Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2026.076608

ARTICLE

From Hardening to Understanding: Adversarial Training vs. CF-Aug for Explainable Cyber-Threat Detection System

Malik Al-Essa[1,*], Mohammad Qatawneh[2,1], Ahmad Sami Al-Shamayleh[3], Orieb Abualghanam[1] and Wesam Almobaideen[4,1]

[1]Computer Science Department, King Abdullah II Faculty for Information Technology, The University of Jordan, Amman, 11942, Jordan

[2]Department of Networks and Cybersecurity, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, 19111, Jordan

[3]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, 19111, Jordan

[4]Department of Electrical Engineering and Computing Sciences, Rochester Institute of Technology, Dubai, 341055, United Arab Emirates

*Corresponding Author: Malik Al-Essa. Email: m.alessa@ju.edu.jo

Received: 23 November 2025; Accepted: 22 December 2025; Published: 09 April 2026

ABSTRACT: Machine Learning (ML) intrusion detection systems (IDS) are vulnerable to manipulations: small, protocol-valid manipulations can push samples across brittle decision boundaries. We study two complementary remedies that reshape the learner in distinct ways. Adversarial Training (AT) exposes the model to worst-case, in-threat perturbations during learning to thicken local margins; Counterfactual Augmentation (CF-Aug) adds near-boundary exemplars that are explicitly constrained to be feasible, causally consistent, and operationally meaningful for defenders. The main goal of this work is to investigate and compare how AT and CF-Aug can reshape the decision surface of the IDS. eXplainable Artificial Intelligence (XAI) is used to analyze the shifts in global feature importance stability under both AT and CF perturbation to link these shifts to the accuracy of the IDS in detecting cyber-threats. This yields a clear picture when boundary hardening (AT) or boundary sculpting (CF-Aug) better serves IDS. Two well-known techniques are used to generate adversarial samples, namely the Fast Gradient Sign Method (FGSM) and the Projected Gradient Descent (PGD) techniques. We have achieved better accuracy with AT and CF-Aug compared to the baseline IDS.

KEYWORDS: eXplainable artificial intelligence (XAI); intrusion detection systems (IDS); counterfactual explanation; adversarial training; deep learning

1 Introduction

AI has become a core engine of technological progress, with steady advances across different domains that include computer vision, healthcare, and autonomous platforms [1–3]. Cybersecurity as one of the domains that started to integrate AI, particularly ML and DL, in its technologies. Data-driven methods now reinforce digital infrastructure against evolving threats through IDS/Intrusion Prevention Systems (IPS), malware and phishing classification, behavior- and anomaly-based monitoring at network and endpoint layers [4]. This adoption wave is propelled by abundant, fine-grained telemetry, network flows, system logs, and endpoint signals, paired with elastic compute and rapid progress in representation learning.

Contemporary campaigns increasingly mix AI-assisted evasion with zero-day exploitation and rapidly adapting malware, which raises the bar for detection systems: they must learn new patterns quickly and provide transparent reasoning for their decisions [5]. However, adversaries actively generate inputs to mislead ML pipelines, escalating an offense-defense arms race that makes robust and interpretable security models a practical requirement rather than an academic ideal [6]. Even though there is an increased adoption of AI in cybersecurity, two limitations still affect this adoption. First, DL models are vulnerable to small, protocol-consistent manipulations that can produce large prediction swings, enabling adversarial activity that remains operationally plausible. Second, many state-of-the-art Deep Neural Network (DNN) architectures operate as black boxes, offering limited visibility into why an alert was triggered, which features controlled the decision, or how the prediction would change under a feasible mitigation. In practice, this combination of vulnerability to adversarial perturbations and black-box nature erodes analyst trust, complicates governance and compliance audits, and slows incident response.

Recent research directions to address these weaknesses include a call for techniques that balance robustness with domain constraints and integrate explainability to actionable controls. Explainability should progress beyond post hoc saliency to include counterfactuals, uncertainty estimates, and feature stability analyses that map cleanly to defense levers (rate limits, ACLs, service hardening). This work advances this agenda by comparing two complementary strategies, AT and CF-Aug, under security-aware constraints, and by assessing not only accuracy but also the explainability of the DL models. Using MoDel Agnostic Language for Exploration and eXplanation (DALEX), an XAI technique, we provide model-agnostic explanations that reveal how each intervention shifts feature importance and decision boundaries under clean and adversarial conditions.

The main contributions of this paper are:

1. We design a unified experimental framework that directly compares AT and CF-Aug for DL-based IDS, under the same architecture, datasets, and evaluation metrics. This allows us to disentangle the effects of boundary hardening (AT) from boundary sculpting (CF-Aug) on cyber-threat detection performance.
2. We introduce an XAI-driven augmentation pipeline in which counterfactual explanations are constrained to be protocol-valid, label-preserving, and operationally meaningful for defenders. These counterfactuals are then reused as training samples, enabling us to quantify how XAI-generated data reshapes the decision surface and improves minority-class detection.
3. We perform a global explainability study using DALEX to compare feature-importance profiles across Baseline, AT, and CF-Aug models. By connecting shifts in global importance to changes in per-class F1, we show that CF-Aug encourages the model to rely more on semantic, security-relevant features, especially for rare attack types such as R2L and U2R in NSL-KDD and Slowloris/SSH-Patator in CICIDS17.
4. We evaluate the framework on two benchmark datasets (NSL-KDD and a cleaned multi-class version of CICIDS17) and report detailed per-class metrics, highlighting the conditions under which AT and CF-Aug provide complementary benefits in terms of overall robustness, minority-class F1, and interpretability.

The remainder of this paper is structured as follows. Section 2 surveys the most relevant work in the literature. Section 3 presents the proposed approach in detail. Section 4 introduces the datasets, describes the implementation setup, and analyzes the experimental results. Finally, Section 5 summarizes the main contributions and discusses potential avenues for future research.

2 Related Work

2.1 Adversarial Learning

Research on AT has developed along both the "offensive" and "defensive" dimensions of the problem. On the one hand, a large body of work focuses on how small, carefully designed changes to legitimate inputs—so-called adversarial samples—can sharply raise the misclassification rate of DL models [7]. Quite a few attack algorithms have been put forth for this purpose. Among them, the Fast Gradient Sign Method (FGSM) [8] is considered one of the most popular gradient-based white-box attacks. However, it may suffer from catastrophic overfitting when being naively integrated into training procedures [9]. FGSM and similar methods leverage the model's loss landscape, usually cross-entropy, to compute perturbations that nudge inputs in the direction of regions where the classifier's decisions are more fragile for a given class.

On the defensive side, research has mostly converged on two main families of methods: AT and provable (or certified) defenses [10]. In a nutshell, AT-based defenses augment the training set by adding adversarially perturbed samples (while keeping their original labels) and then retraining or fine-tuning the classifier so that it learns to correctly classify both clean and perturbed inputs. The goal here is to get a model that is empirically more robust than the original one against a specified threat model. Provable defenses, on the other hand, aim to provide formal guarantees of robustness by computing and optimizing robustness certificates, usually at considerable additional computational cost and reduced scalability. Empirical studies such as [9] also suggest that AT remains the most widely adopted approach in practice, since it can provide strong empirical robustness, scales fairly well to large DNNs, and can be adapted to different attack algorithms. For these practical reasons, we focus on AT and refer the reader to [11] for a recent taxonomy of its many variants.

The interaction of adversarial learning with cybersecurity has attracted a lot of attention, predominantly from the attack perspective. Various works design adversarial samples specifically to evade network intrusion and malware detectors [12,13], showing that even small perturbations can significantly degrade the performance of DL-based security systems. More recently, a smaller but growing body of work has started to explore adversarial defenses for DL in cybersecurity settings. For example, AT has been combined with generative adversarial networks (GANs) to improve intrusion detection and malware classification in [14]. In [15], adversarial samples are generated to extend and rebalance the training data, increasing the fraction of malicious traffic in a binary classification task. Reference [16] applies AT to enhance the robustness of supervised models for the automated detection of cyber threats in industrial control systems. Reference [17] performs a systematic evaluation of various evasion attacks, including an evaluation of AT on multiple DNN architectures.

2.2 XAI

Most DL models today are still black boxes: they produce some final verdict, for instance, benign vs. malicious or normal traffic vs. attack traffic, but fail to provide any clue regarding why they made that decision. For security teams, this lack of insight erodes trust, slows incident response, and makes it increasingly difficult for analysts to confirm or contest alerts in high-stakes environments. In particular, Reference [18] discussed how XAI techniques can be used to explain the decisions made by neural models for malware detection and vulnerability discovery. Similarly, Reference [19] applies XAI methods towards the identification of which input features were of the most importance for the detection of each intrusion category, while Reference [20] proposes an explainable IDS/IPS designed for cloud environments. More recently, researchers have started to interlink XAI with adversarial learning in cybersecurity: Reference [21] uses an adversarial learning framework to understand why some network intrusions are misclassified by a DNN, expressing explanations as the smallest changes needed in the input to flip the model's decision for

misclassified samples. In [22], local, instance-level explanations were combined with AT, where instance-level explanations helped to guide fine-tuning of a DNN that had been previously trained with AT. Reference [23] integrates counterfactual explanations in order to provide defense mechanisms against poisoning and backdoor attacks. In particular, a counterfactual explanation algorithm, originally designed to generate "what-if" explanations, is repurposed for the purpose of systematically modifying inputs in ways that drive the model toward different predictions [24]. Golden Jackal Driven Optimization is used to design a transparent, interpretable intrusion detection system that leverages explainable AI to strengthen and modernize cybersecurity defenses [25]. DoS Attack Detection Enhancement in Autonomous Vehicle Systems with XAI focuses on improving the reliability of detecting DoS attacks while providing transparent, interpretable insights into the Autonomous Vehicles' security decisions [26].

Several recent studies combine AT with explainability techniques for IDS. For example, in our prior work [22], we used XAI to analyze and guide the AT process by identifying vulnerable regions of the feature space, but the explanations were not reused as synthetic training data. In contrast, the present work leverages counterfactual explanations directly as an augmentation mechanism (CF-Aug), allowing us to study how XAI-generated samples reshape the decision boundary and affect robustness. A complementary line of research focuses on building explainable IDS models without explicitly contrasting different robustness strategies. Existing XAI-IDS approaches typically employ post-hoc methods such as feature-importance analyses or local explanations to interpret deep or tree-based detectors, but they do not systematically compare AT and CF-Aug under a unified experimental setup. Our study fills this gap by placing AT and CF-Aug on the same footing and jointly evaluating their impact on both predictive performance and model interpretability. Table 1 summarizes related work and compares it to the proposed approach.

Table 1: Summary of representative adversarial training and XAI-based schemes related to IDS, positioned relative to the proposed approach

Work (Ref.)	Domain/Task	AT	CF	XAI
[14]	Network intrusion detection	Yes	No	No
[27]	Malware detection with GAN-based defenses	Yes	No	No
[15]	Android malware detection	Yes	Partial	No
[16]	ICS cyber-threat detection	Yes	No	No
[17]	DL-based IDS under evasion attacks	Yes	No	No
[18]	Malware and vulnerability detection	No	No	Yes
[19]	Intrusion detection (feature importance)	No	No	Yes
[20]	Cloud-based IDS/IPS	No	No	Yes
[21]	Intrusion detection with adversarial analysis	Yes	No	Yes
[22]	Cybersecurity classification with guided fine-tuning	Yes	No	Yes
[23]	Defenses against poisoning/backdoor attacks	No	Yes	Yes
[25]	Intrusion detection with metaheuristic optimization	No	No	Yes
[26]	DoS detection in autonomous vehicles	No	No	Yes
This work	DL-based IDS on NSL-KDD and CICIDS17	Yes	Yes	Yes

3 Method

The proposed method, as shown in Algorithm 1 (Fig. 1 illustrated the schema for the proposed method), establishes a comparative framework to study how distinct training regimes shape both robustness and global

interpretability in DL. Given a labeled dataset $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with $\mathbf{x}_i \in \mathbb{R}^d$, a baseline DL model $\mathcal{M}_\theta$ is first trained under standard empirical risk minimization and serves as the reference point for subsequent augmentation strategies (Algorithm 1 line 2). To induce robustness to gradient-based perturbations, an adversarial augmentation set $\mathcal{A}$ is generated by applying FGSM and PGD to each training sample under a perturbation factor ϵ (Algorithm 1 line 6). Concretely, FGSM constructs

$$\tilde{\mathbf{x}} = \mathbf{x} + \varepsilon \operatorname{sign}(\nabla_{\mathbf{x}} \ell(\mathcal{M}_\theta(\mathbf{x}), y)), \tag{1}$$

where $\nabla_{\mathbf{x}}$ denotes the gradient computed with respect to $\mathbf{x}$, and $\ell(\mathcal{M}_\theta(\mathbf{x}), y))$ denotes the loss function of the neural model M_θ. while PGD performs K iterative steps with step size α and projection $\Pi_{\mathcal{B}_p(\mathbf{x},\varepsilon)}$ onto the ℓ_p-ball of radius ε:

$$\mathbf{x}^{(t+1)} \leftarrow \Pi_{\mathcal{B}_p(\mathbf{x},\varepsilon)}\left(\mathbf{x}^{(t)} + \alpha \operatorname{sign}\left(\nabla_{\mathbf{x}} \ell(\mathcal{M}(\mathbf{x}^{(t)}), y)\right)\right), \quad \mathbf{x}^{(0)} = \mathbf{x} \tag{2}$$

In all our experiments, Eqs. (1) and (2) are instantiated with the ℓ_∞ norm. For both the NSL-KDD and CICIDS17 datasets, FGSM uses a perturbation factor of $\varepsilon = 0.01$, while PGD uses the same factor with a step size of step_size = 10.

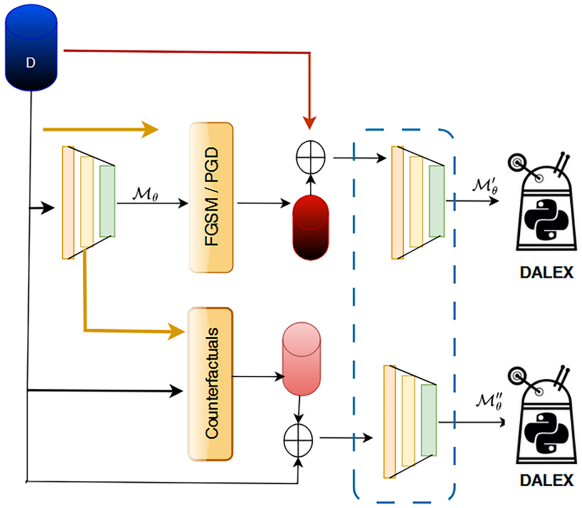


Figure 1: Overall architecture of the proposed explainable intrusion detection framework comparing the baseline, adversarially trained, and counterfactually augmented DL models

Algorithm 1: Proposed method

Input:

$\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$: labeled training set of N samples, where each $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector (e.g., network flow, malware descriptor, email metadata).
Attack generators: FGSM, PGD (implemented as in Eqs. (1) and (2), with hyperparameters ϵ, number of steps, step size, and norm specified in Section 4.3).
Counterfactual generator: a procedure that, given a model and an input $\mathbf{x}$, searches for a minimally modified, domain-feasible sample $\hat{\mathbf{x}}$ that changes the model's prediction, under the constraints described in Section 4.3. DALEX: global explanation technique for trained models, returning

(Continued)

Algorithm 1 (continued)

feature-importance profiles.
ϵ: perturbation value (attack budget) used for adversarial sample generation (FGSM/PGD).

Output:

$\mathcal{M}_\theta$: baseline DL model trained only on the original dataset $\mathcal{T}$.
$\mathcal{M}_\phi$: adversarially trained DL model (trained on $\mathcal{T} \oplus \mathcal{A}$, where $\mathcal{A}$ contains FGSM/PGD adversarial examples generated around $\mathcal{T}$).
$\mathcal{M}_\nabla$: counterfactually trained DL model (trained on $\mathcal{T} \oplus \mathcal{C}$, where $\mathcal{C}$ contains counterfactual examples generated around $\mathcal{T}$).
Global explanations (feature-importance profiles) for each model, used to compare how AT and CF-Aug reshape the decision function.

begin
{Step 1: Train baseline model on the original dataset (no augmentation)
$\mathcal{M}_\theta \leftarrow \mathsf{trainModel}(\mathcal{T})$
{Step 2: Generate adversarial augmentation set $\mathcal{A}$ using $\mathcal{M}_\theta$ and FGSM/PGD
$\mathcal{A} \leftarrow \varnothing$
foreach $(\mathbf{x}, y) \in \mathcal{T}$ **do**
$\tilde{\mathbf{x}} \leftarrow \mathsf{AdversarialAttack}(\mathcal{M}_\theta, \mathbf{x}, y, \varepsilon)$ // e.g., FGSM or multi-step PGD around $\mathbf{x}$
$\mathcal{A} \leftarrow \mathcal{A} \cup \{(\tilde{\mathbf{x}}, y)\}$
{Step 3: Generate CF-Aug set $\mathcal{C}$ using $\mathcal{M}_\theta$
$\mathcal{C} \leftarrow \varnothing$
foreach $(\mathbf{x}, y) \in \mathcal{T}$ **do**
$\hat{\mathbf{x}} \leftarrow \mathsf{GenerateCounterfactual}(\mathcal{M}_\theta, \mathbf{x}, y)$ // minimal change to flip prediction, under domain constraints
$\mathcal{C} \leftarrow \mathcal{C} \cup \{(\hat{\mathbf{x}}, y)\}$ // Infeasible or non-flipping candidates are discarded inside GenerateCounterfactual
{Step 4: Adversarial training (classical adversarial learning on $\mathcal{T}$ plus $\mathcal{A}$)
$\mathcal{T}_\mathcal{A} \leftarrow \mathcal{T} \cup \mathcal{A}$
$\mathcal{M}_\phi \leftarrow \mathsf{trainModel}(\mathcal{T}_\mathcal{A})$
{Step 5: Counterfactual training (XAI-driven augmentation on $\mathcal{T}$ plus $\mathcal{C}$)
$\mathcal{T}_\mathcal{C} \leftarrow \mathcal{T} \cup \mathcal{C}$
$\mathcal{M}_\nabla \leftarrow \mathsf{trainModel}(\mathcal{T}_\mathcal{C})$
{Step 6: Global explainability with DALEX (compare feature-importance profiles)
$\mathsf{Explain}_\theta \leftarrow \mathsf{DALEX}(\mathcal{M}_\theta, \mathcal{T})$
$\mathsf{Explain}_\phi \leftarrow \mathsf{DALEX}(\mathcal{M}_\phi, \mathcal{T}_\mathcal{A})$
$\mathsf{Explain}_\nabla \leftarrow \mathsf{DALEX}(\mathcal{M}_\nabla, \mathcal{T}_\mathcal{C})$
// Compare global feature-importance profiles to study how AT vs. CF-Aug change the learned decision logic

These adversarial samples are retained with their ground-truth labels and aggregated into $\mathcal{A}$, after which a second model $\mathcal{M}_\phi$ is trained on $\mathcal{T} \cup \mathcal{A}$ following the classical AT paradigm, typically improving accuracy (Algorithm 1 line 12). In parallel, the method introduces CF-Aug to encourage semantically meaningful and smoother decision boundaries: for each $(\mathbf{x}, y) \in \mathcal{T}$, a minimally modified $\hat{\mathbf{x}}$ is generated that flips the *baseline* model's prediction while preserving the original label, and these counterfactuals are collected into $\mathcal{C}$ (Algorithm 1 line 10). In this work, generating a counterfactual that "flips the prediction minimally" is formalized as a constrained optimization problem. Given a trained baseline model $\mathcal{M}_\theta$ and an input $\mathbf{x}$, we seek a counterfactual $\hat{\mathbf{x}}$ that changes the model's predicted class while staying as close as possible to $\mathbf{x}$ and remaining domain-feasible. Concretely, we approximate the solution of

$$\min_{\hat{\mathbf{x}}} \ \|\hat{\mathbf{x}} - \mathbf{x}\|_1 \quad \text{s.t.} \quad \arg\max_c \mathcal{M}_\theta(\hat{\mathbf{x}})_c \neq \arg\max_c \mathcal{M}_\theta(\mathbf{x})_c, \ \hat{\mathbf{x}} \in \mathcal{F},$$

where $\mathcal{F}$ denotes the set of feasible samples defined by domain constraints. The optimization is implemented as an iterative search procedure that updates $\hat{\mathbf{x}}$ while monitoring both the model prediction and the distance $\|\hat{\mathbf{x}} - \mathbf{x}\|_1$; after each update, the candidate is projected back onto $\mathcal{F}$ and discarded if it violates any constraint, exceeds a maximum allowed distance $\|\hat{\mathbf{x}} - \mathbf{x}\|_1 \leq \tau$, or fails to flip the prediction within a fixed number of iterations. Only candidates that both flip the prediction and satisfy all feasibility criteria are retained and added to the CF-Aug set $\mathcal{C}$, ensuring that CF-Aug uses realistic, near-boundary samples rather than arbitrary perturbations.

Training a third model $\mathcal{M}_\nabla$ on $\mathcal{T} \cup \mathcal{C}$ promotes decision surfaces that are less sensitive to superficial manipulations, especially when the counterfactual generator enforces domain constraints such as non-negativity, type integrity, and known invariants (Algorithm 1 line 14). After training $\mathcal{M}_\theta$, $\mathcal{M}_\phi$, and $\mathcal{M}_\nabla$, the pipeline employs DALEX to obtain global explanations (e.g., feature-importance profiles); comparing these profiles across the three models (Algorithm 1 line 17) reveals whether robustness-oriented training reallocates importance from brittle or spurious features to more stable, task-relevant features.

4 Empirical Evaluation and Discussion

We assessed the effectiveness of the proposed approach through a series of experiments conducted on two established benchmark datasets for cybersecurity. The details of these datasets are presented in Section 4.1, while Section 4.2 details the implementation aspects of the proposed method. The results are discussed in Section 4.3.

4.1 Datasets

Two datasets are used to evaluate the proposed method, i.e., NSL-KDD and CICIDS17. Both datasets contain different classes (normal class and attack classes). The descriptions for the datasets is shown in Table 2.

Table 2: Data description (#: Number of)

Dataset	#Training set	#Testing set	#Features	#Classes
NSL-KDD	25,192	22,544	41	5
CICIDS17	100,000	100,000	72	9

NSL-KDD

The NSL-KDD dataset [28][1] contains labeled network-flow records covering *normal* traffic and four broad attack categories: Denial of Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R). Each record comprises 41 attributes (37 numeric, 3 categorical, plus the class label); feature definitions are provided in [28]. A distinctive aspect of NSL-KDD is its shift between training and testing distributions: the training split includes 21 attack subtypes, whereas the test split contains 37, introducing 16 previously unseen attacks at evaluation time. This property stresses generalization to novel behaviors, an ability required by practical IDS. The label distribution is highly skewed, with R2L and U2R representing rare classes.

Although NSL-KDD is dated and does not perfectly reflect today's production networks, it remains a de facto multi-class benchmark that enables controlled comparison across IDS methods and careful analysis of minority-class detection. Following common practice, we adopt the predefined NSL-KDD splits. We use the `KDDTrain+20Percent` subset as the training set and `KDDTest+` as the held-out test set. This preserves the benchmark's original difficulty and allows a fair comparison with prior work while keeping the training size computationally manageable. All models are trained on `KDDTrain+20Percent` and evaluated on `KDDTest+`. A stratified 20% subset of the training set is further held out as a validation set for hyperparameter tuning.

CICIDS17

The CICIDS17 dataset [29] is one of the most commonly used benchmarks in cybersecurity research. It was captured in a realistic network testbed that included a variety of devices and operating systems, arranged into separate victim and attacker networks that communicated over the Internet. The recorded traffic covers widely used protocols such as HTTP, HTTPS, FTP, SSH, and SMTP. To better reflect real-world usage, the authors employed profiling agents—trained on traffic generated by real human users—to automatically produce both benign and malicious network flows. The dataset contains multiple attack scenarios, including brute-force attacks, Heartbleed, botnet communications, several forms of DoS and DDoS, infiltration attempts, and web-based attacks. Each flow in CICIDS17 is represented by 78 numerical features and a single class label, all extracted using the CICFlowMeter tool and described in detail in [29].

For our experiments, we rely on the refined version of CICIDS17[2] released by [30]. This cleaned version corrects issues in the original dataset by removing meaningless artifacts, fixing dataset errors, and discarding mislabeled traces. In particular, it retains 72 features from the original feature set, excluding those deemed irrelevant or prone to causing overfitting. The refined dataset was previously used in [31] for binary classification; in this work, we extend its use to a multi-class classification setting. From this dataset, we construct two disjoint subsets of 100,000 samples each: one for training and one for testing. We perform *stratified sampling without replacement* using a fixed random seed, preserving the original class proportions in both splits (approximately 80% benign and 20% attack traffic). Within the training subset, we randomly select 20% of the samples (again using stratified sampling and the same seed) as a validation set. This procedure ensures that (i) the training and test sets do not overlap, (ii) each split reflects the original class distribution of CICIDS17, and (iii) our results are reproducible. The final processed dataset includes 8 attack types: PortScan, DoS Hulk, DDoS, DoS GoldenEye, DoS Slowloris, FTP-Patator, SSH-Patator, and DoS Slowhttptest.

4.2 Implementation Details

The implementation was carried out in Python 3.12 using Keras 2.7 for model construction, with TensorFlow as the computational backend. All numerical features were normalized via min–max scaling,

[1] https://www.unb.ca/cic/datasets/nsl.html
[2] downloads.distrinet-research.be/WTMC2021

which linearly maps each feature to the interval $[0, 1]$. Normalization was fit on the training split and applied consistently to test sets to prevent leakage. This step mitigates disproportionate influence from features with different units or scales and typically improves convergence stability for gradient-based optimization. Neural-network hyperparameters were tuned with the Tree-structured Parzen Estimator (TPE) as implemented in `Hyperopt`. We allocated 20% of the training set as a stratified validation subset to preserve class proportions, following a Pareto-style allocation to balance exploration and validation reliability. We executed thirty optimization trials and selected the configuration minimizing validation loss. The search space is summarized in Table 3. The resulting hyperparameters are used for all DL models in this study.

Table 3: Hyperparameter search space

Hyperparameter	Values
Mini-batch size	$\{2^5, 2^6, 2^7, 2^8, 2^9\}$
Learning rate	$[10^{-4}, 10^{-3}]$
Dropout	$[0, 1]$
Number of neurons per hidden layer	$\{2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}\}$

The DNN architecture comprises three fully connected layers, with the number of neurons per layer chosen by the hyperparameter search. To enhance generalization and reduce overfitting, the network includes dropout and batch normalization. Hidden layers use the Rectified Linear Unit (ReLU) activation for nonlinearity, and the output layer employs a softmax activation to produce class probabilities. Training proceeded for a maximum of $T = 150$ epochs with early stopping based on validation loss. Let $\mathcal{M}_t$ denote the DL model after epoch t and $L_{\text{val}}(t)$ the validation loss; we select

$$t^\star = \arg\min_{t \le T} L_{\text{val}}(t), \qquad \widehat{\mathcal{M}} = \mathcal{M}_{t^\star}.$$

We used patience $p = 10$: training halts if L_{val} fails to improve for 10 consecutive epochs, after which the best checkpoint is restored. The cap of 150 epochs provides adequate headroom while avoiding unnecessary computation; empirical learning curves indicated that MacroF1 gains beyond ~140 epochs were negligible across runs.

DALEX is employed to derive *global explanations* of the DL model, characterizing behavior over the entire dataset rather than at the level of individual predictions. By summarizing feature importance, inspecting model response patterns, and revealing structure-induced effects, DALEX offers a dataset-wide view of the model's decision logic. This perspective helps surface systematic biases, assess alignment with domain knowledge, and validate that salient signals are consistent with practitioner expectations. We integrate the DALEX Python package (v1.2.0) to quantify the global relevance of input features. We create adversarial samples using the FGSM and PGD methods as implemented in the Adversarial Robustness Toolbox (ART) library.

4.3 Results and Discussion

The experimental study is designed with several objectives. First, we evaluate how well the proposed method performs in practice (Section 4.3.1). Second, we conduct an ablation analysis to compare two enhancement strategies for strengthening a DL intrusion detection model: (i) training with CF-augmented samples, and (ii) AT. This comparison targets the overall predictive accuracy of the DL model for detecting attacks.

In addition to performance, we examine how these training strategies influence model behavior and decision logic. To that end, we apply the DALEX XAI framework to study and visualize how CF-Aug training and AT reshape the model's reasoning, and we contrast both against the baseline model $\mathcal{M}_\theta$ trained on the original dataset. This analysis (Section 4.3.2) allows us to quantify not only which approach yields better classification results, but also which approach produces a model whose decisions are more interpretable and more stable.

4.3.1 Analysis of the Performance for the Proposed Method

We perform an ablation study to assess the effectiveness of the proposed approach on two widely used intrusion-detection benchmarks: NSL-KDD and CICIDS2017. Table 4 summarizes the number of training samples per class for the NSL-KDD and CICIDS17 datasets under the three training regimes. In the Baseline setting, the models are trained only on the original data distribution. In the Adversarial-Augmented (Adv-Augmented) setting, each original sample is paired with an adversarial counterpart, effectively doubling the number of samples per class; consequently, this configuration yields the largest training sets for both benchmarks. In contrast, the CF-Aug setting increases the dataset size more selectively: only those samples for which valid counterfactuals can be generated—given the chosen hyperparameters and the counterfactual generator's feasibility constraints—are added to the training pool. As a result, the total number of samples under CF-Augmented lies between the Baseline and Adv-Augmented cases, with a noticeable relative increase for rare classes (e.g., R2L and U2R in NSL-KDD and several minority attack types in CICIDS17). This leads to a more balanced training distribution without incurring the full cost of duplicating the entire dataset.

Table 4: Number of training samples per class under baseline, Adv-augmented, and CF-Aug settings for NSL-KDD and CICIDS17

Dataset	Class	Baseline	Adv-Augmented	CF-Augmented
NSL-KDD	Normal	13,449	26,898	14,155
	DoS	9234	18,468	10,323
	Probe	2289	4578	3589
	R2L	209	418	665
	U2R	11	22	466
Total	–	25,192	50,384	29,198
CICIDS17	Normal	79,288	158,576	85,631
	DoS Hulk	7582	15,164	11,774
	PortScan	7609	15,218	14,006
	DDoS	4551	9102	11,013
	DoS GoldenEye	362	724	4976
	FTP-Patator	191	382	3037
	SSH-Patator	143	286	4937
	DoS Slowloris	191	382	5053
	DoS Slowhttptest	83	166	6896
Total	–	100,000	200,000	147,296

Table 5 reports the per-class F1 scores for NSL-KDD and CICIDS17 across the four training strategies. For NSL-KDD, the results show that all variants perform similarly on the majority classes (Normal and

DoS), with PGD-based AT achieving the best F1 for the Normal class (0.83) and tying with the Baseline and FGSM models on DoS. The benefits of the proposed CF-based augmentation are most evident on the minority classes. For R2L and U2R, CF-Augment yields the highest F1 scores (0.31 and 0.17, respectively), substantially improving over the Baseline (0.22 and 0.12). This indicates that counterfactual samples help the model capture more informative decision boundaries for rare attack types, without sacrificing performance on the frequent classes.

Table 5: F1 per class for NSL-KDD and CICIDS17 datasets. The best results are in bold

Dataset	Class	Baseline	Adv-Training (FGSM)	Adv-Training (PGD)	CF-Aug
NSL-KDD	Normal	0.82	0.81	**0.83**	0.82
	DoS	**0.89**	**0.89**	0.88	0.88
	Probe	**0.78**	0.76	0.77	0.75
	R2L	0.22	0.22	0.27	**0.31**
	U2R	0.12	0.14	0.15	**0.17**
CICIDS17	Normal	0.96	0.96	**0.97**	**0.97**
	DoS Hulk	0.97	**0.98**	**0.98**	**0.98**
	PortScan	0.81	0.84	0.86	**0.87**
	DDoS	**0.99**	0.97	**0.99**	0.98
	DoS GoldenEye	0.36	0.38	**0.44**	0.38
	FTP-Patator	0.70	**0.97**	0.94	0.84
	SSH-Patator	0.53	0.69	0.61	**0.69**
	DoS slowloris	0.21	0.20	0.22	**0.40**
	DoS Slowhttptest	**0.47**	0.17	0.41	0.27

For CICIDS17, all methods attain high F1 on the Normal and DoS Hulk traffic, with PGD and CF-Aug slightly outperforming the others (0.97 for Normal and 0.98 for DoS Hulk). On more challenging or less frequent attack categories, the behavior diverges more clearly. CF-Aug achieves the best F1 for PortScan (0.87) and DoS slowloris (0.40), and matches the highest score for SSH-Patator (0.69), highlighting its effectiveness in improving detection of certain sophisticated or low-frequency attacks. Conversely, AT, especially with FGSM, excels on FTP-Patator (0.97), while PGD-based AT offers the best performance for DoS GoldenEye (0.44). However, on DoS Slowhttptest the Baseline model still yields the highest F1 (0.47), suggesting that adversarial objectives can occasionally distort the decision boundary in ways that hurt specific classes.

Overall, these results suggest a complementary effect between AT and CF-Aug. AT tends to enhance robustness and F1 for several medium-frequency attack types, whereas CF-Aug is particularly beneficial for hard or underrepresented classes, improving their detection without degrading performance on the dominant classes. This supports the claim that counterfactual samples provide targeted, label-preserving variations that help the model better discriminate fine-grained attack behaviors. Fig. 2 compares the overall accuracy (OA), WeightedF1, and MacroF1 obtained on NSL-KDD and CICIDS17 for the four training regimes (Baseline, AT with FGSM, AT with PGD, and CF-Aug). Across both datasets, AT with PGD and the CF-Aug models consistently outperform the Baseline and FGSM-based AT in all three metrics. On NSL-KDD, AT (PGD) and CF-Aug yield the highest OA and WeightedF1, while CF-Aug achieves the best MacroF1, indicating a more balanced treatment of minority classes. A similar trend is observed on CICIDS17, where AT (PGD) and CF-Aug steadily increase OA and WeightedF1 over the Baseline, and both reach the highest MacroF1 values.

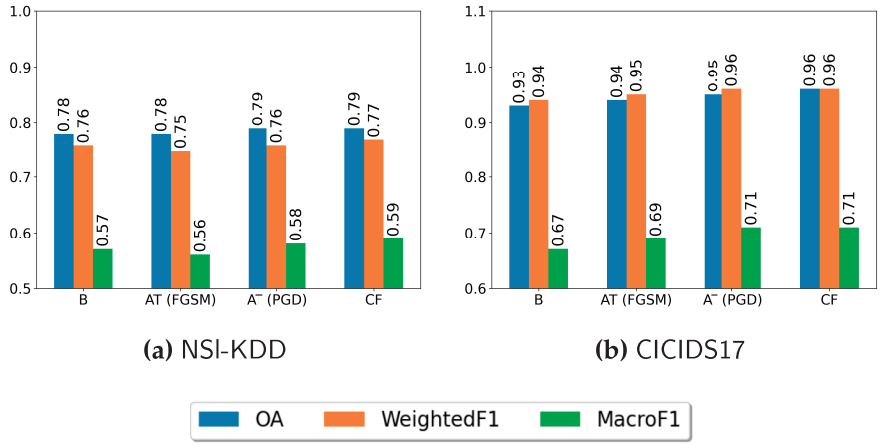


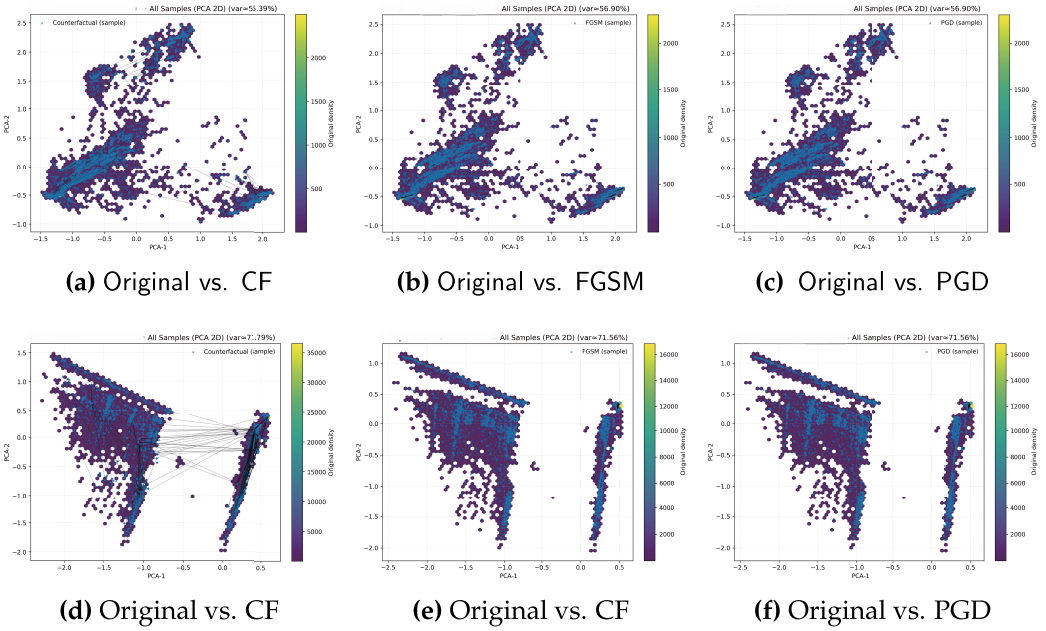
Figure 2: Overall accuracy (OA), WeightedF1, and MacroF1 for NSL-KDD and CICIDS17

The improvement in MacroF1 under CF-Aug is driven primarily by gains on rare classes such as R2L and U2R in NSL-KDD and low-frequency DoS/Probe attacks in CICIDS17 (see Table 5). Because counterfactual samples are generated near the decision boundary, they tend to concentrate around ambiguous regions where minority classes are under-represented, effectively increasing the density of informative training points for these classes. In contrast, AT doubles the dataset size by adding highly correlated adversarial variants of all samples. While this improves OA and WeightedF1, it can also reallocate capacity towards the majority or easier classes, which is reflected in slight drops for some rare categories (e.g., DoS Slowhttptest) despite the overall positive trend. Together, these results suggest that AT (PGD) and CF-Aug provide complementary benefits: AT mainly hardens the decision boundary around frequent behaviors, whereas CF-Aug focuses on minority, near-boundary behaviors, leading to more accurate and more balanced intrusion-detection performance.

All results reported in this work are obtained from a single train/validation/test split with a fixed random seed. While this is common practice in the IDS literature, it does not capture variability due to random initialisation and sampling. A more exhaustive study with multiple independent runs (reporting mean and standard deviation of OA and F1 scores) is an important direction for future work. Nevertheless, we note that the relative behaviour of the Baseline, AT, and CF-Aug models is consistent across our experiments and in line with the qualitative differences highlighted by the explanation analysis.

4.3.2 Analysis of the XAI-Based Strategy

Fig. 3 visualizes the geometric relationship between the original training samples and the different types of synthetic data used in our experiments. Each plot shows a two–dimensional PCA projection of the feature space, with original points overlaid with either CF-based samples or adversarial samples generated by FGSM and PGD.



(a) Original vs. CF **(b)** Original vs. FGSM **(c)** Original vs. PGD

(d) Original vs. CF **(e)** Original vs. CF **(f)** Original vs. PGD

Figure 3: PCA projections of original and synthetic samples for NSL-KDD (top row) and CICIDS17 (bottom row). Each panel shows the original training samples (blue) together with (**a,d**) counterfactual (CF) samples, (**b,e**) FGSM adversarial samples, and (**c,f**) PGD adversarial samples. Axes correspond to the first two principal components computed on the original training data

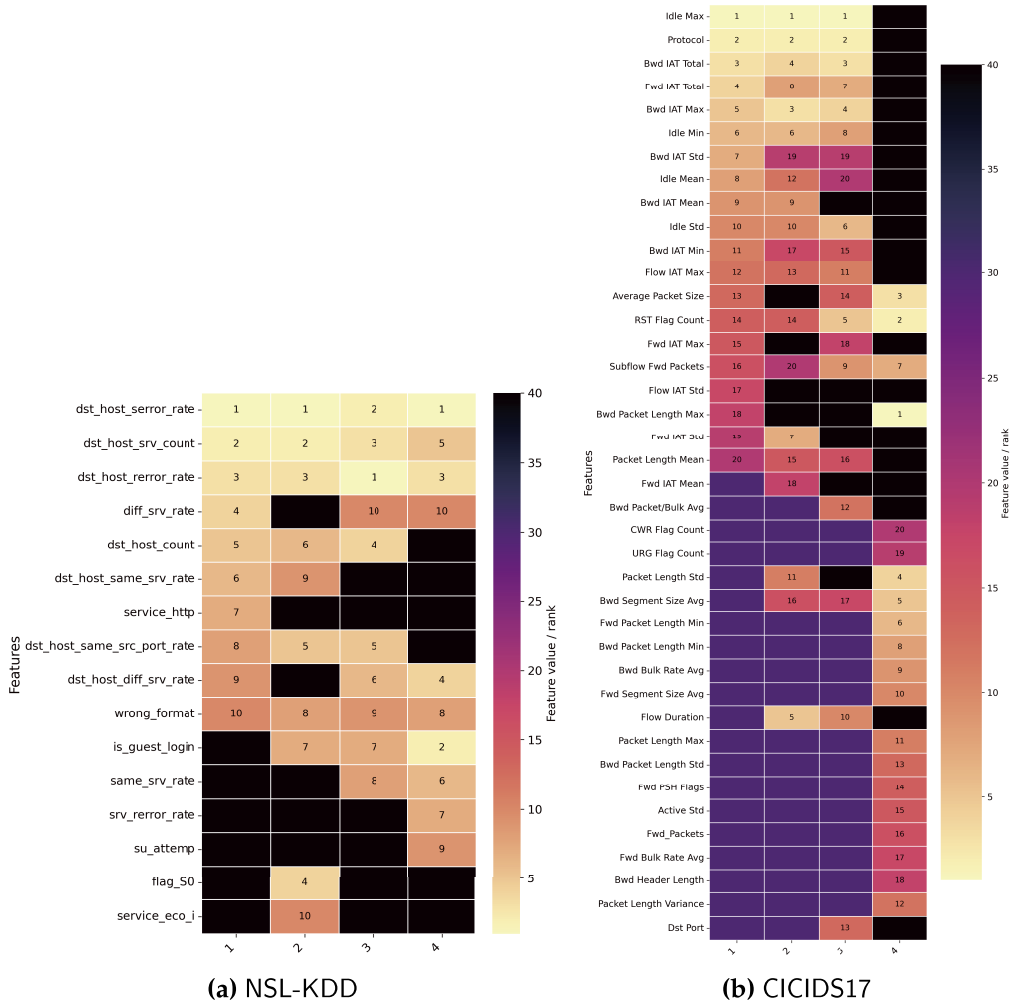
For NSL-KDD (top row, Fig. 3a–c), the CF samples (Fig. 3a) largely trace the same manifold as the original data, forming tight clouds that remain close to the corresponding original points. This indicates that counterfactuals perform local, label-preserving moves in feature space, exploring nearby decision boundaries without drifting far from the data distribution. By contrast, FGSM and especially PGD (Fig. 3b,c) produce more pronounced displacements: the perturbed samples spread further along specific directions of the PCA axes, revealing that adversarial samples tend to push inputs toward regions where the classifier is uncertain or misclassifies, even when those regions lie somewhat off the main data manifold.

A similar pattern is observed for CICIDS17 (bottom row, Fig. 3d–f). The CF samples again stay close to the dense regions of the original distribution, effectively thickening the existing clusters rather than creating new ones. In comparison, adversarial samples generated by FGSM and PGD introduce more extreme shifts and sharper "fringes" around the original clusters, reflecting higher-magnitude, targeted perturbations. Overall, these visualizations support our interpretation that CF-Augment enriches the training data with realistic, near-manifold variations, whereas AT exposes the model to more aggressive worst-case perturbations that deliberately probe the boundaries of the learned decision regions. To complement the qualitative impression from Fig. 3, we compute a simple quantitative measure of displacement in the PCA space. Let $\mathbf{z}_i \in \mathbb{R}^2$ denote the projection of an original sample $\mathbf{x}_i$ onto the first two principal components and $\tilde{\mathbf{z}}_i^{(m)}$ the projection of its corresponding synthetic sample generated by method $m \in \{\text{CF}, \text{FGSM}, \text{PGD}\}$. We define the average PCA displacement for method m as

$$d_m = \frac{1}{N} \sum_{i=1}^{N} \left\| \tilde{\mathbf{z}}_i^{(m)} - \mathbf{z}_i \right\|_2.$$

Across both NSL-KDD and CICIDS17 we obtain $d_{\mathrm{CF}} < d_{\mathrm{FGSM}} < d_{\mathrm{PGD}}$, confirming that counterfactual samples remain closest to their original counterparts in the projected space, whereas PGD-based adversarial samples move farthest away.

Fig. 4 presents heatmaps of the top–20 ranked features for the four models (Baseline, AT(FGSM), AT(PGD), and CF-Aug) on NSL-KDD and CICIDS17. Each row corresponds to a feature and each column to a model; the cell value denotes the rank of that feature within the corresponding model (1 = most important, 20 = least important among the top–20), while empty cells indicate that the feature does not appear in that model's top–20 list.



(a) NSL-KDD **(b)** CICIDS17

Figure 4: DALEX-based ranking of the top 20 features. (1) Baseline, (2) AT (FGSM), (3) AT (PGD), (4) CF-Aug

For CICIDS17 (Fig. 4), the Baseline and both adversarially trained models exhibit very similar importance profiles. Their top–10 is dominated by *Idle Max*, *Protocol*, and *BWD IAT Total*, together with related connection-count features. In contrast, these features drop out of the top–10 for the CF-Aug model, which instead assigns the highest importance to packet- and flow-level descriptors, including *Average Packet Size*, *Packet Length Std*, *Bwd Segment Size Avg*, and *RST Flag Count*. This shift suggests that CF-Aug encourages the detector to rely less on coarse host aggregates—which can be more easily manipulated by changing

traffic mix—and more on fine-grained traffic characteristics that are tightly coupled to the morphology of malicious flows.

On NSL-KDD (Fig. 4), all four models agree on the central role of the *dst_host* family of attributes: *dst_host_serror_rate*, *dst_host_srv_count*, and *dst_host_rerror_rate* consistently appear among the top three features across training regimes. However, the CF-Aug model diverges in the lower portion of the ranking. In addition to these dominant host-level rates, CF-Aug elevates more semantic features such as *is_guest_login*, *srv_rerror_rate*, and *su_attempt*, which are strongly associated with R2L and U2R behaviors. This is consistent with our per-class results, where CF-Aug provides the largest gains for these minority attack classes. Overall, the heatmaps indicate that AT largely preserves the baseline inductive biases, whereas CF-Aug actively reshapes the feature-importance landscape in a way that emphasizes features relevant to underrepresented and harder-to-detect threats.

These differences can be traced back to the way the two augmentation schemes interact with the decision boundary. Counterfactuals are explicitly optimized to flip the prediction with minimal, domain-feasible changes, so for minority classes, they tend to perturb precisely those few semantic attributes that separate R2L/U2R (or rare CICIDS17 attacks) from normal and majority classes. When the model is retrained on $\mathcal{T} \cup \mathcal{C}$, these boundary-probing counterfactuals repeatedly expose the classifier to such semantic cues in ambiguous regions of the feature space, naturally increasing their global importance. In contrast, FGSM/PGD adversarial samples are generated by following the loss gradient around all training points within an ℓ_∞ factor; they do not target specific minority decision fronts, and they primarily nudge the model to stabilize its existing predictions without substantially reweighting, which features those predictions depend on. As a result, AT tends to maintain the baseline inductive biases visible in the heatmaps, whereas CF-Aug shifts the model toward features that are most informative for underrepresented attack classes. A systematic comparison of our DALEX-based analysis with alternative XAI methods, such as SHAP and LIME, is left as future work.

5 Conclusion

In this paper, we examined how two complementary data-centric strategies, AT and CF-Aug, reshape the accuracy and interpretability of DL-based IDS. Using NSL-KDD and CICIDS17 as benchmark datasets, we instantiated a common experimental pipeline in which a baseline DNN was first trained on the original data, and then contrasted with variants retrained on adversarially generated samples (FGSM- and PGD-based) and on label-preserving counterfactual samples. This design allowed us to disentangle the effects of boundary hardening via adversarial perturbations from those of boundary sculpting via counterfactual explanations, and to assess their impact on predictive performance. The empirical results indicate that AT and CF-Aug consistently improve the overall performance compared to the baseline in terms of multiple metrics: OA, WeightedF1, and MacroF1. The explainability analysis further elucidates how these training regimes alter model behavior. Global feature-importance profiles indicate that both AT and CF-Aug can shift emphasis away from brittle or spurious signals toward more stable, domain-relevant features. However, CF-Aug tends to produce models whose importance rankings are more stable under perturbations and more closely aligned with security intuition, while AT can occasionally distort the decision surface for specific classes despite improving aggregate robustness. Taken together, quantitative metrics and explanation profiles underscore that no single strategy universally outperforms others; rather, AT and CF-Aug provide complementary benefits that can be chosen or combined depending on whether the primary objective is worst-case robustness, minority-class detection, or analyst-facing interpretability. A number of avenues for future work arise naturally from this study. On the robustness side, extending the evaluation to broader threat models, including black-box, transfer, and problem-space attacks, would provide a more complete picture

of the protection afforded by AT and CF-Aug. On the explainability side, taking global analysis, combining it with local explanations, and adding causal constraints might yield richer, action-oriented insights for Security Operations Center workflows.

Acknowledgement: Not applicable.

Funding Statement: Not applicable.

Author Contributions: Conceptualization: Malik Al-Essa, Mohammad Qatawnweh, and Orieb Abualghanam; Methodology: Malik Al-Essa, Mohammad Qatawneh, Ahmad Sami Al-Shamayleh, Orieb Abualghanam, and Wesam Almobaideen; Software: Malik Al-Essa, Orieb Abualghanam, and Wesam Almobaideen; Validation: Malik Al-Essa, Mohammad Qatawneh, Ahmad Sami Al-Shamayleh, Orieb Abualghanam, and Wesam Almobaideen; Formal analysis: Malik Al-Essa, Mohammad Qatawneh, Ahmad Sami Al-Shamayleh, Orieb Abualghanam, and Wesam Almobaideen; Investigation: Malik Al-Essa, and Mohammad Qatawnweh; Resources: Malik Al-Essa, Mohammad Qatawneh, Ahmad Sami Al-Shamayleh, Orieb Abualghanam, and Wesam Almobaideen; Data curation: Malik Al-Essa, and Orieb Abualghanam; Writing—original draft preparation: Malik Al-Essa, Mohammad Qatawneh, Ahmad Sami Al-Shamayleh, Orieb Abualghanam, and Wesam Almobaideen; Writing—review and editing: Malik Al-Essa, Mohammad Qatawneh, Ahmad Sami Al-Shamayleh, Orieb Abualghanam, and Wesam Almobaideen; Visualization: Malik Al-Essa, Mohammad Qatawneh, and Ahmad Sami Al-Shamayleh; Supervision: Wesam Almobaideen, and Mohammad Qatawneh; Project administration: Wesam Almobaideen, and Mohammad Qatawneh. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Data is openly available in a public repository.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Malik J, Akhunzada A, Al-Shamayleh AS, Zeadally S, Almogren A. Hybrid deep learning based threat intelligence framework for industrial IoT systems. J Ind Inf Integr. 2025;45(1):100846. doi:10.1016/j.jii.2025.100846.
2. Assmi H, Guezzaz A, Benkirane S, Azrour M, Jabbour S, Innab N, et al. A robust security detection strategy for next generation IoT networks. Comput Mater Contin. 2025;82(1):443–66. doi:10.32604/cmc.2024.059047.
3. Ghazal TM, Hasan MK, Raju KN, Khan MA, Alshamayleh A, Bhatt MW, et al. Data space privacy model with federated learning technique for securing IoT communications in autonomous marine vehicles. J Intell Robotic Syst. 2025;111(3):84. doi:10.1007/s10846-025-02298-1.
4. Al-Essa M, Appice A. Dealing with imbalanced data in multi-class network intrusion detection systems using xgboost. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham, Switzerland: Springer; 2021. p. 5–21.
5. Ali WA, Roccotelli M, Boggia G, Fanti MP. Intrusion detection system for vehicular ad hoc network attacks based on machine learning techniques. Inf Secur J Global Perspect. 2024;33(6):659–77. doi:10.1080/19393555.2024.2307638.
6. Xu J, Wang Y, Chen H, Shen Z. Adversarial machine learning in cybersecurity: attacks and defenses. Int J Manag Sci Res. 2025;8(2):26–33. doi:10.53469/ijomsr.2025.08(02).04.
7. Asha S, Vinod P. Evaluation of adversarial machine learning tools for securing AI systems. Clust Comput. 2022;25(1):503–22. doi:10.1007/s10586-021-03421-1.
8. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proceedings of the 3rd International Conference on Learning Representations; 2015 May 7–9; San Diego, CA, USA. p. 1–11.

9. Andriushchenko M, Flammarion N. Understanding and improving fast adversarial training. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 16048–59.
10. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, et al. Intriguing properties of neural networks. In: Proceedings of the 2nd International Conference on Learning Representations; 2014 Apr 14–16; Banff, AB, Canada. p. 1–10.
11. Bai T, Luo J, Zhao J, Wen B, Wang Q. Recent advances in adversarial training for adversarial robustness. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Survey Track; 2021 Aug 19–26; Virtual. p. 4312–21. doi:10.24963/ijcai.2021/591.
12. Alhajjar E, Maxwell P, Bastian N. Adversarial machine learning in Network Intrusion Detection Systems. Expert Syst Appl. 2021;186(2):115782. doi:10.1016/j.eswa.2021.115782.
13. Pierazzi F, Pendlebury F, Cortellazzi J, Cavallaro L. Intriguing properties of adversarial ML attacks in the problem space. In: Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP); 2020 May 18–21; San Francisco, CA, USA. p. 1332–49.
14. Yin C, Zhu Y, Liu S, Fei J, Zhang H. Enhancing network intrusion detection classifiers using supervised adversarial training. J Supercomput. 2020;76(9):6690–719. doi:10.1007/s11227-019-03092-1.
15. Wang C, Zhang L, Zhao K, Ding X, Wang X. AdvAndMal: adversarial training for android malware detection and family classification. Symmetry. 2021;13(6):1081. doi:10.3390/sym13061081.
16. Anthi E, Williams L, Rhode M, Burnap P, Wedgbury A. Adversarial attacks on machine learning cybersecurity defences in industrial control systems. J Inf Secur Appl. 2021;58(8):102717. doi:10.1016/j.jisa.2020.102717.
17. Khamis RA, Matrawy A. Evaluation of adversarial training on different types of neural networks in deep learning-based IDSs. In: Proceedings of the 2020 International Symposium on Networks, Computers and Communications (ISNCC); 2020 Oct 20–22; Montreal, QC, Canada. p. 1–6.
18. Warnecke A, Arp D, Wressnegger C, Rieck K. Evaluating explanation methods for deep learning in security. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS&P); 2020 Sep 7–11; Genoa, Italy. p. 158–74.
19. Wang M, Zheng K, Yang Y, Wang X. An explainable machine learning framework for intrusion detection systems. IEEE Access. 2020;8:73127–41. doi:10.1109/access.2020.2988359.
20. Sangeetha SKB, A NB. Intrumer: a multi module distributed explainable IDS/IPS for securing cloud environment. Comput Mater Contin. 2025;82(1):579–607. doi:10.32604/cmc.2024.059805.
21. Marino DL, Wickramasinghe CS, Manic M. An adversarial approach for explainable AI in intrusion detection systems. In: Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society, IECON 2018; 2018 Oct 21–23; Washington, DC, USA. p. 3237–43.
22. Malik AE, Andresini G, Appice A, Malerba D. An XAI-based adversarial training approach for cyber-threat detection. In: Proceedings of the 2022 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing; 2022 Sep 12–15; Falerna, Italy. p. 1–8.
23. Kuppa A, Le-Khac NA. Adversarial XAI methods in cybersecurity. IEEE Trans Inf Forensics Secur. 2021;16:4924–38. doi:10.1109/tifs.2021.3117075.
24. Pawlicki M, Pawlicka A, Kozik R, Choraś M. Explainability vs. security: the unintended consequences of xAI in cybersecurity. In: Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems; 2024 Jul 2–20; Singapore. p. 1–7.
25. Shah SHA, Akhtar LH, Ali MN, Kim BS. Golden jackal driven optimization for a transparent and interpretable intrusion detection system using explainable AI to revolutionize cybersecurity. Egypt Inform J. 2025;32(1):100837. doi:10.1016/j.eij.2025.100837.
26. Ali MN, Imran M, Bahoo G, Kim BS. DoS attack detection enhancement in autonomous vehicle systems with explainable AI. In: Proceedings of the 2024 IEEE International Conference on Future Machine Learning and Data Science (FMLDS); 2024 Nov 20–23; Sydney, NSW, Australia. p. 228–33.

27. Wang J, Chang X, Wang Y, Rodríguez RJ, Zhang J. LSGAN-AT: enhancing malware detector robustness against adversarial examples. Cybersecurity. 2021;4(1):1–15. doi:10.1186/s42400-021-00102-9.
28. Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications; 2009 Jul 8–10; Ottawa, ON, Canada. p. 1–6.
29. Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018); 2018 Jan 22–24; Funchal, Madeira, Portugal. p. 108–16.
30. Engelen G, Rimmer V, Joosen W. Troubleshooting an intrusion detection dataset: the CICIDS2017 case study. In: Proceedings of the 6th IEEE European Symposium on Security and Privacy Workshops, EuroS&PW 2021; 2021 May 27; San Francisco, CA, USA. p. 7–12.
31. Andresini G, Pendlebury F, Pierazzi F, Loglisci C, Appice A, Cavallaro L. INSOMNIA: towards concept-drift robustness in network intrusion detection. In: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security; 2021 Nov 15; Virtual Event, Republic of Korea. p. 111–22.