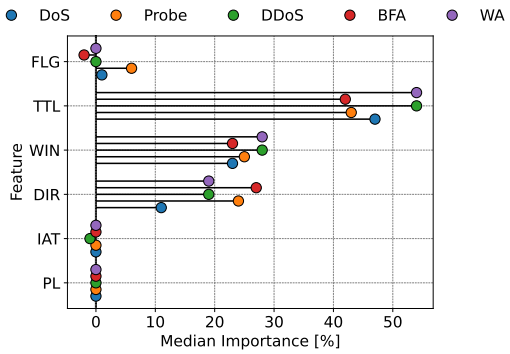

Fig. 3: Per-feature importance computed across 100 malicious biflows and separately for each attack type.

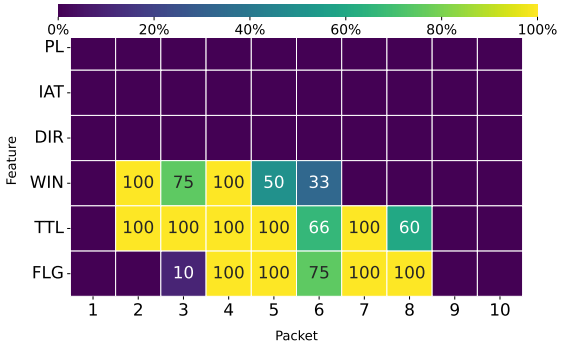

Fig. 4: Heatmap of the Evasion Rate [%] across the Feature–Packet space for random perturbations applied to 100 malicious samples.

A. NIDS Performance and SHAP-based Explanations

First, we assess performance of NIDS, using *F1*, *Precision*, and *Recall* on both classes, achieving 99.96%, 99.97%, and 99.94%, respectively. This indicates that NIDS is able to obtain strong discrimination between classes with minimal errors. To investigate why a biflow is correctly classified as malicious, we analyze SHAP values aggregated across a subset of 100 correctly classified malicious samples. Figure 2 illustrates the *median importance* for the six features considered (`PL`, `IAT`, `DIR`, `WIN`, `TTL`, `FLG`) across the first 10 packets, along with the aggregated importance of each packet (`PKT`). Results reveal a clear pattern: *the model heavily relies on the first few packets to make its decisions*. In particular, the fourth packet is the most informative, with `TTL`, `DIR`, and `WIN` positively contributing to the NIDS decision. Among these, the `TTL` has the highest median importance, followed by the `WIN` and `DIR` of the second packet. Conversely, from the seventh packet onward, feature importance becomes negligible, indicating that the model has learned to extract sufficient information from the initial portion of the biflow. This behavior highlights the critical role that early traffic dynamics play in distinguishing between benign and malicious biflows.

To further investigate the decision-making process of the NIDS, we compute per-feature SHAP attributions for each attack category. To this end, we select 100 representative biflows per attack, ensuring a fair and balanced comparison across categories. This per-attack analysis confirms that the NIDS primarily relies on the initial packets of a biflow to make its decisions, with the first four packets being particularly important. To examine these patterns in more detail, Fig. 3 illustrates the aggregated importance of each feature across all 10 packets. This visualization reveals consistent behaviors regardless of the attack type. Specifically, `PL` and `IAT` consistently show near-zero attributions, while `FLG` exhibits only marginal importance, though slightly more pronounced for `Probe` attacks ($\approx$6%). Overall, `TTL` emerges as the most critical feature for all attack types, with $>40\%$ importance, followed by `WIN` (23–28%) and `DIR` (11–27%).

B. SHAP-Guided Attacks to NIDS

To evaluate the effectiveness of our proposed XAI-guided strategy, we adopt a targeted perturbation approach driven by feature importance in the *self-consistent setting*. Specifically, SHAP values are first utilized to rank the packet-level features based on their contribution to the model's prediction. Following our methodology, we iteratively apply minimal perturbations to the SHAP top-ranked features, stopping the process as soon as the NIDS's prediction flips from malicious to benign. This procedure allows us to assess whether the most influential features identified by SHAP are actually actionable for adversarial manipulation. Experimental results highlight that only a limited subset of packet-level features is capable of altering the model's decision. In particular, modifying the `TTL` consistently yields very high evasion success rates across multiple packets (e.g., 100% for packets 2, 4, and 5). Similarly, the second most influential feature, `WIN`, achieves a 100% evasion rate when perturbed in packets 2, 4, and 6.

To further evaluate the consistency of the proposed strategy, we analyze the results obtained in the *generalization setting*. We use the SHAP values computed on the 100 samples from the previous analysis to guide the perturbation of a disjoint set of completely different samples. This allows us to assess whether the features identified as most influential remain actionable on unseen data. Perturbing only the 4^{th} `TTL` flips the prediction in 94% of cases, indicating that SHAP provides transferable and reliable insights across a broader input distribution. For the remaining samples, even perturbing the second and third most important features for SHAP yields no success, leading to failed adversarial manipulations. Further analysis of the class-conditional distribution for the fourth packet's `TTL` reveals that benign traffic predominantly exhibits a value of 128—a common default initial `TTL`—while this value is absent in malicious samples. However, in the adversarial setting, the `TTL` must be increased significantly beyond 128 to successfully flip the prediction from malicious to benign. This indicates that while the NIDS relies heavily on this feature, it remains relatively robust to minor perturbations due to the compensating contributions of other features.

C. Non-Targeted Random Perturbations

To assess the effectiveness of XAI-guided perturbations, we compare this procedure against perturbations applied randomly, without any prior knowledge (cf. Sec. III). Notably, for this experiment, we leverage the same 100 biflows used in the analysis reported in Sec. V-B. By adopting this fully unguided, random perturbation approach, the overall evasion success rate reaches only 54%, meaning that just over half of the targeted samples are successfully misclassified as benign.

Figure 4 depicts a heatmap of the evasion rate across the feature-packet space under random perturbations. Even without guidance, certain features demonstrate a higher impact on the model's prediction, closely mirroring the ranking observed with SHAP attributions: (i) `TTL` perturbations achieve the highest success rates from the second to the fifth packet, and in the seventh (100%), while dropping to 66% in the sixth packet; (ii) `WIN` modifications reach 100% success in the second and fourth packets, decreasing to 75% in the third, and having no effect from the seventh onward. Interestingly, `FLG` perturbations exhibit the most variable behavior, reaching 100% success when applied to packets four through eight (excluding the sixth). In contrast, altering `PL`, `DIR`, and `IAT` consistently fails to alter predictions, yielding a 0% success rate. Overall, the effectiveness of random perturbations drops sharply at the first and sixth packets, with only limited success observed for the seventh and eighth `TTL` and `FLG`, ultimately falling to 0% from the ninth packet onward for all features.

These results validate the utility of XAI-guided feature perturbation. By identifying the most influential features beforehand, we can strictly focus the perturbations on the fields most likely to subvert the NIDS's decision. Although random perturbations naturally expose some of these vulnerabilities, relying on XAI guidance enables highly targeted, efficient attacks, significantly minimizing the required perturbation attempts compared to an unguided, brute-force approach.

VI. Conclusion

In this work, we leveraged XAI explanations to perform informed adversarial attacks. We evaluated whether such explanations enable more effective feature manipulation and improve evasion success against DL-based NIDS in SDN environments. By computing SHAP values, we identified temporal patterns in the model's decision-making process, demonstrating that early-packet features—particularly `TTL`, `WIN`, and `DIR`—play a dominant role across different attack categories. Building on this insight, we evaluated the NIDS's vulnerability by exploiting both XAI-guided and random perturbations. Results show that XAI-guided adversarial attacks achieved a 94% evasion rate by modifying a single high-impact feature. Conversely, random perturbations succeeded in 54% of cases, showing that *XAI explanations can effectively guide the identification of the most actionable features for adversarial evasion*.

Future research could pursue various promising directions: (i) extending the evaluation to a multi-class setting to enable per-attack responses and more interpretable, class-specific SHAP attributions; (ii) improving model robustness by investigating advanced training methods (e.g., TRADES, randomized smoothing) and analyzing whether such approaches alter the set of vulnerable features, as well as considering multi-feature adversarial perturbations; and (iii) leveraging XAI to proactively identify and protect the most vulnerable features during the training phase.

References

[1] "Microsoft Digital Defense Report: 600 million cyberattacks per day around the globe," https://news.microsoft.com/en-cee/2024/11/29/microsoft-digital-defense-report-600-million-cyberattacks-per-day-around-the-globe/, 2024, accessed: 2025-06-12.

[2] D. Kreutz *et al.*, "Software-defined networking: A comprehensive survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2014.

[3] Z. Ahmad *et al.*, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. on Emerg. Telecommun. Technol.*, vol. 32, no. 1, p. e4150, 2021.

[4] K. Kharoubi *et al.*, "Network intrusion detection system using convolutional neural networks: Nids-dl-cnn for iot security," *Cluster Computing*, vol. 28, no. 4, p. 219, 2025.

[5] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[6] A. B. Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[7] D. L. Marino *et al.*, "An adversarial approach for explainable AI in intrusion detection systems," in *IEEE IECON*, 2018, pp. 3237–3243.

[8] E. Tcydenova *et al.*, "Detection of adversarial attacks in AI-based intrusion detection systems using explainable AI," *Human-Centric Comput Inform Sci*, vol. 11, 2021.

[9] P. Barnard *et al.*, "Robust network intrusion detection through explainable artificial intelligence (XAI)," *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.

[10] M. Tserenkhuu *et al.*, "Intrusion detection system framework for SDN-based IoT networks using deep learning approaches with xai-based feature selection techniques and domain-constrained features," *IEEE Access*, vol. 13, pp. 136 864–136 880, 2025.

[11] S. J. Cherian, "A novel intelligent classifier with explainable AI based intrusion detection system in SDN," in *IEEE INCSST*, 2025, pp. 1–7.

[12] S. Okada *et al.*, "XAI-driven black-box adversarial attacks on network intrusion detectors," *International Journal of Information Security*, vol. 24, no. 3, p. 103, 2025.

[13] B. Pham-Thai *et al.*, "An approach to attack classification for programmable network infrastructure using machine learning and deep learning solutions," *IEEE Access*, vol. 14, pp. 4886–4916, 2026.

[14] B. Chang Chung *et al.*, "A deepshap-based adversarial attack on machine learning-based network intrusion detection," *IEEE Access*, vol. 14, pp. 2566–2575, 2026.

[15] S. M. Lundberg *et al.*, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[16] A. Chakraborty *et al.*, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[17] M. S. Elsayed *et al.*, "Insdn: A novel sdn intrusion dataset," *IEEE Access*, vol. 8, pp. 165 263–165 284, 2020.

[18] M. Lopez-Martin *et al.*, "Network traffic classifier with convolutional and recurrent neural networks for internet of things," *IEEE Access*, vol. 5, pp. 18 042–18 050, 2017.

[19] A. Nascita *et al.*, "Machine and deep learning approaches for IoT attack classification," in *IEEE INFOCOM*, 2022, pp. 1–6.