Analysis and Mitigation of Adversarial Noise Attacks in Autonomous Optical Networks

Marc Ruiz*, Jaume Comellas, and Luis Velasco

Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

**e-mail: marc.ruiz-ramirez@upc.edu*

ABSTRACT

Autonomous optical networking increasingly relies on Network Digital Twins (NDTs) to automate complex operations through machine learning (ML). However, these ML pipelines are highly vulnerable to evasion attacks, where attackers might introduce subtle adversarial noise to mislead critical decision-making. This paper studies the impact of different adversarial techniques, and proposes a security framework utilizing model diversity and knowledge augmentation to detect and mitigate such threats. By employing an ensemble of general and specialized classifiers, the system identifies inconsistencies in optical constellation samples. Results demonstrate that this approach, paired with periodic fine-tuning, successfully cancels the effects of adversarial noise, ensuring robust network operation.

Keywords: autonomous optical networks, ML pipelines, adversarial attacks

1. Introduction

The journey toward Autonomous Networking (AN) is driven by the industry's pursuit of seamless, "zero-touch" operations. Guided by strategic frameworks from the TM Forum, major operators are currently navigating a roadmap that aims to achieve Level 5 Autonomy (ANL5)—full automation across multiple domains and technologies—by the year 2030 [1]. This evolution relies heavily on embedding native intelligence through Artificial Intelligence (AI) and Machine Learning (ML) to grant networks essential self-learning, self-detection, and self-healing capabilities. Within the optical layer, a primary focus of these ML applications is the estimation of Quality of Transmission (QoT) for optical lightpaths, which is a cornerstone of intelligent network management [2].

A central component in achieving AN is the Network Digital Twin (NDT), which uses telemetry data, models, and interfaces to analyze, diagnose, and provide critical inputs to Software Defined Networking (SDN) control and management entities. A prominent example is the OCATA NDT, which uses deep neural networks (DNN) to model signal propagation and accurately estimate QoT metrics from optical constellation (OC) analysis [3]. Indeed, the combination of distributed telemetry for continuous collection of OC samples and centralized AI-based lightpath analysis at the NDT results into ML pipelines that enable autonomous optical network operation [4].

Despite their benefits, the distributed nature of ML pipelines and the move toward disaggregated optical systems have significantly expanded the attack surface. One potential threat is evasion, where an adversary maliciously alters input data to produce misleading outputs, potentially causing the system to make incorrect autonomous decisions. For instance, an attacker (Mallory) performing a Man-in-the-Middle (MitM) attack can intercept and tamper lightpaths between a transmitter and receiver. Then, by combining this tampering with an evasion attack on the collected OC samples, the breach becomes extremely difficult to detect, undermining the integrity of the entire control plane [5].

In view of the above, it is of paramount importance to revisit and strength lightpath analysis procedures to incorporate novel countermeasures to mitigate and/or cancel the effect of potential sophisticated AI-based attacks. In this paper, we firstly present an evasion technique to introduce adversarial noise in OC samples to mislead reference, vulnerable AI-based lightpath analysis. Then, we present an improved AI-based lightpath analysis that combines model diversity and knowledge augmentation to detect and cancel the effect of adversarial noise addition.

2. Reference Scenario

Figure 1 illustrates the reference network scenario where a lightpath between sites *A* and *Z* is established. Without loss of generality, we assume that the main control plane elements needed for autonomous lightpath operation are hosted and running in one centralized control plane site. These control plane elements are: *i*) the *SDN controller* in charge of providing connectivity, as well as other procedures and policies related with QoT assurance; *ii*) the *telemetry manager*, that receives telemetry from distributed *telemetry agents* with the current status of the underlying system; and *iii*) the *OCATA NDT* that models the domain optical layer in order to perform AI-based lightpath analysis. Thus, OC samples are continuously collected and conveniently encoded and processed to add security and privacy before they are sent to the control plane site, where are decoded and processed at the OCATA NDT.

979-8-3195-4420-9/26/$31.00

Among the different algorithms and procedures that can be run in the OCATA NDT (see [3] for some examples), in this paper we assume a simplified AI-based lightpath analysis that consists in a pre-trained DNN image classifier [6] fine-tuned with OC samples. This model allows predicting relevant lightpath characteristics, e.g., the end-to-end distance, from the extracted features of the OC samples. Thus, a comparison between the actual and expected characteristics allows detecting inconsistencies that can be related with failures, anomalies, or attacks. In this case, the SDN controller is notified about that mismatch for consequent decision making.

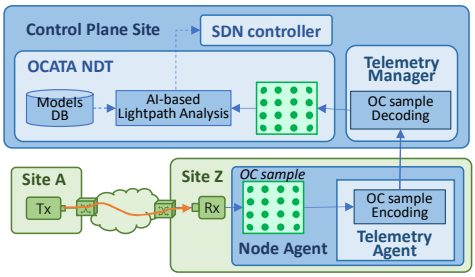

Figure 1 ML Pipeline

3. Adversarial Noise Attack Methodology

In this section we present the methodology to generate adversarial OC samples to degrade the accuracy of AI-based lighpath analysis models based on classification. Without loss of generality, we assume that the ML classifier used for lightpath analysis is gathered by the attacker before performing any evasion attack Figure 2), exploiting a model disclosure vulnerability of the NDT system. Thus, an adversarial sample X' is obtained as the result of adding a small perturbation δ to an original OC sample X (eq. 1). Two different methods for adversarial perturbation generation are considered: the Fast Gradient Sign (FGS) method [7] and the Targeted Clean-Label (TCL) method [8].

The FGS method represents a foundational adversarial methodology that utilizes gradient information to introduce perturbations into input data. This technique is employed at inference time to intercept a legitimate signal and inject calculated noise based on the gradient obtained through the backpropagation of the input through the classifier model. This process is effective because the gradient provides a precise mathematical direction in which the input can be adjusted to maximize the model's loss function. By moving the input data in the direction that the model's error increases most rapidly, the attacker can systematically mislead the classifier into producing an incorrect prediction.

The intensity and magnitude of these gradient-based modifications are governed by the hyperparameter ε. The relationship between ε and the attack efficacy is generally proportional: as ε increases, the probability of a successful misclassification (attack success rate) rises. However, this leads to a critical trade-off, as higher values of ε produce more prominent distortions. To ensure that the attack remains silent and bypasses security detection, the attacker must identify an optimal balance where the perturbations are subtle enough to remain imperceptible to human observers while still being robust enough to deceive the detection system. More formally, the perturbation δ_{FSM} can be generated according to eq. (2), where the sign of the gradient ∇ of the loss function J with respect to X is computed. This gradient is determined by the set of parameters θ of the reference classifier model and the true label y of the original sample.

While FGS offers a computationally efficient means of inducing misclassification, it is limited by its untargeted nature; the objective is merely to provoke an error, regardless of the resulting output class. Furthermore, FGS exhibits significant instability at higher perturbation magnitudes, leading it to predict a single class universally for all inputs. To mitigate these limitations, the TCL method provides a more sophisticated, optimization-based framework. TCL allows the attacker to manipulate the classifier into predicting a specific predetermined target class. By optimizing the input in the feature space rather than just the image space, TCL achieves a more precise and robust attack that remains subtle to try to bypass any detection mechanism. Thus, the primary objective of TCL is to generate a perturbed instance X' that resides close to the target label y' in the model's feature space f, while maintaining a high degree of similarity to the original base instance X in the input space. The TCL attack is formulated as a solution to the optimization problem denoted by eq. (3), where parameter β serves as a trade-off controller: a higher value prioritizes imperceptibility, whereas a lower one favors higher attack success rates by allowing more aggressive noise addition.

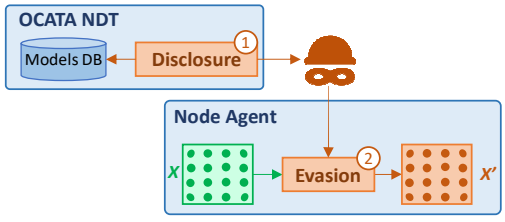

Figure 2 Adversarial Noise Attack

$$X' = X + \delta_i, \forall\, i = \{FSM, TCL\} \tag{1}$$

$$\delta_{FSM} = \varepsilon \cdot sign(\nabla J(\theta, X, y)) \tag{2}$$

$$\delta_{TCL} = \underset{\delta}{argmin} \|f(X') - f(X(y'))\|_2^2 + \beta \cdot \|\delta\|_2^2 \tag{3}$$

4. Detection and Mitigation of Adversarial Noise Attacks

To counter the evasion vulnerabilities inherent in the gradient-based attacks, we propose a detection and mitigation framework focused on model diversity and knowledge augmentation (Figure 3). This approach employs alternative models configured with different characteristics to complicate the success of adversarial noise addition. Moreover, it enables fine-tuning of ML classifiers that is periodically executed to incorporate the knowledge of adversarial attacks, making classifiers more robust and reliable.

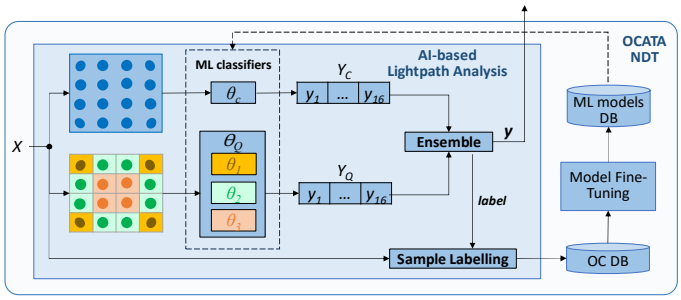

Figure 3 AI-based Lightpath Analysis Architecture

This section focuses on the procedure executed each time a new OC sample is received, consisting in predicting the true lightpath distance as a categorical class based on the dispersion of the symbols in the OC sample. The prediction is based on ML classifiers that process the symbols in of a single CP, and predict the class. Then, an ensemble approach selecting the majority class among all CPs returns a single prediction. To add diversity, two different approaches for the ML classifiers are considered. On the one hand, a single general model θ_c is used for all the CPs. On the other hand, a set of models Θ_Q containing CP-specialized models for different OC regions are selectively used, namely: *i*) a model for corner CPs, ii) a model for side CPs, iii) a model for central CPs.

Algorithm 1 shows the procedure, that receives a given OC sample X, the ML classifiers θ_c and Θ_Q, as well as other parameters, and returns the predicted class. After some initializations, the whole OC sample (16-QAM is considered) is divided into a set of CP symbols (line 1-2 in Algorithm 1). Then, the predicted class with both approaches is computed and stored (lines 3-6). After this computation, the majority class y and proportion p of CPs returning that class is computed (lines 7-8). Once the predictions are computed, they are processed to return a single diagnosis. Moreover, OC sample is labelled and stored for retraining/fine-tuning purposes. Thus, three different possibilities are considered for that labelling. First, when both generic and CP-specialized models return the same majority class, we consider that prediction as normal (lines 9-11). Second, when there is a mismatch between predictions but proportions are above a given threshold *thr*, the sample is stored as suspicious (lines 12-15). Finally, in case that predictions do not match and confidence is below a threshold, the sample is considered and stored as adversarial, and the returned prediction is null (lines 16-17). In this way, improving the robustness of ML classifiers by means of fine-tuning with adversarial samples is enabled.

Algorithm 1 - AI-based Lightpath Analysis

IN: X, θ_c, Θ_Q *thr*, *DB* **OUT:** y

1: $Y_C \leftarrow \emptyset$; $Y_Q \leftarrow \emptyset$
2: $\{x_{1,\ldots,}x_{16}\} \leftarrow$ splitCP(X)
3: **for** $x \in \{x_{1,\ldots,}x_{16}\}$ **do**
4: $Y_C \leftarrow Y_C \cup$ predict(x, θ_c)
5: $\theta_q \leftarrow$ getCPmodel(x, Θ_Q)
6: $Y_Q \leftarrow Y_Q \cup$ predict(x, θ_q)
7: $<y_c, p_c> \leftarrow$ getMajorityClass(Y_C)
8: $<y_q, p_q> \leftarrow$ getMajorityClass(Y_Q)
9: **if** $y_c == y_q$ **then**
10: push(*DB*, X, 'normal')
11: **return** y_c
12: **if** min(p_c, p_q)>*thr* **then**
13: push(*DB*, X, 'suspicious')
14: **if** $p_c \geq p_q$ **then return** y_c
15: **else return** y_q
16: push(*DB*, X, 'adversarial)
17: **return** ∅

5. Illustrative Results

For evaluation purposes, we reproduced a large number of lightpath configurations, ranging from 80km to 2000km, with the OC sample dataset available in [9]. The reference ML classifiers (both general and CP-specialized) are based on the MobileNetV2 architecture [6], adapted and tuned to classify OC samples in the available dataset in 5 classes depending on the lighpath distance range: *i*) [80km, 400km], *ii*) (400km, 800km], *iii*) (800km, 1200km], *iv*) (1200km, 1600km], and *v*) (1600km, 2000km].

Figure 4a-b shows the performance of FGS method. First of all, we focus on the best tuning of ε parameter in order to have large attack accuracy (% of successful attacks) with small perturbations hard to detect. In view of the figures, we can conclude that ε =1e-3 provides high attack accuracy (inducing large misclassifications) with a distribution of classes that is relatively close to the normality intervals in the absence of attacks (orange dotted lines in Figure 4b). Note that ε > 1e-3 does not improve attack effect, whereas it causes a large imbalance in classes prediction, which would be easy to detect due to clear abnormal patterns in classification results.

Figure 4c shows the performance of TCL method as a function of β. The colored areas denote the visual perception of adversarial samples; β values larger than 100 are required to get similar imperceptibility than that of

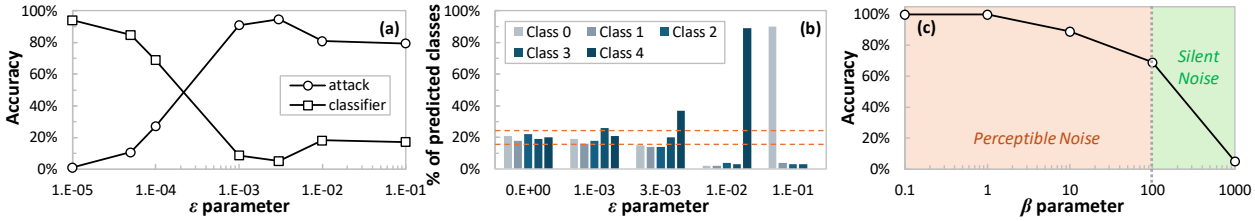

Figure 4 FGS Performance (a-b), TCL Performance (c)

FGS configured with ε =1e-3. Since the accuracy of attack drops below 80% when proper β value are is, we decide to select FGS as reference adversarial noise technique. Figure 4c shows an example of a CP (-3 -3i) of a lightpath of 1600 km before and after applying FGS. The addition of few symbols with a similar dispersion than the original ones confirm the effectiveness of the selected approach.

Finally, Figure 6 shows the performance of AI-based lightpath analysis. First, the resultant proportion p using either general model θ_c or CP-specialized models θ_Q in the absence of attack is shown. Note that both approaches provide large p values (>95%), as well as a remarkably high consensus proportion (all CPs return the same class), leading us to setup thr = 80% (green dotted line). Then, when an adversarial noise attack is injected (FGS with ε =1e-3), the ensemble procedure returns p that clearly drops below thr, which makes the attack it easy to detect. Eventually, the consideration of fine-tuning with adversarial samples is considered. In this case, the attack is not detected because it is not affecting classifier performance (similar to that in the absence of attack), therefore completely cancelling the malicious purpose of the attacker.

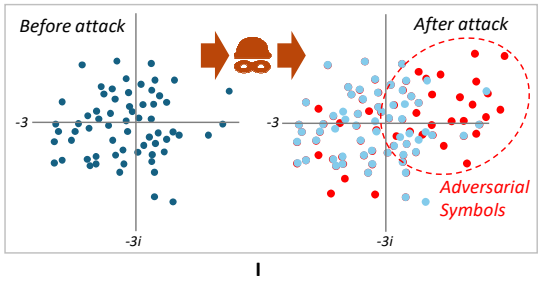

Figure 5 Example of Adversarial Noise

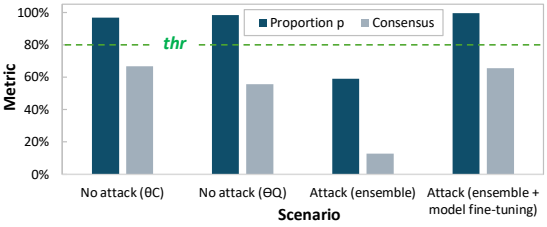

Figure 6 AI-based Lightpath Analysis Performance

6. Conclusions

This paper analyzed security threats to autonomous optical network ML pipelines, identifying how adversarial noise attacks can compromise NDT integrity. To mitigate these risks, we proposed a framework using model diversity and knowledge augmentation based periodic fine-tuning to effectively cancel the impact of attacks.

ACKNOWLEDGEMENT

The authors want to thank Martina Vichino and Adrián Cerezuela for their valuable contributions. The research leading to these results has received funding from the European Union's Horizon Europe research and innovation programme under G.A. No. 101092766 (ALLEGRO), the project PID2024-157824OB-I00 funded by MICIU/ AEI/10.13039/501100011033 /FEDER, UE and from ICREA.

REFERENCES

[1] O. González de Dios *et al.*, “Automation of multi-layer multi-domain transport networks and the role of AI,” IEEE JOCN, vol. 17, 2025.

[2] L. Zhang *et al.* "A survey on QoT prediction using machine learning in optical networks," Elsevier OFT, vol. 68, 2022.

[3] M. Devigili *et al.*, “Applications of the OCATA Time Domain Digital Twin: from QoT Estimation to Failure Management,” IEEE JOCN, vol. 16, 2024.

[4] P. Gonzalez *et al.*, “Experimental Validation of Lightpath Operation based on Continuous Digital-Twin Model Tuning,” in Proc. OFC, 2026.

[5] M. Ruiz *et al.*, “Network digital twinning solutions to increase the security of multi-domain optical transport networks,” IEEE JOCN, vol. 18, 2026.

[6] M. Sandler *et al.*, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. CVPR, 2019.

[7] A. Lad *et al.*, “Fast Gradient Sign Method Variants in White Box Settings: A Comparative Study,” in Proc. ICICT, 2024.

[8] C. Zhu *et al.*, “Transferable Clean-Label Poisoning Attacks on Deep Neural Nets,” in Proc. ICML, 2019.

[9] M. Ruiz *et al.*, “Optical Constellation Analysis (OCATA)”, CORA Dataverse Repository, 2022. doi:10.34810/data146