future internet

MDPI

Article

Adversarial Robustness Evaluation for Multi-View Deep Learning Cybersecurity Anomaly Detection

Min Li [1,2], Yuansong Qiao [1] and Brian Lee [1,*]

[1] Software Research Institute, Technological University of the Shannon, Midlands Midwest, University Road, N37 HD68 Athlone, Ireland; a00278508@student.tus.ie (M.L.); yuansong.qiao@tus.ie (Y.Q.)
[2] School of Artificial Intelligence, Jingchu University of Technology, No.33 Xiangshan Road, Jingmen 448000, China
* Correspondence: brian.lee@tus.ie

Abstract

In the evolving cyberthreat landscape, a critical challenge for intrusion detection systems (IDSs) lies in defending against meticulously crafted adversarial attacks. Traditional single-view detection frameworks, constrained by their reliance on limited and unidimensional feature representations, are often inadequate for identifying maliciously manipulated samples. To address these limitations, this study proposes a key hypothesis: a detection architecture that adopts a multi-view fusion strategy can significantly enhance the system's resilience to attacks. To validate the proposed hypothesis, this study developed a multi-view fusion architecture and conducted a series of comparative experiments. A two-pronged validation framework was employed. First, we examined whether the multi-view fusion model demonstrates superior robustness compared to a single-view model in intrusion detection tasks, thereby providing empirical evidence for the effectiveness of multi-view strategies. Second, we evaluated the generalization capability of the multi-view model under varying levels of attack intensity and coverage, assessing its stability in complex adversarial scenarios. Methodologically, a dual-axis training assessment scheme was introduced, comprising (i) continuous gradient testing of perturbation intensity, with the ε parameter increasing from 0.01 to 0.2, and (ii) variation in attack density, with sample contamination rates ranging from 80% to 90%. Adversarial test samples were generated using the Fast Gradient Sign Method (FGSM) on the TON_IoT and UNSW-NB15 datasets. Furthermore, we propose a validation mechanism that integrates both performance and robustness testing. The model is evaluated on clean and adversarial test sets, respectively. By analyzing performance retention and adversarial robustness, we provide a comprehensive assessment of the stability of the multi-view model under varying evaluation conditions. The experimental results provide clear support for the research hypothesis: The multi-view fusion model is more robust than the single-view model under adversarial scenarios. Even under high-intensity attack scenarios, the multi-view model consistently demonstrates superior robustness and stability. More importantly, the multi-view model, through its architectural feature diversity, effectively resists targeted attacks to which the single-view model is vulnerable, confirming the critical role of feature space redundancy in enhancing adversarial robustness.

Keywords: cybersecurity anomaly detection; multi-view fusion; adversarial robustness

check for updates

Academic Editor: Ugo Fiore

Received: 7 September 2025
Revised: 29 September 2025
Accepted: 4 October 2025
Published: 8 October 2025

Citation: Li, M.; Qiao, Y.; Lee, B. Adversarial Robustness Evaluation for Multi-View Deep Learning Cybersecurity Anomaly Detection. *Future Internet* **2025**, *17*, 459. https://doi.org/10.3390/fi17100459

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the rapid expansion of networked applications, both the sophistication and magnitude of cyberattacks have escalated. The ongoing evolution of security threats in cyberspace essentially reflects a dynamic adversarial process between defensive strategies and offensive techniques. Intrusion detection systems (IDSs), as a fundamental component of network security, are instrumental in defending against such threats by continuously monitoring network traffic, identifying anomalous activities, and issuing timely alerts for potential security breaches. Nevertheless, conventional rule-based IDSs—which depend on static, expert-defined rule sets—are plagued by limited adaptability, particularly when confronted with zero-day exploits and complex attack variants. In response to these limitations, research attention has increasingly shifted toward leveraging machine learning (ML), especially deep learning (DL) approaches, which have demonstrated promise in enabling more intelligent and adaptive IDS solutions.

DL leverages multi-layer neural network architectures to autonomously extract intricate features from high-dimensional network traffic data [1], thereby enhancing detection accuracy and efficiency relative to conventional ML techniques. For instance, generative adversarial networks (GANs) can mitigate data imbalance by generating high-fidelity synthetic samples to enrich the training set [2]. Convolutional neural networks (CNNs) are adept at capturing hierarchical structured features from traffic data, while long short-term memory networks (LSTMs) excel at modeling temporal dependencies [3]. Additionally, autoencoders (AEs) facilitate the unsupervised identification of latent anomalies within network traffic [4]. Collectively, these advances have significantly bolstered the adaptability of IDSs in complex and evolving environments. However, the "black box" nature of DL models exposes their vulnerabilities in adversarial environments [5]. Adversaries can exploit this opacity by crafting adversarial samples through subtle perturbations, which can mislead models and result in detection failures. The landscape of adversarial attacks continues to diversify, with the shared objective of evading detection mechanisms while preserving attack stealthiness. Notably, traditional machine learning evaluation protocols primarily emphasize model generalization on independent and identically distributed (IID) data, neglecting robustness. As a result, enhancing the adversarial robustness of IDS models—without compromising detection performance—has become a core issue in the cybersecurity field.

Nevertheless, this uni-modal representation inherently lacks redundancy, rendering such models susceptible to adversarial manipulations and targeted attacks focused on specific feature domains. By contrast, multi-view learning integrates multiple feature representations, supplying the model with greater informational redundancy and diversity, which in theory contributes to improved adversarial robustness [6]. Within intrusion detection applications, adversaries may circumvent detection in a single view, whereas aggregating information from multiple perspectives enables the uncovering of concealed malicious activity. Despite these theoretical benefits, the robustness of multi-view fusion models under varying intensities and scopes of adversarial attacks remains an open research question that requires further investigation.

In adversarial machine learning, early studies showed that deep models are highly sensitive to even small changes in high-dimensional space, making them vulnerable in security settings (Szegedy et al. [7]; Goodfellow et al. [8]). More recently, Alotaibi et al. [9] provided a systematic review of adversarial attacks on intrusion detection systems (IDSs), outlining the real challenges these attacks create for network security and summarizing the main defense strategies. Taken together, these studies point to the importance of testing how robust IDSs are under different attack scenarios and suggest that combining multiple perspectives may help strengthen their resilience against adversarial threats.

Building on the aforementioned context, this study formulates a hypothesis: multi-view fusion architectures can enhance the adversarial robustness of IDSs by leveraging feature diversity at the model level. To test this hypothesis, we constructed a comprehensive and hierarchical evaluation framework. Using the TON_IoT and UNSW-NB15 datasets as benchmarks, adversarial samples were crafted with the FGSM. This study assessed the performance of both single-view and multi-view models under benign and adversarial conditions, incorporating varying levels of perturbation strength (ε ranging from 0.01 to 0.2) and attack coverage (80% and 90%). This work specifically addresses key research questions: is the adversarial robustness of multi-view architectures universally greater than that of single-view counterparts across different attack scenarios? Through multi-faceted experimental investigations, this study elucidates the underlying mechanisms contributing to the robustness of multi-view learning.

The main contributions of this paper are summarized as follows:

(1) **Development of a Multi-Dimensional, Multi-Level Adversarial Robustness Evaluation Framework:** This study introduces an innovative evaluation framework designed to systematically assess the adversarial robustness of deep learning models for IDSs. The proposed framework integrates multiple dimensions—including perturbation strength (ε ranging from 0.01 to 0.2) and attack coverage (80% and 90%)—utilizing the TON_IoT and UNSW-NB15 datasets as benchmarks. Unlike conventional evaluation approaches that typically examine a single attack intensity or a fixed proportion of adversarial samples, this framework allows for a more comprehensive evaluation of intrusion detection model robustness across different levels of attack intensity and coverage, thereby offering a more reliable foundation for practical IDS deployment and security assessment.

(2) **Multi-Level Validation of the Robustness of Multi-View DL Models:** While prior studies have predominantly concentrated on enhancing the adversarial robustness of single-view models, this work introduces a novel multi-view fusion strategy to validate the robustness advantages of multi-view architectures in adversarial settings. Experimental findings reveal that by aggregating complementary information from feature spaces, multi-view models demonstrate a marked ability to withstand targeted attacks that compromise single-view models. This integration not only bolsters the resistance of intrusion detection systems to adversarial threats but also enhances their stability and generalization performance across diverse threat scenarios.

(3) **Two-Dimensional Adversarial Training Strategy:** This study presents a two-dimensional adversarial training framework that jointly modulates attack intensity (via progressive adjustment of ε) and attack coverage (by varying the proportion of adversarial samples) during model training, thereby constructing a more challenging and diverse augmented training set. This strategy strengthens the model's resilience to known adversarial patterns. In contrast to conventional adversarial training—which typically targets a single perturbation strength—this method achieves broader and more representative coverage of adversarial examples, equipping the model to better adapt to the complexity and diversity of real-world adversarial environments.

(4) **Dual Test for Performance and Robustness:** In the testing phase, this study adopts a dual verification mechanism that assesses model performance on both clean and adversarial test sets. By analyzing both performance retention and adversarial robustness, the stability of multi-view models across different evaluation environments is thoroughly examined. This approach effectively overcomes the limitations of conventional evaluation methods that overlook adversarial scenarios, thereby offer-

ing a rigorous and systematic methodology for the validation of reliable intrusion detection systems.

The remainder of this paper is organized as follows. Section 2 provides background and reviews relevant work in the field. Section 3 elaborates on our proposed methodology in detail, describing the frameworks and principles. Section 4 outlines the preparation steps for our experiments, briefly describing the TON_IoT and UNSW-NB15 datasets used in this study, as well as the data preprocessing performed to complete the experimental evaluation. Section 5 details the experimental setup. Finally, Section 6 provides a detailed analysis of the experimental comparison results and introduces the application and significance of the research results to networks.

2. Related Work

IDSs serve as the cornerstone of network security defense. In recent years, ML and DL technologies have been shown to enhance the detection capabilities of IDSs. However, traditional ML methods, which rely heavily on manual feature selection and engineering, exhibit limitations in terms of efficiency and scalability, particularly when processing high-dimensional data . Issa et al. and Shahriar et al. highlighted that such reliance on feature engineering hampers the effectiveness of ML-based IDSs in complex scenarios [10,11]. Experimental evidence further supports the superior detection accuracy of deep learning models in handling sophisticated attack patterns. Despite these advances, the robustness of deep learning-based IDSs in real-world network environments remains insufficiently validated, especially in the face of adversarial attacks. Zhang et al. underscored the necessity for further research into the adversarial robustness of deep learning models deployed in IDSs [12]. So, improving the adversarial robustness of deep models has emerged as a key research focus and has garnered significant attention in recent years [13]. Sicong Han et al. provided a comprehensive review of adversarial examples in deep learning, examining both their existence and the challenges they pose to model robustness [14].

Adversarial attacks compromise DL models by introducing subtle perturbations into input data, leading to misclassification and reduced detection performance—a major challenge in IDS research [9]. The FGSM, a widely used white-box attack technique, serves as a standard tool for evaluating IDS robustness. Sauka et al. explored the use of FGSM in IDSs, generating adversarial samples via gradient-based methods to assess the robustness of deep learning models [15]. Similarly, Liu et al. highlighted FGSM's efficiency and prevalence in white-box robustness evaluations in their review [16]. In addition to white-box scenarios, black-box attacks represent an even greater threat to IDSs, as they more closely reflect real-world adversarial conditions—where attackers possess limited knowledge of model internals. Alotaibi et al. discussed the practical risks posed by black-box attacks, emphasizing their significance for IDSs deployed in real network environments [9]. Yandong Li et al. further proposed a black-box adversarial attack algorithm capable of effectively compromising both standard and defended deep neural network models, demonstrating considerable versatility and threat in security-critical applications such as IDSs [17]. Naturally, the present study focuses mainly on white-box settings. In future research, we intend to further assess model robustness under black-box attack scenarios.

To counter adversarial attacks, various robustness enhancement strategies have been proposed. Among these, adversarial training has emerged as a mainstream approach. Liu et al. demonstrated that incorporating adversarial examples into the training data can significantly improve model stability when faced with perturbations [16]. Shahriar et al. explored the application of data augmentation techniques in IDSs and found that approaches such as Gaussian augmentation can enhance detection performance by increasing data diversity. However, they also noted that current robustness enhancement methods remain

limited in their ability to defend against diverse and complex attack scenarios, often exhibiting unstable performance [11]. Furthermore, Awad et al. highlighted that existing studies rarely address robustness evaluation in multi-dimensional attack settings, thereby constraining the generalizability and real-world applicability of these methods [18].

This study centers on multi-view learning. By integrating multi-source feature representations, multi-view learning enhances feature redundancy and has demonstrated notable advantages in adversarial robustness within domains such as image recognition and natural language processing. However, research on the robustness of multi-view learning under adversarial attacks remains limited, with most studies concentrated on image or vision-related applications [19]. In the context of intrusion detection systems (IDSs), some works have investigated the fusion of multi-modal traffic features to improve detection accuracy. For instance, Asmaa Halbouni et al. introduced a CNN-LSTM hybrid model that combines temporal and spatial features, significantly enhancing the adaptability and detection accuracy of IDSs in complex network environments [20]. Nevertheless, most IDS-related studies primarily emphasize performance optimization in benign conditions and lack a systematic evaluation of the robustness of multi-view models under adversarial settings. This study addresses this gap by experimentally validating the robustness advantages of multi-view fusion strategies in adversarial attack scenarios.

3. Methodology

3.1. Overview

This study aims to develop a comprehensive adversarial robustness assessment for both multi-view and single-view models, focusing on the evaluation of multi-view fusion model robustness in adversarial environments within IDSs. The research objectives are twofold: (i) to verify whether multi-view fusion models demonstrate superior robustness in intrusion detection tasks, thereby providing empirical evidence for the efficacy of multi-view fusion strategies; (ii) to examine the generalizability of multi-view models across varying attack intensities and coverage, assessing their stability and reliability in complex adversarial scenarios.

To address these objectives, this study establishes a multi-dimensional, multi-layered adversarial robustness evaluation framework, comprising the following core methodologies:

Selection of Baseline Models: In this study, we selected the following benchmark models:

Single-view models: Developed using different neural architectures based on a single data perspective. These models serve as baselines to compare the effectiveness and robustness advantages of multi-view fusion.

Zero-perturbation and zero-adversarial-example models: Trained without the inclusion of adversarial samples or perturbations, these models are used to evaluate the stability and reliability of multi-view architectures under complex adversarial conditions [21].

Robustness Evaluation of the Multi-View Fusion Strategy: The adversarial robustness of single-view and multi-view models is systematically evaluated using FGSM-generated samples, enabling a comprehensive evaluation of the robustness benefits provided by multi-view fusion.

Enhanced Dual-Dimensional Adversarial Training Framework: In the training phase, this study constructs a comprehensive adversarial training environment leveraging a dual-dimensional strategy. Specifically, the framework simultaneously controls two key factors: the "intensity" of attacks and the "breadth" of adversarial exposure.

Attack intensity: Progressive adversarial training is employed to address attack intensity, where FGSM-generated adversarial samples with varying perturbation strengths (ε values from 0.01 to 0.2) are systematically integrated with clean samples in the training set.

Attack breadth: The breadth of adversarial exposure is controlled by varying the proportion of adversarial samples within the training data (e.g., 80% or 90%), which effectively simulates different levels of adversarial coverage that may be encountered in real-world scenarios.

By jointly varying both dimensions, the framework constructs a challenging and diverse training environment. This enables a more comprehensive assessment of the model's generalization ability and adversarial robustness, ensuring that the trained model is resilient across a wide spectrum of attack intensities and coverages.

Dual-Test Environment Verification Strategy: In the test phase, two complementary evaluation environments are established: clean data testing and adversarial sample testing. Based on the trained models with adversarial defense capabilities, a dual verification strategy is applied. First, model performance is evaluated on the clean test set to assess its performance retention ability, that is, to detect whether the model retains the ability to recognize normal input after being trained with adversarial samples. Then, the model's adversarial robustness is tested using a test set composed of adversarial samples generated with a fixed perturbation strength. This dual evaluation mechanism enables simultaneous assessment of both the model's general detection performance and its robustness under adversarial conditions.

To further validate the universality and generalizability of the conclusions, experiments were conducted on both the TON_IoT and UNSW-NB15 datasets. Following the third and fourth evaluation strategies described earlier, both single-view and multi-view models were deployed to comprehensively evaluate the robustness advantages of the proposed multi-view fusion strategy.

3.2. Model Design

In this section, we introduce the design structure of the model used in the experiment.

To assess the robustness advantages of multi-view models in intrusion detection tasks, this study establishes a comparative experimental framework encompassing both single-view and multi-view paradigms. The evaluated models include the following: (i) single-view baseline models (autoencoder [AE], convolutional neural network [CNN]); (ii) multi-view fusion models (multi-view AE, multi-view CNN); and (iii) Deep Generalized Canonical Correlation Analysis (DGCCA) model. As illustrated in Figure 1, the overall architecture is organized into three layers.

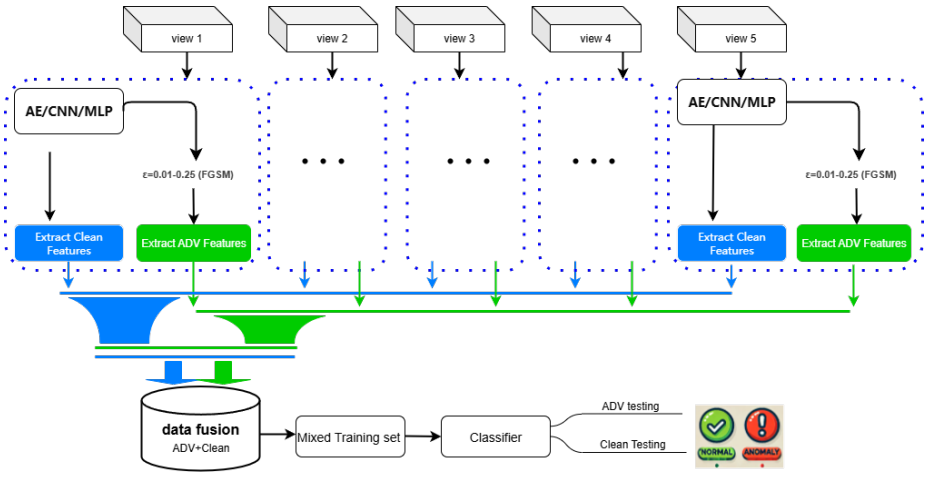

Figure 1. Model framework (MV AE/MV CNN/DGCCA).

(1) **View Layer**
This layer consists of five distinct data views (View 1 to View 5), each representing a feature space. Each view independently processes the raw data using its feature extraction module—such as an AE, CNN, or multi-layer perceptron (MLP)—to extract both clean and adversarial features from the raw input. For single-view models, a single extraction module processes the entire dataset.

(2) **Adversarial Sample Generation and Feature Extraction**
Perturbation Application: During training, adversarial perturbations are introduced to the input data to create adversarial samples. The Fast Gradient Sign Method (FGSM) is adopted in this study as the adversarial attack approach. The perturbation is defined as follows [19]:

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)), \tag{1}$$

where $\mathbf{x}$ denotes the original input sample, y represents the corresponding label, J is the model loss function, and ϵ is the perturbation magnitude parameter that controls the deviation between the adversarial sample and the original input.
Feature Extraction: For each view, both clean and adversarial data are processed by the corresponding AE, CNN, or MLP modules to generate clean features and adversarial features, respectively.

(3) **Feature Fusion Layer**
In the multi-view model, clean and adversarial features from each view are combined by adjusting the proportion of adversarial examples in the training set (80% or 90%), thereby constructing training environments with different levels of attack coverage.

(4) **Classification Layer**
The fused features are fed into a unified classifier to perform the intrusion detection task. The performance of single-view models (AE, CNN, MLP) and multi-view models (AE, CNN, DGCCA) is systematically compared to evaluate the benefits of multi-view fusion in terms of feature complementarity and adversarial robustness.

Algorithms 1–3 describe in detail the algorithm implementation based on multi-view AE, CNN, and DGCCA, respectively.

Algorithm 1 Multi-view AE with FGSM-based Robust Training

Require: Preprocessed dataset D with label column `label`, subset sizes S, encoding dims E, FGSM perturbation strengths $\varepsilon \in \{0.01, 0.03, ..., 0.2\}$
Ensure: Average F1 score, ROC-AUC, PR-AUC per ε
1. Remove non-feature columns, impute missing values with mean, normalize with MinMaxScaler.
2. Stratified split D into training/validation/test sets with ratio 8:1:1.
3. Partition features into 5 views according to dataset-specific grouping rule Φ_d.
4. For each view, train shallow AutoEncoder and extract encoded features.
5. Generate adversarial samples using FGSM on each view:
$X_{adv} = X + \varepsilon \cdot \text{sign}(\nabla_X L(X, \hat{X}))$, $X_{adv} \in [0, 1]$.
6. Concatenate all view encodings into X_{enc} (clean) and X_{adv} (adversarial).
7. Construct mixed training set by replacing 80% of clean samples with adversarial ones.
8. Train MLP classifier with 1 hidden layer on fused features.
9. Tune threshold t^* on validation set to maximize F1 score.
10. Evaluate classifier on test set using F1 score, ROC-AUC, and PR-AUC.
11. For each ε, repeat steps 5–10 with 30 random seeds (30 runs) and report results as mean ± standard deviation.
12. **Return** average metrics (F1, ROC-AUC, PR-AUC) for each ε.

Algorithm 2 Multi-view CNN with FGSM-Adversarial Training

Require: Preprocessed dataset D, subset sizes S, perturbations ε, FGSM ratio α
Ensure: F1 score, ROC-AUC, PR-AUC for each ε
1. Clean missing values, normalize features, and split D into train/val/test sets.
2. Divide features into 5 subsets
3. For each subset, compute reshape shape (h, w) to convert vectors into 2D tensors.
4. For each $\varepsilon \in \{0.01, 0.03, ..., 0.2\}$ and each run:
   4.1 For each view v_i:
      Reshape $X_{train}^{v_i}$ to $(n, h_i, w_i, 1)$.
      Train small CNN to perform binary classification.
      Generate FGSM adversarial samples $X_{adv}^{v_i}$:
         Function GENERATE_FGSM_SAMPLES($model$, X, y, ε):
            Compute $\nabla_X \mathcal{L}(y, f(X))$ and apply $X_{adv} = X + \varepsilon \cdot \text{sign}(\nabla_X)$
            Return clipped X_{adv} to $[0, 1]$
         End Function
      Create mixed training set with α ratio adversarial samples.
      Store reshaped and mixed inputs for final model.
   4.2 Build shared multi-input CNN model:
      Each input $\rightarrow$ CNN branch $\rightarrow$ concatenated embedding.
      Fully-connected layers for binary prediction.
   4.3 Train on multi-view inputs using binary cross-entropy loss.
   4.4 On validation set, find threshold t^* that maximizes F1 score.
   4.5 Evaluate on test set: compute F1, ROC-AUC, and PR-AUC.
5. Repeat steps 1–4 with 30 different random seeds (30 runs), and report results as mean ± standard deviation.
6. **Return** performance metrics per ε.

Algorithm 3 DGCCA with FGSM-Adversarial Training

Require: Multi-view data $\{X^{(v)}\}_{v=1}^{V}$, label vectors Y,
Network config $(layer_sizes, outdim)$, ε values, adv ratio α
Ensure: Performance metrics (F1, ROC-AUC, PR-AUC) for each ε
1. Preprocess dataset: missing value imputation, normalization, stratified split.
2. Partition features into V views (5 views).
3. For each $\varepsilon \in \{0.01, ..., 0.2\}$:
   For each run (total r times):
      Initialize DGCCA model $\mathcal{M}$ with V MLP subnets and CCA loss.
      Generate FGSM adversarial views:
         Function FGSM_ATTACK($x, \nabla_x L, \varepsilon$): return $x + \varepsilon \cdot \text{sign}(\nabla_x L)$
         Select α% of samples and perturb each view with FGSM gradient.
         Combine clean + perturbed views as training input.
      Train $\mathcal{M}$ with adversarial data to minimize CCA loss.
      If enabled, apply post-training linear GCCA to obtain projections $\{U^{(v)}\}$.
      On validation set, evaluate using:
         Majority vote of view outputs $\rightarrow$ predicted label.
         Compute F1 score, ROC-AUC, PR-AUC.
      Record scores.
4. Report metrics as mean ± standard deviation over 30 runs for each ε.
5. **Return**: Performance trends across ε values.

3.3. Adversarial Training Pipeline

This section outlines the adversarial training process adopted for the various models. As illustrated in Figure 2, the process is structured into three distinct stages:

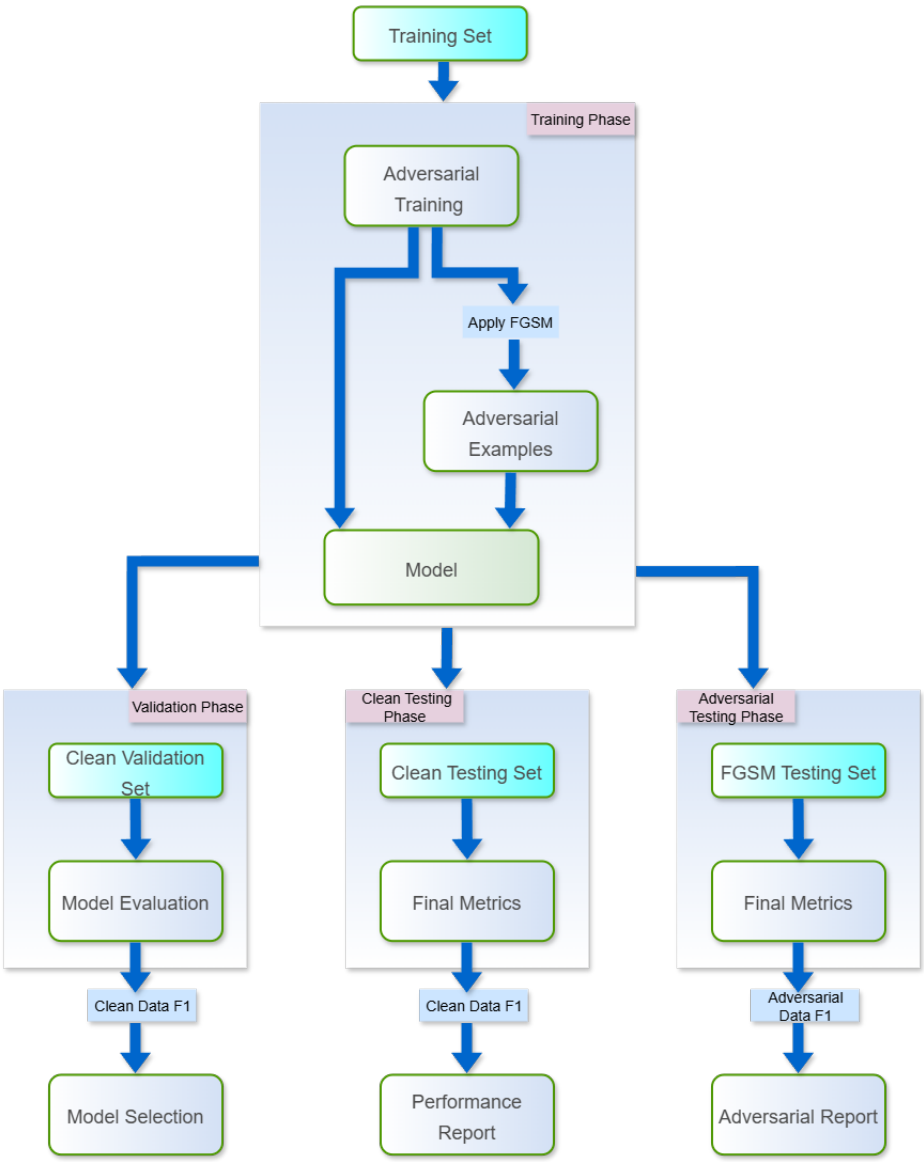

Figure 2. Adversarial training pipeline.

(1) **Training Phase**
Adversarial Sample Generation: The FGSM attack is applied to the original training set, with the perturbation intensity parameter (ϵ) adjusted to generate adversarial samples at varying attack strengths.
Construction of Mixed Training Set: Adversarial samples are combined with clean data at predefined ratios (e.g., 80%, 90%) to simulate environments with different levels of attack coverage, resulting in an augmented training set.

(2) **Validation Phase**
The clean validation set is used to evaluate model performance (e.g., F1 score), and the candidate model with the best performance on benign data is selected. This ensures that the basic detection capability is maintained and not compromised by adversarial training.

(3) **Testing Phase**
Clean Testing: The model's performance is evaluated on an unperturbed test set to assess its effectiveness and reliability in real-world scenarios.

Adversarial Testing: The model's robustness is quantified using an adversarial test set generated by FGSM, with analyses conducted across varying perturbation strengths (ϵ) and attack coverage levels to evaluate stability under adversarial conditions.

This pipeline is trained under a dual-dimensional "strength-breadth" control strategy, where ϵ regulates the perturbation intensity ("strength" dimension) and the mixture ratio determines the attack coverage ("breadth" dimension). Through a dual-testing environment, we systematically evaluate the generalization and robustness of the multi-view fusion model in complex adversarial settings.

3.4. Adversarial Attacks and Sample Generation

This section clarifies the core concepts relevant to this study. Adversarial examples are maliciously crafted inputs designed to mislead deep neural network classifiers by introducing minute perturbations to the original data—perturbations that are typically imperceptible to humans. This phenomenon was first identified by Szegedy et al. [7], who demonstrated that even negligible perturbations can cause well-trained neural networks to make severe misclassifications. Adversarial attacks can be categorized based on the attacker's knowledge of the target model [22]:

White-box attacks: The attacker has full access to the model's architecture and parameters.

Black-box attacks: The attacker has limited or no knowledge of the model internals.

From a different perspective, attacks can also be classified by other criteria. They can be classified by the number of attack iterations, and include one-step attacks (e.g., FGSM) vs. iterative attacks (e.g., PGD). They can also be classified by the attack objective, and include targeted attacks vs. untargeted attacks.

Among the various adversarial example generation methods, FGSM stands out due to its simplicity and computational efficiency. FGSM leverages the gradient information of neural networks to create adversarial samples efficiently. Here, adversarial examples are generated using the FGSM algorithm, which applies calculated perturbations to input data to effectively compromise intrusion detection systems (IDSs). Figure 3 illustrates the full adversarial attack process, detailing the transformation from an initial malicious sample to a successfully evaded detection.

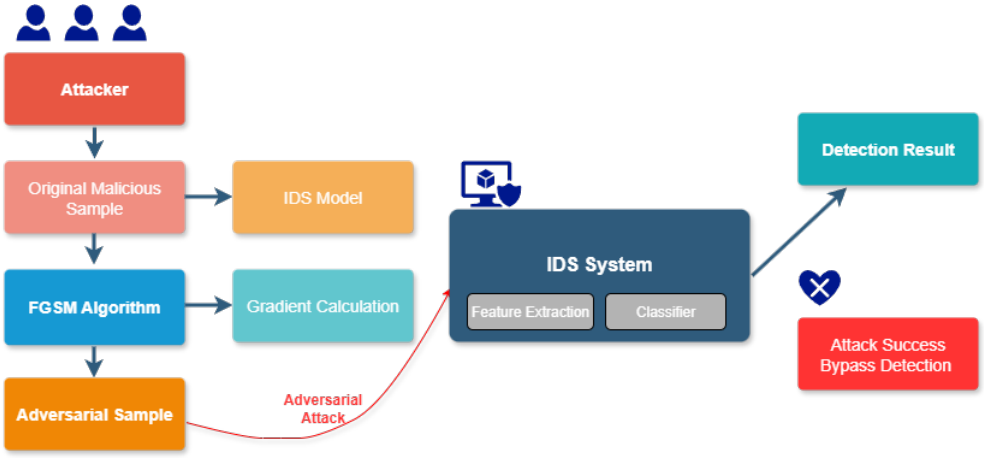

Figure 3. FGSM-based adversarial attack process

This algorithm is founded on the linear perturbation hypothesis introduced by Goodfellow et al. [8]. It generates adversarial samples by computing the model's gradient and applying sign-based perturbations. FGSM exemplifies the characteristics of a one-step, white-box attack: the adversary must have access to the model's architecture in order to obtain gradient information, and only a single gradient calculation is required to induce

misclassification or bypass detection in the IDS. Analogous to the vulnerability in image classification models observed by Szegedy et al. [7], IDS models are also susceptible to minor perturbations owing to the linear accumulation effects present in high-dimensional feature spaces.

In this study, FGSM is deliberately chosen as the representative adversarial method to isolate the effects of our *dual-axis evaluation framework* (varying perturbation strength ε and attack coverage) and to clearly examine the contribution of multi-view fusion to robustness. Our framework is attack-agnostic and can be readily extended to optimization-based attacks (e.g., PGD, C&W, AutoAttack). A broader exploration of these attacks is left as an important direction for future work.

4. Preparation for the Experiment

4.1. Data Set

This study is experimentally validated using two representative cybersecurity datasets.

The first dataset is the TON_IoT Win10 subset, developed by Moustafa at the University of New South Wales Canberra in 2021 as a contemporary testbed for intrusion detection in IoT environments [23]. This subset comprises 35,975 records, with 30.7% labeled as attack samples and 69.3% as normal, encompassing 125 attributes distributed across five distinct views: memory, process, processor, hard disk activity, and network traffic (Figure 4). The dataset originates from a test environment deployed on multiple virtual machines—including Windows, Linux, and Kali Linux—integrating IoT telemetry, network traffic, and system logs, thereby providing a comprehensive feature space for evaluating the robustness of intrusion detection systems (IDSs).

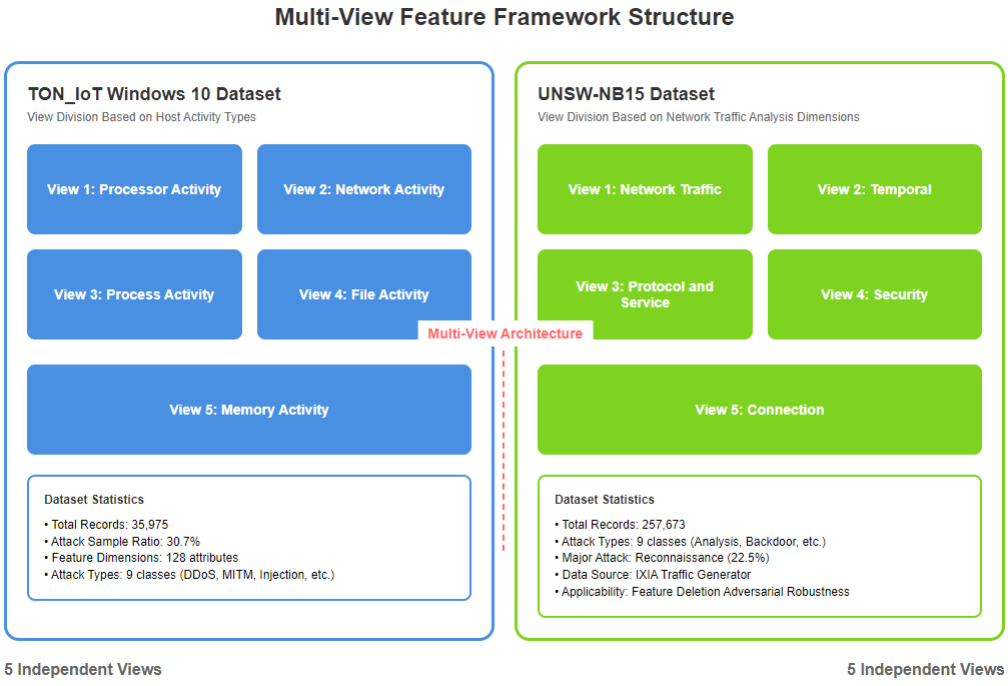

Figure 4. Multi-view feature framework structure.

From an adversarial perspective, the heterogeneity of the TON_IoT Win10 dataset enables the simulation of complex real-world attack scenarios, where adversaries may craft adversarial samples through subtle perturbations to bypass detection. Its multi-view structure (illustrated in Figure 4) facilitates multi-dimensional feature extraction, mitigating the risk posed by perturbations confined to a single view and ultimately enhancing IDS robustness against adversarial threats.

The second dataset used in this study is UNSW-NB15 [24,25] , released in 2015 and comprising 257,673 records representing realistic modern network activities, both normal and malicious. This dataset includes nine attack categories: Fuzzers, Analysis, Backdoors, Denial of Service (DoS), Exploits, Generic, Reconnaissance, Shellcode, and Worms. Among these, Reconnaissance attacks constitute the largest portion, accounting for 22.5% of the dataset. Raw network traffic was captured using the tcpdump tool, while feature extraction was performed with Argus and Bro-IDS tools. The dataset encompasses 49 features grouped into several categories, including basic features, content-based features, temporal features, and additional derived features.

From an adversarial standpoint, several network traffic features within UNSW-NB15—such as protocol type and packet size—are vulnerable to adversarial perturbations, making this dataset highly relevant for evaluating the robustness of intrusion detection systems under adversarial attack scenarios.

4.2. Multi-View Framework Design

This study employs a multi-view analysis framework to facilitate multi-view feature fusion. To this end, dataset-specific view partitioning schemes were devised to align with the intrinsic characteristics of each dataset. The framework architecture, as depicted in Figure 4, demonstrates the independent view partitioning strategies applied to both datasets.

For the TON_IoT dataset, host activities are categorized by type, resulting in five dedicated views: processor activity, network activity, process activity, file activity, and memory activity. This partitioning effectively captures multi-dimensional threat signatures characteristic of Windows environments.

In contrast, the UNSW-NB15 dataset adopts a network traffic-centric partitioning strategy, generating five distinct views based on different network analysis dimensions: traffic statistics, temporal features, protocol service, security assessment, and connection status. This multi-faceted partitioning significantly enhances the granularity and depth of network behavior analysis.

Together, these view architectures illustrate the adaptability and effectiveness of multi-view feature fusion across diverse cybersecurity contexts.

4.3. Data Preprocessing

To ensure consistency and comparability across all experiments, this section outlines the unified and adaptive data preprocessing strategy applied to all models in this study.

For the TON_IoT Win10 dataset, missing values were addressed using the missingData() function. This function first removes the type column and converts all object-type columns to numerical values. Any missing values resulting from this conversion are then imputed using a mean substitution approach to maintain data integrity.

Regarding the UNSW-NB15 dataset, a unified data transformation pipeline was implemented via the preprocess-data () function. Numerical features are normalized using the MinMaxScaler, scaling values to the range [0, 1].

A common preprocessing step for both datasets involves data partitioning. Stratified sampling is employed to split each dataset into training, validation, and test sets following an 8:1:1 ratio. For feature normalization, MinMaxScaler is applied consistently across both datasets. Notably, since the MinMaxScaler is integrated within the preprocess_data() function for `UNSW-NB15`, the ColumnTransformer framework is utilized to concurrently process categorical and numerical feature transformations.

Data adaptation techniques were implemented to accommodate the specific requirements of different model architectures. For single-view CNN models, one-dimensional feature vectors are reshaped into multi-dimensional tensors to match the expected input

format. Single-view autoencoders apply a feature selection strategy, restricting the input to the first selected features (FEATURES-SELECTED) to manage model complexity.

In the case of multi-view models, the datasets are partitioned into five semantically meaningful views based on domain knowledge. Specifically, the Windows 10 dataset is divided according to system activity types, while the `UNSW-NB15` dataset is segmented based on network traffic characteristics. Multi-view CNNs further require each view to be reshaped into two-dimensional tensors using the auto_reshape() function. For DGCCA models, labeled data are converted into PyTorch(v1.12.1) tensor format and processed according to the corresponding multi-view partitioning strategy.

5. Experiment and Results

5.1. Experiment Outline

In this section, we conduct a detailed evaluation of the robustness of single-view and multi-view models within IDS under adversarial conditions, aiming to verify the effectiveness and generalizability of the multi-view fusion strategy. The experimental framework involves a multi-dimensional, multi-level comparative analysis based on two representative datasets (TON_IoT Win10 and UNSWNB15), two adversarial dimensions (adversarial sample ratio and perturbation intensity), and two test environments (clean test set and adversarial test set). The following sections (Sections 5.2 and 5.3) will detail the experimental deployment and hyperparameter settings on the two datasets, respectively. These parameters are chosen within the typical ranges in previous IDS research and determined based on preliminary validation runs to balance convergence stability and detection performance. The experimental procedure is outlined as follows:

5.1.1. Evaluation Under Different Adversarial Sample Ratios (80%, 90%)

Using enhanced training sets with varying adversarial sample ratios, we compare the performance of single-view baseline models and multi-view models. During training, adversarial samples generated with a perturbation strength of $epsilon = 0.05$ are mixed into the original dataset at the specified ratios to construct the augmented training sets. In the testing phase, we evaluate the performance retention of all models across both clean and adversarial test environments, as well as analyze the robustness advantages exhibited by the multi-view models under different levels of attack coverage. Specifically, we set the attack coverage at 80% and 90%. An 80% level reflects a partial yet substantial compromise of the dataset, while 90% represents a near-complete adversarial takeover. These two settings enable us to evaluate robustness under both targeted attacks and large-scale adversarial conditions.

5.1.2. Valuation Based on Perturbation Intensity (ϵ from 0.01 to 0.2)

Focusing on the perturbation intensity dimension, we compare the performance of single-view baseline models and multi-view models. During training, augmented datasets with 80% and 90% FGSM adversarial sample ratios are utilized, incorporating adversarial samples generated at various ϵ levels. During the testing phase, the performance degradation $DeltaF1$ of all models is measured under both clean and adversarial test environments. $DeltaF1$ is defined as the range of F1 score variation across perturbation intensities with ϵ values from 0.01 to 0.2. This analysis assesses the robustness and stability of multi-view models under varying perturbation intensities.

5.1.3. Dual Verification of Model Robustness and Stability

Employing a dual verification mechanism that includes both clean and adversarial test sets, the F1 scores of all models are evaluated to determine whether multi-view models

exhibit superior robustness compared to single-view models in adversarial settings, as well as to examine their performance stability across different testing conditions.

As described in the experimental setup, the experiments conducted on the TON_IoT Win10 dataset were divided into two primary phases: the first phase evaluated model performance on a clean test set, while the second phase assessed the model's generalization ability under adversarial examples. Each phase incorporated two adversarial sample ratios (80% and 90%) and two perturbation strength schemes—a fixed perturbation strength ($\epsilon = 0.05$) and a dynamic perturbation strength varying between 0.01 and 0.2. This design systematically explores the combined effects of adversarial sample ratio and perturbation strength on model robustness, providing insights into the model's defense effectiveness under varied conditions and informing parameter optimization for FGSM adversarial training.

Subsequently, an in-depth comparative analysis was performed by calculating the average F1 scores across the two datasets to evaluate model generalization and robustness under different training strategies.

To verify the universality and generalizability of the findings, the same experimental protocols were applied to a second dataset. Given that each dataset consists of two phases and each phase involves four experimental settings, a total of eight experiments were conducted per dataset, amounting to sixteen experiments in total for this study.

To ensure the validity and reproducibility of the experimental results, the experimental environment and hardware configuration were standardized throughout this study. All experiments were conducted on a workstation equipped with an AMD Ryzen 7 5800U CPU (8 cores, 16 threads, 1.9–4.4 GHz) and an integrated AMD Radeon Graphics GPU. Additional computational resources were provided by the Center for High Performance Computing (SRI) at the Technological University of the Shannon: Midlands Midwest.

The software environment consisted of Windows 10 Pro (64-bit) as the operating system. Model development and training were implemented using Python 3.8.10 (Anaconda distribution) in conjunction with the PyTorch 1.12.1 deep learning framework. Auxiliary libraries, including scikit-learn 1.0.2, pandas 1.4.3, and NetworkX 2.6.3, were employed to support data analysis and model evaluation.

5.2. Experimental Deployment Based on the TON_IoT Dataset

This section describes the detailed deployment of experiments based on five models (single-view AE, single-view CNN, multi-view AE, multi-view CNN, and DGCCA).We conduct a series of comparative experiments based on a dual-validation mechanism—evaluating model performance on both clean and adversarial test sets—to determine whether multi-view models consistently outperform single-view models in terms of overall performance and robustness.

In the first stage, augmented training sets are constructed using FGSM adversarial samples with 80% and 90% ratios at a fixed perturbation strength of $\epsilon = 0.05$. The models are then evaluated on a clean test set (100% original samples) and an adversarial test set (50% adversarial samples). In the second stage, we further investigate model robustness under varying perturbation strengths (ϵ ranging from 0.01 to 0.2). By analyzing performance across this range, we assess the stability and adaptability of multi-view models relative to single-view baselines, thereby verifying the effectiveness of multi-view fusion in enhancing adversarial robustness.

In this experimental setup, the core configuration of the AE model includes MinMaxScaler normalization. The model is trained using the Adam optimizer with a learning rate of 0.001, the mean squared error (MSE) loss function, and 100 training epochs (batch size 32, initially set to 64). The encoder has 128, 64, 32, and 8 units, with batch normalization,

ELU activation, and a dropout rate of 0.3. The decoder consists of 32, 64, 128, and 125 units, ending with a sigmoid activation. In the single-view setting, the model takes the full feature set as input, while in the multi-view setting, separate subsets of features are processed independently. Anomalies are identified by minimizing the reconstruction error.

For CNN models, it is worth mentioning the multi-view feature structure. Each view's input is reshaped into a two-dimensional tensor before being passed into the convolutional layers. Each view is processed through a convolutional block consisting of a convolutional layer with 32 or 128 filters (kernel size = 2, ReLU activation), followed by adaptive max pooling and a fully connected layer. The outputs from all views are then concatenated and passed through a fully connected layer with 64 units and a dropout rate of 0.5 to produce a binary classification output. The model is trained using the Adam optimizer with a learning rate of 0.001, consistent with the AE model.

The DGCCA model processes each view separately using a multi-layer perceptron (MLP) consisting of four layers: input $\rightarrow$ 32 $\rightarrow$ 16 $\rightarrow$ 8 $\rightarrow$ output (2 units). Each MLP uses LogSigmoid activation and BatchNorm1d for normalization. DGCCA is used to integrate the representations across views, and inter-view correlations are maximized using the SVD-based GCCA loss. Parameter optimization is performed using the Adam optimizer.

Tables 1 and 2 summarize the performance of all models on the clean test set and the adversarial test set (with $\epsilon = 0.05$), respectively. Both tables include results under two training configurations, in which 80% and 90% of the training data consist of FGSM-generated adversarial examples. These results allow for a systematic comparison of the models' generalization to natural samples and robustness to adversarial samples under different training conditions. Figure 5 further illustrates how each model's performance changes across a range of perturbation strengths (ϵ from 0.01 to 0.2), providing additional insight into model stability under varying attack intensities. All results are reported as mean $\pm$ standard deviation over 30 runs with different random seeds.

Table 1. F1 scores of models based on TON_IoT dataset (clean testing set).

Models	**F1 Score**		
	$\varepsilon = 0$ (0% adv)	**$\varepsilon = 0.05$ (80% adv)**	**$\varepsilon = 0.05$ (90% adv)**
Single-view AE	0.72	0.4778 ± 0.0003	0.4718 ± 0.0162
Single-view CNN	0.892	0.8539 ± 0.0123	0.8386 ± 0.0130
Multi-view AE	0.871	0.8521 ± 0.0083	0.8231 ± 0.0091
Multi-view CNN	0.925	0.8631 ± 0.0104	0.8543 ± 0.0112
DGCCA	0.86	0.8418 ± 0.0004	0.8020 ± 0.0223

Table 2. F1 scores of models based on TON_IoT dataset (adversarial example testing set).

Models	**F1 Score**		
	$\varepsilon = 0$ (0% adv)	**$\varepsilon = 0.05$ (80% adv)**	**$\varepsilon = 0.05$ (90% adv)**
Single-view AE	0.72	0.4722 ± 0.0140	0.4719 ± 0.00172
Single-view CNN	0.892	0.6184 ± 0.0130	0.6010 ± 0.0128
Multi-view AE	0.871	0.7817 ± 0.0075	0.7362 ± 0.0079
Multi-view CNN	0.925	0.7823 ± 0.0024	0.7396 ± 0.0054
DGCCA	0.86	0.6433 ± 0.0120	0.5480 ± 0.0146

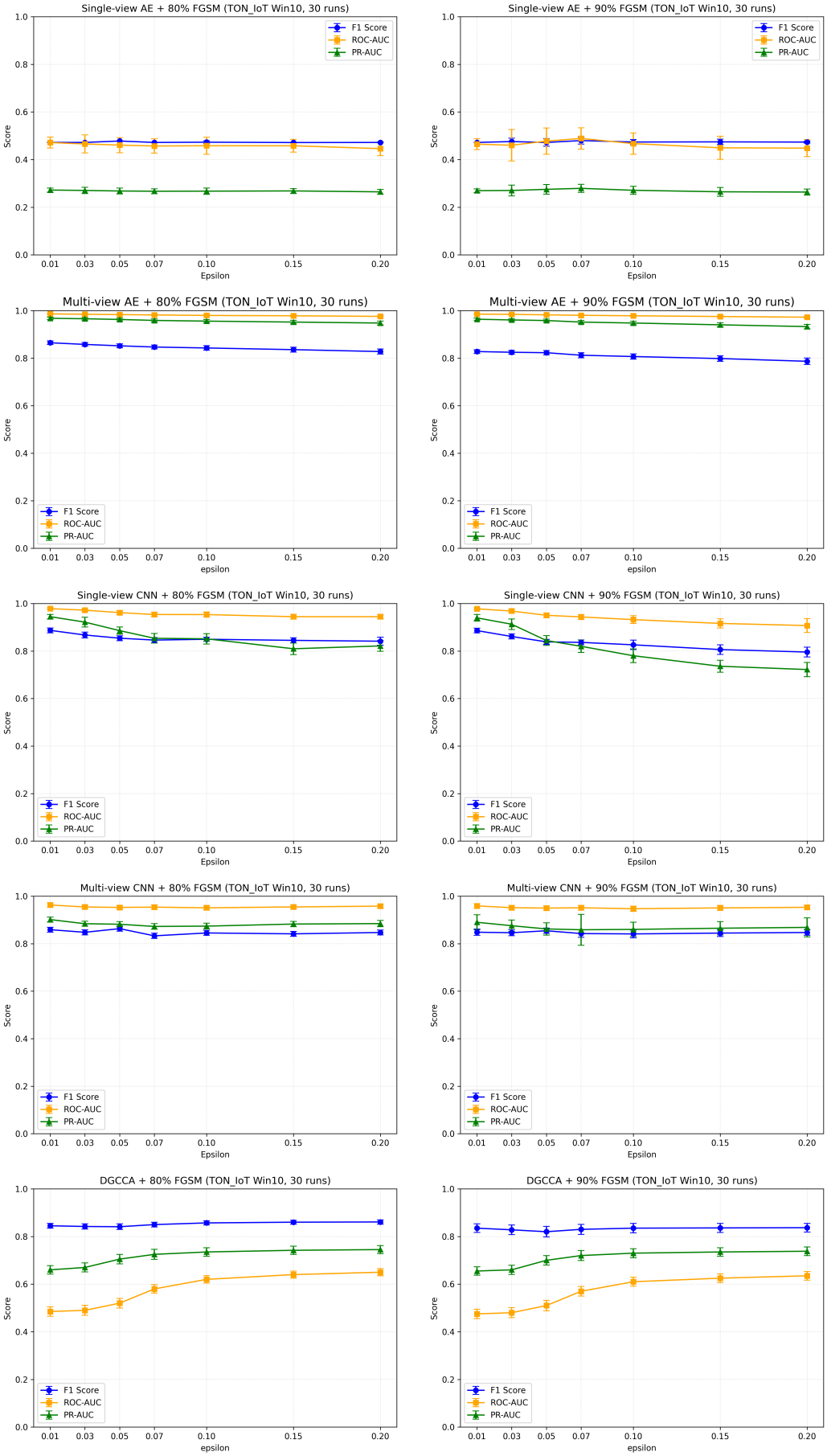

Figure 5. Curve plots for different adversarial samples and ε values (TON_IoT Win10 dataset).

5.3. Experimental Deployment Based on the UNSW-NB15 Dataset

This section outlines the experimental deployment on the UNSW-NB15 dataset and describes the data preprocessing procedures and parameter configurations adapted to the specific characteristics of each model type. In the single-view setting, all features are processed collectively, whereas in the multi-view setting, features are divided into multiple subsets corresponding to distinct views.

For the AE model, a symmetrical multi-layer fully connected architecture is adopted. Both the encoder and decoder consist of four layers with dimensions of 128, 64, 32, and 8/8, 32, 64, and 128, respectively. Each layer is followed by batch normalization, ELU activation, and a dropout layer with a rate of 0.3. Training was conducted for 100 epochs using a batch size of 32. The Adam optimizer was applied with a learning rate of 0.001, and the mean squared error (MSE) was used as the loss function. In the multi-view setting, an independent autoencoder was trained for each view to ensure the model's ability to reconstruct and detect anomalies within each feature subspace.

Similarly, for the CNN model, the input features of each view are reshaped to accommodate one-dimensional convolution operations. Each view branch comprises a Conv1D layer (32 filters, kernel size = 2, ReLU activation), followed by batch normalization, MaxPooling1D, and a Flatten layer. The outputs from all views are then concatenated and passed through a fully connected layer (dense with 64 units and ReLU activation), followed by a dropout layer with a rate of 0.5. The final layer uses a sigmoid activation function to perform binary classification. The training settings for the CNN model are consistent with those of the autoencoder: 100 training epochs, a batch size of 32, and the Adam optimizer with a learning rate of 0.001. The number of convolutional filters is fixed at 32 across all branches, and standard MaxPooling1D is uniformly applied.

For the DGCCA model, a separate multi-layer perceptron (MLP) is constructed for each view, with the following architecture: input $\rightarrow$ 32 $\rightarrow$ 16 $\rightarrow$ 8 $\rightarrow$ 2. All hidden layers employ the LogSigmoid activation function, and BatchNorm1D is applied after each layer to improve training stability. The DGCCA module fuses the multi-view feature representations by maximizing the inter-view correlation using a customized DGCCA loss function based on singular value decomposition (SVD). The model is trained using the Adam optimizer with a learning rate of 0.001 for 100 epochs and a fixed batch size of 32, maintaining consistency with the other models for fair comparison.

Tables 3 and 4 present the performance of each model on clean test samples and adversarial samples (with $\epsilon = 0.05$), respectively, under two adversarial training settings: 80% and 90% adversarial sample ratios. Figure 6 illustrates the performance trends of all models as the perturbation strength (ϵ) increases from 0.01 to 0.2, highlighting the sensitivity and robustness of different architectures under varying attack intensities. All results are reported as mean $\pm$ standard deviation over 30 runs with different random seeds.

Table 3. F1 scores of models based on UNSW-NB15 dataset (clean testing set).

Models	**F1 Score**		
	$\varepsilon = 0$ **(0% adv)**	$\varepsilon = 0.05$ **(80% adv)**	$\varepsilon = 0.05$ **(90% adv)**
Single-view AE	0.797	0.6121 ± 0.022	0.3629 ± 0.015
Single-view CNN	0.806	0.7641 ± 0.0156	0.7588 ± 0.0209
Multi-view AE	0.848	0.8243 ± 0.014	0.8162 ± 0.014
Multi-view CNN	0.856	0.8476 ± 0.012	0.8237 ± 0.014
DGCCA	0.8	0.7589 ± 0.015	0.7530 ± 0.0218

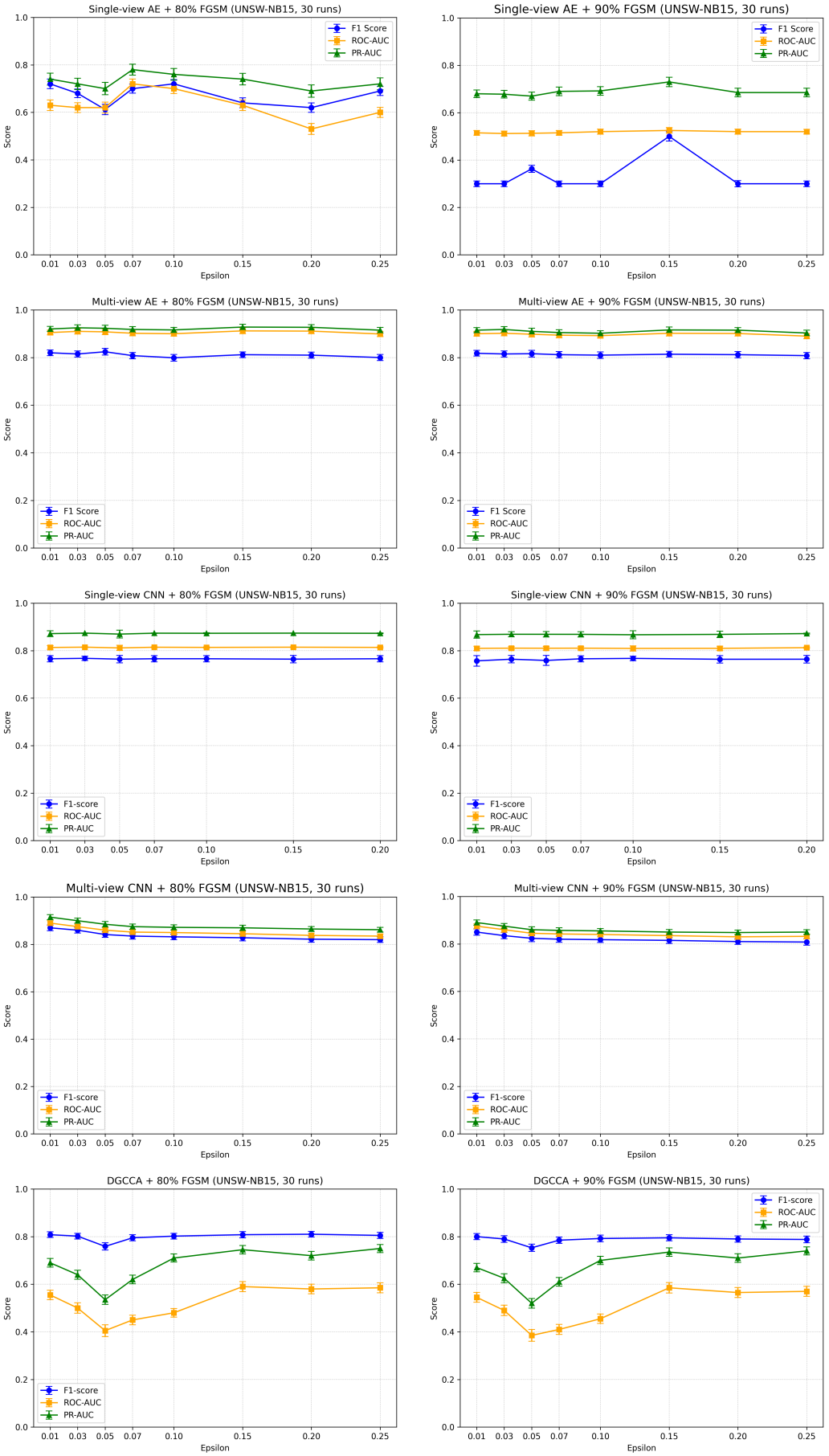

Figure 6. Curve plots for different adversarial samples and ε values (UNSW_NB15 dataset).

Table 4. F1 scores of models based on UNSW-NB15 dataset (adversarial example testing set).

Models	F1 Score		
	$\varepsilon = 0$ (0% adv)	$\varepsilon = 0.05$ (80% adv)	$\varepsilon = 0.05$ (90% adv)
Single-view AE	0.797	0.6232 ± 0.017	0.3514 ± 0.013
Single-view CNN	0.806	0.6481 ± 0.0096	0.6221 ± 0.0085
Multi-view AE	0.848	0.75128 ± 0.0054	0.7324 ± 0.0071
Multi-view CNN	0.856	0.7568 ± 0.0051	0.7344 ± 0.0060
DGCCA	0.8	0.6161 ± 0.13	0.4879 ± 0.14

6. Conclusions

To evaluate the effectiveness of multi-view fusion strategies in enhancing robustness against adversarial intrusion detection, this study conducted a series of systematic experiments covering the following four aspects:

Diverse model architectures (five types of models);

Training with varying adversarial sample ratios (80% and 90% FGSM);

Multiple perturbation strengths (ϵ ranging from 0.01 to 0.2);

Comprehensive validation (dual testing on clean and adversarial test sets).

This section will provide a detailed analysis based on all the deployed experimental results.

6.1. Experimental Conclusion Analysis

6.1.1. Analysis Based on the TON_IoT Dataset

We first investigated the impact of varying adversarial sample ratios (80% and 90%) in the training set on model robustness, under a fixed perturbation strength of $\epsilon = 0.05$. The experimental results presented in Tables 1 and 2, corresponding to the clean and adversarial test sets, respectively, provide the following validation conclusions:

(1) After introducing perturbations and varying proportions of adversarial examples, all models exhibited different levels of performance degradation on both the clean and adversarial test sets, highlighting the well-known robustness–accuracy trade-off in adversarial training. The degradation generally ranged between 3% and 7%, with some models—particularly the single-view AE—experiencing much larger drops, which means that they nearly experienced complete performance failure. A sharper decline indicates lower stability under adversarial conditions. In contrast, multi-view models demonstrated more moderate performance degradation, suggesting better adversarial robustness. Among them, the multi-view CNN achieved the most balanced results across both the clean and adversarial test sets, representing the best trade-off between performance and robustness among all evaluated models. These findings indicate that the multi-view fusion strategy, especially when combined with a well-structured CNN architecture, not only enhances robustness against adversarial attacks but also maintains better stability on clean distributions.

(2) Under varying perturbation intensities, the F1 score of the multi-view CNN decreases by less than 0.04 within the range $\epsilon \in [0.01, 0.2]$, maintaining a relatively stable trend. The multi-view AE similarly exhibits limited performance fluctuation, with a ΔF1 not exceeding 0.03. While the single-view CNN shows a downward trend, it still performs better than DGCCA. In contrast, the PR-AUC and ROC-AUC curves of the single-view AE and DGCCA fluctuate significantly under medium to high perturbation levels. Overall, the multi-view models exhibit more stable performance across different perturbation intensities, with smoother metric curves and slower degradation, indicating stronger resistance to adversarial perturbations.

6.1.2. Analysis Based on the UNSW-NB15 Dataset

To achieve the experimental objectives and validate the model's generalization ability, the same experiments were conducted on the UNSW-NB15 dataset. Based on the results shown in Tables 3 and 4, the following conclusions can be drawn:

(1) A consistent trend with the previous dataset is observed: after adversarial training, the F1 scores of all models exhibit varying degrees of decline. However, in both the clean and adversarial test sets, multi-view models consistently demonstrate stronger resistance to adversarial interference, confirming their robustness superiority.
(2) Under different perturbation intensities, the F1 score of the multi-view CNN fluctuates by only about 0.02 across the ϵ range, maintaining the most stable performance. The multi-view AE also shows excellent stability. In contrast, the single-view CNN experiences a significant drop in performance, especially when $\epsilon > 0.10$. The single-view AE shows a highly unstable curve and is easily affected by perturbations. Although the DGCCA model remains relatively stable, its overall performance is limited, with considerable fluctuations observed in its ROC-AUC curve. These results further demonstrate that multi-view fusion models maintain superior stability and robustness under adversarial conditions.

6.2. Comprehensive Conclusion

Through cross-validation on two datasets and a dual-dimensional experimental design involving perturbation intensity and adversarial sample ratio, this study yields the following key conclusions:

(1) The multi-view fusion strategy significantly outperforms the single-view baseline models. In particular, the multi-view CNN achieves the highest F1 scores across both datasets, multiple attack ratios, and various perturbation intensities.
(2) The multi-view models exhibit superior resistance to adversarial perturbations. They demonstrate the smallest degradation in F1, ROC-AUC, and PR-AUC scores as ϵ increases, indicating good robustness under escalating attack strengths.
(3) The multi-view models retain strong generalization ability. Compared to single-view models, they maintain comparable performance on both clean and adversarial test sets, suggesting that adversarial training does not compromise their original detection capabilities.
(4) DGCCA is a special case among multi-view models. Unlike other multi-view architectures that focus on feature extraction within each individual view followed by simple concatenation, DGCCA explicitly models inter-view correlations by projecting all views into a shared latent subspace. While this design captures structural dependencies between views, it may also amplify inter-view interference. As a result, DGCCA tends to exhibit weaker overall performance on both sets compared to some single-view baselines.

6.3. Theoretical Contributions and Practical Value

The evaluation conducted in this study uncovers the underlying robustness mechanisms of multi-view models under adversarial environments. It confirms the effectiveness of feature-space-diversified fusion strategies, thereby offering new theoretical insights into the field of secure machine learning. From a practical standpoint, the robustness of multi-view fusion can be explained by three main factors:

Representation Redundancy—Redundant features provide backup information. Even if adversarial noise affects one view, other views can still supply reliable signals, reducing overall vulnerability.

Feature Complementarity—Different views contribute complementary information from diverse feature spaces, making it harder for perturbations crafted in one subspace to transfer effectively to others.

Gradient Misalignment—Gradients from different views are not perfectly aligned, which means a single perturbation direction is less likely to disrupt all views at once.

From a broader perspective, the proposed evaluation framework and adversarial training methodology are transferable to other security-critical domains—such as financial fraud detection and industrial control system monitoring—demonstrating broad applicability for assessing and enhancing the robustness of existing detection systems.

6.4. Future Work

Based on the current research, several directions can be pursued in future work. First, the existing multi-view fusion strategies primarily rely on spatially parallel feature extraction. Future studies could investigate modeling inter-view interactions and structural dependencies, enabling the fusion process to capture complementarity and correlations between views more effectively. This enhancement is expected to not only improve adversarial robustness but also enhance the model's discriminative power. Second, while this study employs FGSM for adversarial sample generation due to its simplicity and efficiency, future work could incorporate stronger and more diverse attack methods, such as PGD (Projected Gradient Descent), AutoAttack, or Carlini & Wagner (C&W) attacks, to further evaluate model robustness under more challenging conditions. Third, comparing multi-view fusion methods with newer IDS models, such as ensemble approaches or transformer-based tabular networks, can give a more complete picture of system performance and make the evaluation of robustness more reliable. Finally, future research could focus on evaluating the cross-domain and cross-dataset generalization robustness of multi-view fusion strategies, further validating their applicability in complex, real-world intrusion detection scenarios.

Author Contributions: M.L.: conceptualization of this study, methodology, code, validation, writing—original draft, visualization. Y.Q.: conceptualization of this study, supervision. B.L.: funding acquisition, conceptualization of this study, methodology, writing—review and editing, supervision. All authors have read and agreed to the published version of the manuscript

Funding: This publication has resulted from research conducted with the financial support of the Technological University of the Shannon under the President Doctoral Scholarship 2021, and the Horizon Europe Framework Program (HORIZON), under the grant agreement 101119681, Resilmesh.

Data Availability Statement: The dataset used in this study is free and publicly available on the Internet. We can obtain the TON_IoT dataset and UNSW-NB15 dataset through the following link: https://research.unsw.edu.au/projects/toniot-datasets (accessed on 3 April 2022)

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aldhaheri, A.; Alwahedi, F.; Ferrag, M. A.; Battah, A. Deep learning for cyber threat detection in IoT networks: A review. *Internet Things Cyber-Phys. Syst.* **2024**, *4*, 110–128. [CrossRef]
2. Lin, Z.; Shi, Y.; Xue, Z. IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection Systems. In Proceedings of the Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, Chengdu, China, 16–19 May 2022.
3. Zhang, J.; Sun, H.; Wang, J. Malicious domain name detection model based on CNN-LSTM. In Proceedings of the Third International Conference on Computer Communication and Network Security (CCNS 2022), Hohhot, China, 15–17 July 2022.
4. Bai, J.; Ju, L. Intrusion Detection Algorithm Based on Adversarial Autocoder. *J. Southwest China Norm. Univ. Nat. Sci. Ed.* **2021**, *46*, 77–83. [CrossRef]

5. Pelekis, S.; Koutroubas, T.; Blika, A.; Berdelis, A.; Karakolis, E.; Ntanos, C.; Spiliotis, E.; Askounis, D. Adversarial machine learning: A review of methods, tools, and critical industry sectors. *Artif. Intell. Rev.* **2025**, *58*, 226. [CrossRef]
6. Li, Y.; Xu, Y.; Liu, Z.; Hou, H.; Zheng, Y.; Xin, Y.; Zhao, Y.; Cui, L. Robust detection for network intrusion of industrial IoT based on multi-CNN fusion. *Measurement* **2020**, *154*, 107450. [CrossRef]
7. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
8. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**, arXiv:1412.6572.
9. Alotaibi, A.; Rassam, M.A. Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. *Future Internet* **2023**, *15*, 62. [CrossRef]
10. Issa, M.; Aljanabi, M.; Muhialdeen, H. Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. *J. Intell. Syst.* **2024**, *33*, 20230248. [CrossRef]
11. Sajid, M.; Malik, K.R.; Almogren, A.; Malik, T.S.; Khan, A.H.; Tanveer, J.; Rehman, A.U. Enhancing intrusion detection: A hybrid machine and deep learning approach. *J. Cloud Comput.* **2024**, *13*, 123. [CrossRef]
12. Zhang, Y.; Muniyandi, R.C.; Qamar, F. A Review of Deep Learning Applications in Intrusion Detection Systems: Overcoming Challenges in Spatiotemporal Feature Extraction and Data Imbalance. *Appl. Sci.* **2025**, *15*, 1552. [CrossRef]
13. Barni, M.; Kallas, K.; Nowroozi, E.; Tondi, B. On the Transferability of Adversarial Examples against CNN-based Image Forensics. In Proceedings of the ICASSP 2019—IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 8286–8290. [CrossRef]
14. Han, S.; Lin, C.; Shen, C.; Wang, Q.; Guan, X. Interpreting Adversarial Examples in Deep Learning: A Review. *ACM Comput. Surv.* **2023**, *55*, 328. [CrossRef]
15. Sauka, K.; Shin, G.-Y.; Kim, D.-W.; Han, M.-M. Adversarial Robust and Explainable Network Intrusion Detection Systems Based on Deep Learning. *Appl. Sci.* **2022**, *12*, 6451. [CrossRef]
16. Liu, H.; Lang, B. Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Appl. Sci.* **2019**, *9*, 4396. [CrossRef]
17. Li, Y.; Li, L.; Wang, L.; Zhang, T.; Gong, B. NATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019.
18. Awad, Z.; Zakaria, M.; Hassan, R. An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems. *Sci. Rep.* **2025**, *15*, 14177. [CrossRef]
19. Chowdhury, K. Adversarial Machine Learning: Attacking and Safeguarding Image Datasets. In Proceedings of the 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 12–13 December 2024; pp. 1360–1365. [CrossRef]
20. Halbouni, A.; Gunawan, T.S.; Habaebi, M.H.; Halbouni, M.; Kartiwi, M.; Ahmad, R. CNN-LSTM: Hybrid Deep Neural Network for Network Intrusion Detection System. *IEEE Access* **2022**, *10*, 99837–99849. [CrossRef]
21. Li, M.; Qiao, Y.; Lee, B. A Comparative Analysis of Single and Multi-View Deep Learning for Cybersecurity Anomaly Detection. *IEEE Access* **2025**, *13*, 83996–84012. [CrossRef]
22. Lella, E.; Macchiarulo, N.; Pazienza, A.; Lofù, D.; Abbatecola, A.; Noviello, P. Improving the Robustness of DNNs-based Network Intrusion Detection Systems through Adversarial Training. *arXiv* **2023**, arXiv:10193009.
23. Moustafa, N. TON_IoT Datasets. UNSW Canberra. 2022. Available online: https://research.unsw.edu.au/projects/toniot-datasets (accessed on 3 April 2022).
24. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems. In Proceedings of the Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 10–12 November 2015.
25. Moustafa, N.; Slay, J. The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset. *Inf. Secur. J. A Glob. Perspect.* **2016**, *25*, 18–31. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.