International Journal of Information Security (2026) 25:125
https://doi.org/10.1007/s10207-026-01299-x

REGULAR CONTRIBUTION

Detecting adversarial evasion in deep learning intrusion detection systems using explainable AI

Elijah M. Maseno[1] · **Yanxia Sun**[1] · **Zenghui Wang**[2]

Received: 16 April 2026 / Accepted: 4 July 2026
© The Author(s) 2026

Abstract

Deep learning based network intrusion detection systems (IDS) can achieve strong traffic classification performance, but their resilience to adversarial manipulation remains a critical concern. This study evaluates the adversarial robustness of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models in a multiclass intrusion detection setting using the Train_Test_Network dataset with ten traffic classes. The models were trained on true sliding flow-window sequences under a unified preprocessing pipeline to support fair comparison. Adversarial robustness was first assessed under a white-box Fast Gradient Sign Method (FGSM) setting and then broadened through additional FGSM and Projected Gradient Descent (PGD) stress testing. SHapley Additive exPlanations (SHAP) were further used to analyse explanation instability under clean and adversarial conditions, and explanation-drift features were evaluated as a secondary adversarial detection signal. Under clean evaluation, both models achieved strong and nearly identical performance, with accuracies of 0.9614 for LSTM and 0.9615 for GRU and weighted F1-scores of 0.9597 and 0.9598, respectively. Under the main FGSM condition, performance declined substantially: the LSTM achieved adversarial accuracy of 0.6094 and weighted F1-score of 0.6290 with an evasion rate of 37.38%, while the GRU achieved adversarial accuracy of 0.5130 and weighted F1-score of 0.5690 with an evasion rate of 47.02%. The broader robustness sweep showed that iterative PGD exposed stronger fragility than FGSM alone. SHAP analysis indicated that adversarial perturbation altered both prediction outcomes and local explanation structure. A learned explanation-driven detector improved over the rule-based baseline, while larger-scale validation confirmed that explanation drift remained informative, though not perfectly separable, at broader scale. Overall, the results show that strong clean performance does not imply adversarial robustness, and that explanation drift provides a useful auxiliary signal for adversarial monitoring in recurrent IDS models.

Keywords Intrusion detection systems · Adversarial machine learning · Explainable artificial intelligence · SHAP · Recurrent neural networks

1 Introduction

Intrusion Detection Systems (IDS) remain a critical component of modern cybersecurity because they support the timely identification of malicious activity that may bypass preventive controls such as firewalls, access control policies, and endpoint protection. In contemporary enterprise and institutional environments, network traffic is high in volume, heterogeneous in structure, and increasingly shaped by evolving attack techniques. Under these conditions, IDS platforms are expected not only to detect known threats but also to distinguish subtle deviations from legitimate behaviour in near real time. This requirement has increased the importance of intelligent detection methods that can operate effectively

✉ Yanxia Sun
ysun@uj.ac.za

Elijah M. Maseno
226118296@student.uj.ac.za

Zenghui Wang
wangzengh@gmail.com

[1] Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg, South Africa

[2] Department of Electrical Engineering, University of South Africa, Johannesburg, South Africa

Published online: 15 July 2026

across complex and dynamic traffic patterns [1, 2]. The growing complexity of networked systems has contributed to the adoption of machine learning and deep learning methods in IDS research. Compared with signature-based mechanisms, learning-based models can capture statistical regularities and higher-order feature interactions from traffic data, making them attractive for detecting both known and previously unseen attacks. Among deep learning approaches, recurrent neural architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks have received considerable attention due to their ability to model temporal dependencies in sequential data and retain informative patterns through gating mechanisms [3, 4]. In IDS settings, these properties are especially useful when traffic features are represented in a structured sequential form, since attack behaviour may depend on nontrivial relationships among multiple attributes rather than isolated feature values alone. As a result, LSTM and GRU models are increasingly explored as effective classifiers for multiclass intrusion detection tasks.

Despite their predictive strength, deep learning-based IDS models introduce an important security risk: they are vulnerable to adversarial manipulation. Adversarial machine learning has shown that carefully crafted perturbations, even when small in magnitude, can alter model predictions and cause misclassification at inference time [5, 6]. In the context of intrusion detection, this threat is especially serious because an attacker may exploit model sensitivity to evade detection while preserving traffic characteristics that remain operationally plausible. Such evasion attacks challenge a key assumption behind learning-based IDS deployment, namely that strong performance on clean benchmark data translates into robust behavior under hostile conditions. For security applications, this assumption is dire. A model that performs well in standard evaluation but fails under adversarial perturbation may provide a misleading sense of defensive capability.

Alongside predictive accuracy, explainability has become increasingly important in analyst-facing security systems. In operational environments, IDS outputs often need to be interpreted by analysts who must validate alerts, prioritise incidents, and determine appropriate response actions. Post hoc explainability methods such as SHapley Additive exPlanations (SHAP) are therefore widely used to identify the features that most strongly influence individual predictions and to increase transparency in model-assisted decision support [7]. In principle, such explanations can improve trust, accountability, and analytical efficiency by helping analysts understand why a model has classified a traffic instance as benign or malicious. However, explainability in security contexts should not be treated only as a usability layer. If adversarial perturbations modify not only the final prediction but also the explanation pattern associated with that prediction, then explanation reliability itself becomes a security-relevant concern.

This issue reveals an important gap in the current literature. Much of the existing work on adversarial robustness in IDS focuses on predictive degradation, attack success rates, or defence strategies aimed at preserving classification accuracy under perturbation. Separately, the growing literature on explainable AI in cybersecurity has largely examined how explanation techniques can improve transparency, trust, and human interpretation of model outputs. Comparatively fewer studies have addressed the intersection of these two questions: whether adversarial attacks systematically distort explanations, and whether such distortion can itself be exploited as a signal for detecting manipulated inputs. In other words, while adversarial robustness and explainability have each been studied independently in IDS research, the reliability of explanations under adversarial pressure remains insufficiently examined. This is a consequential omission because explanation instability may indicate that the model's decision process has been materially altered, even when conventional output-level cues are ambiguous.

The present study addresses this gap by evaluating the adversarial robustness of LSTM and GRU based IDS models within a multiclass temporal flow-window setting and by examining whether SHAP explanation instability can support adversarial monitoring. Specifically, the study subjects recurrent IDS classifiers to white-box adversarial perturbation under a main FGSM setting and a broader FGSM and PGD robustness sweep, and compares their behaviour under clean and adversarial conditions. It then analyses how SHAP attribution patterns change between original and perturbed inputs and uses two complementary explanation-deviation signals, namely explanation drift and top-k feature overlap, as the basis for explanation-driven adversarial detection. To improve beyond rigid thresholding, the study further evaluates a lightweight learned detector over these explanation-deviation features. By linking adversarial robustness analysis with explanation reliability, the study reframes explainability from a purely interpretive tool into a practical auxiliary signal for identifying manipulated traffic instances. This perspective is particularly relevant for operational IDS settings, where both predictive performance and explanation stability must remain dependable under hostile conditions.

Accordingly, the contributions of this study are as follows:

1. A multiclass adversarial robustness evaluation of LSTM and GRU based IDS models using true sliding flow-window sequences under white-box FGSM and broader FGSM and PGD stress testing.

2. A SHAP based analysis of how adversarial perturbation alters explanation structure through global attribution drift and top-k feature reordering.
3. An explanation-driven adversarial detection framework that combines rule-based SHAP monitoring with a lightweight learned detector over explanation-deviation features.
4. A comparative assessment of LSTM and GRU in terms of predictive robustness, class-level fragility, and explanation-based monitoring effectiveness.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the methodology. Section 4 presents the experimental setup 5 provides the results. Section 6 presents the discussion. Section 7 concludes the study and identifies directions for future research.

2 Related work

Intrusion detection research has evolved from rule-based and classical anomaly-detection paradigms toward machine learning and, more recently, deep learning, largely because contemporary network environments generate heterogeneous, high-volume, and rapidly changing traffic patterns that are difficult to model with static signatures alone. Foundational survey work has consistently shown that learning-based IDS methods are attractive because they can capture nonlinear relationships and generalize beyond fixed attack templates, but it has also warned that benchmark accuracy alone is an insufficient basis for judging practical IDS effectiveness in operational settings [1, 2, 8]. More recent surveys similarly indicate that deep learning has become a dominant line of inquiry in IDS because of its capacity to process complex feature spaces and support multiclass attack detection, especially in IoT and large-scale network environments [9, 10].

Within this shift, recurrent architectures have received sustained attention because traffic observations can exhibit structured dependencies across features, packets, flows, or temporally ordered representations. LSTM networks are particularly attractive because their memory cells and gating mechanisms support dependency modeling over extended contexts, while GRU networks provide a related but comparatively streamlined architecture that often reduces parameter complexity without abandoning recurrent gating behavior [3, 4]. IDS studies using recurrent models have reported strong performance in both binary and multiclass settings, with representative examples including recurrent neural network-based IDS pipelines, LSTM based intrusion detection, and more recent hierarchical LSTM frameworks for network traffic analysis [11–14]. GRU based IDS studies have also expanded, especially in IoT contexts, where efficiency and sequence handling are both important design considerations [15].

Conceptually, prior work converges on the view that recurrent deep learning architectures are well suited to IDS because they can model relationships that are not easily captured by shallow learners. However, much of this literature is still optimized for clean-data evaluation and frequently emphasizes predictive metrics such as accuracy, precision, recall, F1-score, or ROC-AUC under benign conditions. Two limitations follow. First, high clean performance is often treated as evidence of practical suitability even though the deployment environment is adversarial. Second, LSTM and GRU are commonly compared as alternative sequence learners in terms of classification performance, but far less often under a shared adversarial threat model. As a result, current literature shows that both architectures are viable for IDS, yet it provides less clarity on whether they fail differently, or remain differentially resilient, once an attacker deliberately perturbs the input space [16] [9, 17].

As machine learning and deep learning became more central to IDS design, adversarial machine learning emerged as a corresponding concern. The adversarial IDS literature now makes clear that learning-based detectors can be systematically manipulated through carefully crafted perturbations that alter the model's inference behavior while preserving substantial similarity to the original input from the attacker's perspective. Survey evidence shows that this problem spans attack types, datasets, learning algorithms, and evaluation protocols, making adversarial robustness a core issue rather than a niche concern in IDS research [9]. Comparative studies similarly show that IDS models can be meaningfully degraded by adversarial perturbations and that robustness varies across model classes and attack assumptions [17].

A useful conceptual distinction in this literature is between feature-space attacks and more realistic problem-space attacks. Feature-space attacks such as FGSM, PGD, and related gradient-based methods remain popular because they are computationally convenient, analytically transparent, and effective for stress testing trained models. However, more recent work argues that feature-space perturbations alone may oversimplify adversarial behavior in network security, where traffic semantics, protocol feasibility, and generation constraints matter. [16], for example, emphasize realistic adversarial threat modeling for NIDS and show that the path from mathematical perturbation to operationally plausible network manipulation is nontrivial. Follow-on case studies have reinforced the need to evaluate robustness more carefully on multiclass IDS tasks and within settings that better reflect how attacks might unfold in practice [18, 19].

Defense-oriented research has largely responded through adversarial training, robust optimization, architecture redesign, data augmentation, or ensemble strategies.

These approaches can improve resilience under specific attack assumptions, but they also bring trade-offs in computation, generalization, and attack coverage. For instance, [20] combine adversarial robustness and explainability in a deep learning-based NIDS and show that adversarial training can improve resilience against attacks such as FGSM and PGD. More recent work has also started to explore hybrid or credibility-based defense mechanisms that attempt to strengthen robustness without relying solely on retraining [21]. Even so, the dominant pattern in the literature remains classifier-centric, the defensive objective is usually to preserve predictive performance under attack, whereas less attention is given to whether the attack leaves exploitable traces in the model's explanation behavior [20, 21].

The expansion of deep learning in IDS has intensified interest in explainable AI because security analysts often need more than a class label or confidence score to make defensible operational decisions. XAI methods are therefore increasingly used to reveal feature importance, clarify model reasoning, and improve trust in analyst-facing security systems. Recent systematic reviews confirm that explainability has become a major research theme in IDS, especially as institutions seek to balance predictive sophistication with transparency, accountability, and human interpretability [10, 22]. Broader cybersecurity reviews reach a similar conclusion, noting that explainability is increasingly treated as necessary for black-box security models deployed in critical environments [23].

In practice, SHAP and LIME are among the most widely used post hoc explanation methods in IDS. Their appeal lies in the ability to approximate local decision logic and expose which features most strongly contribute to a given classification. Representative IDS studies have used SHAP and LIME to interpret black-box classifiers, improve analyst understanding of IoT intrusion detection outputs, and compare explanation quality across model families [24–26]. Other recent work has extended this trend by integrating explanation layers into diverse IDS pipelines, including deep learning and hybrid architectures, further reinforcing that interpretability is now a mainstream objective rather than a peripheral add-on [27].

However, the dominant framing of XAI in IDS remains largely interpretive. Most studies ask whether explanations help humans understand or trust model predictions, whether they reveal salient features, or whether they can improve forensic analysis and response. This body of work is valuable, but it leaves an important unanswered question for adversarial settings. If an input has been strategically manipulated, then the issue is not only whether the explanation is understandable, but also whether the explanation remains dependable. Recent surveys on XAI in IDS and cybersecurity increasingly recognize this tension and note that transparency alone is insufficient when explanation mechanisms themselves may be unstable, misaligned, or vulnerable to manipulation [10, 22, 23].

A smaller but highly relevant line of research asks whether explanation behavior can be used not only to interpret model predictions but also to detect adversarial inputs. This perspective is important because adversarial perturbations may not always be obvious from output scores alone, yet they may still induce measurable changes in attribution patterns. A notable example is the use of SHAP signatures to distinguish clean from adversarial inputs, where explanation patterns themselves become the basis for a secondary detector rather than merely a descriptive tool [28]. This idea is conceptually attractive for IDS because manipulated traffic instances may remain close to normal input manifolds in feature space while still producing unstable or atypical explanation structure.

At the same time, explanation-based defense is not straightforward because explanation methods can themselves be attacked. [29] show that post hoc explanation techniques such as SHAP and LIME can be fooled, producing plausible but misleading interpretations even when the underlying model behavior is problematic. More recent reviews of adversarial XAI likewise emphasize that explanation reliability under attack is unresolved and that interpretable outputs should not be assumed secure merely because they appear coherent to a human observer [30]. Related survey work on the dual nature of XAI in IDS and broader explainability taxonomies similarly underscores that the same mechanisms used to improve transparency can also introduce new attack surfaces or false confidence if not evaluated critically [31, 32].

Taken together, these studies imply a productive tension. On one hand, adversarial perturbations may distort explanation structure in ways that reveal manipulation. On the other hand, explanation mechanisms may themselves be susceptible to deception. The resulting challenge is therefore not to assume explanations are inherently trustworthy, but to evaluate whether explanation behavior exhibits stable, measurable, and security-relevant deviations under attack. In IDS, this issue is particularly consequential because explanation methods are increasingly positioned as tools for analyst trust, model governance, and forensic support. Yet explicit explanation-driven adversarial detection remains much less developed in IDS than in image-classification oriented adversarial ML research [28] [30] [22].

The literature therefore supports three broad conclusions. First, deep learning has become central to IDS research, and recurrent models such as LSTM and GRU are established choices for modeling complex intrusion patterns. Second, adversarial machine learning has demonstrated that high predictive performance on clean benchmarks is not a sufficient indicator of security robustness. Third, explainable AI

has become increasingly important in IDS, but explanation reliability under attack remains insufficiently understood, particularly when explanations are treated as part of the defensive workflow rather than as a post hoc reporting layer [9, 10, 16].

What remains underdeveloped is the intersection of these strands in a comparative recurrent IDS setting. Although prior studies demonstrate the usefulness of LSTM and GRU for intrusion detection, direct comparative analyses of their adversarial robustness under a unified evaluation protocol remain limited. The same is true for explanation-centered defenses. IDS research has used SHAP and related methods to improve interpretability, and adversarial machine learning research has shown that explanation instability can help reveal manipulated inputs, but comparatively fewer studies bring these ideas together to determine whether SHAP explanation drift can serve as a practical monitoring signal across recurrent IDS architectures. This gap motivates the present study, which evaluates LSTM and GRU based IDS models under white-box adversarial perturbation using a common temporal flow-window pipeline and investigates whether changes in SHAP attribution patterns and top-k feature stability can function as indicators of adversarial manipulation. In this way, the study moves explainability beyond a purely interpretive role and positions it as a potential operational component of IDS defense.

This study is motivated by that gap. Rather than treating explainability only as a post hoc aid for interpreting model predictions, it examines whether SHAP explanation drift and top-k feature instability can be used to detect adversarially manipulated inputs in LSTM and GRU based IDS models. To strengthen this perspective, the study further evaluates a lightweight learned detector built on explanation-deviation features, allowing explanation-based monitoring to be assessed not only through fixed thresholds but also through data-driven discrimination. In doing so, the study contributes to a more security-aware understanding of explainability in IDS: explanations are not only tools for human interpretation, but may also function as measurable indicators of adversarial interference. This framing provides the foundation for the methodology presented in the next section.

3 Methodology

The proposed framework consists of five main phases: dataset preprocessing and temporal flow-window construction, recurrent IDS model development, adversarial robustness evaluation, SHAP-based explanation analysis, and explanation-driven adversarial detection, as illustrated in Fig. 1. The overall design was structured to support a controlled comparison between Long Short-Term Memory and Gated Recurrent Unit models while simultaneously evaluating their robustness to adversarial perturbation and the usefulness of explanation instability as a secondary detection signal.

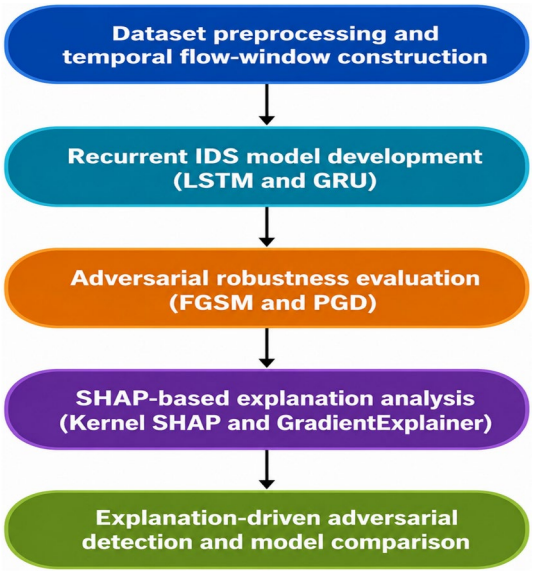

Fig. 1 Main Methodological Phases of the Proposed Adversarial Robustness and SHAP-Based Detection Framework

3.1 Dataset preprocessing

This study used a publicly available dataset, the TON_IoT network dataset which captures the complex status of the current digital world [33]. The final column in this dataset is treated as the target label, and all remaining columns are treated as predictor variables. The class inventory in the implementation includes Backdoor attack, Denial of Service (DoS) attack, Distributed Denial of Service (DDoS) attack, Injection attack, Man-In-The-Middle (MITM) attack, Normal, Password cracking attack, Ransomware attack, Scanning attack, and Cross-site Scripting (XSS) attack. Accordingly, the learning problem in this work is formulated as a 10-class supervised classification task in which each network-flow record is assigned to exactly one class. This design moves beyond binary benign-versus-malicious discrimination and instead requires the model to differentiate among multiple attack behaviours and normal traffic. That formulation is appropriate for the selected dataset family because ToN-IoT was explicitly designed to support multiclass intrusion analysis through a label field and attack-type subclasses rather than only coarse anomaly labelling [34]. The target label was extracted from the multiclass attack-type field and encoded into integer class

indices using a label encoder. Numerical variables were imputed using the median and standardized, while categorical variables were imputed using the most frequent value and ordinal encoded. Numerical standardization was performed using the standard score transformation in Eq. (1):

$$z = \frac{x-u}{s} \tag{1}$$

where x denotes the original feature value, μ is the training-set mean, and s is the training-set standard deviation.

The data were organized into true temporal flow-window sequences. Specifically, records were sorted by timestamp together with grouping fields derived from network context, namely source IP and destination IP, after which overlapping sliding windows of length 5 and stride 1 were constructed. Each resulting sample therefore represented a short ordered temporal segment of consecutive flow observations rather than an artificial partition of a single record. The label assigned to each window corresponded to the class of the final record in the sequence. This design preserved local temporal order and enabled the recurrent models to operate on temporally structured flow behavior.

After temporal window construction, the resulting sequence tensor was partitioned into training and test subsets using a stratified 80:20 split with a fixed random state of 42 in order to preserve class proportions. The final recurrent input representation can be expressed as Nx T x F, where N denotes the number of samples, $T = 5$ denotes the temporal window length, and F denotes the processed feature dimension per time step.

3.2 Recurrent IDS model development

Two recurrent neural architectures were evaluated in this work, namely Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Both models were implemented using a matched architectural design so that any observed differences in clean classification performance, adversarial degradation, and explanation instability could be attributed primarily to the recurrent unit rather than to downstream architectural variation.

The recurrent input to both models had shape (T, F), where $T = 5$ denotes the temporal window length and F denotes the number of processed features per time step. The LSTM-based IDS used a single LSTM layer with 64 hidden units, followed by a dropout layer with rate 0.3, a dense hidden layer with 32 ReLU units, and a final softmax output layer with one unit per class. The GRU-based IDS followed the same downstream design, differing only in the replacement of the LSTM layer with a GRU layer of 64 hidden units.

The architecture of the LSTM model can be summarized as follows:

- Input layer with shape $(5, F)$.
- LSTM layer with 64 hidden units
- Dropout layer with rate 0.3
- Dense hidden layer with 32 units and ReLU activation
- Output dense layer with (C) units and softmax activation.

The architecture of the GRU model can be summarized as follows:

- Input layer with shape $(5, F)$
- GRU layer with 64 hidden units
- Dropout layer with rate 0.3
- Dense hidden layer with 32 units and ReLU activation
- Output dense layer with (C) units and softmax activation

where (C) denotes the number of traffic classes.

Both models were compiled using the Adam optimizer and categorical cross-entropy loss, with accuracy as the primary optimization metric. Training was conducted under identical settings in order to preserve experimental comparability:

- Optimizer: Adam
- Loss: categorical cross-entropy
- Metric: accuracy
- Epochs: 10
- Batch size: 64
- Validation split: 0.1

3.3 Adversarial robustness evaluation

To evaluate robustness under deliberate test-time perturbation, the trained IDS models were assessed under a white-box untargeted evasion threat model. Under this setting, the adversary is assumed to have access to the classifier and to the gradients of the loss with respect to the input features. The objective is to perturb the standardized feature-space representation of a sample so as to induce misclassification, rather than to poison training data or force prediction into a specific target class. As the baseline adversarial condition, adversarial examples were first generated using the Fast Gradient Sign Method (FGSM) through the Adversarial Robustness Toolbox (ART). ART was used to wrap the trained Keras models using a classifier interface compatible with gradient-based evasion attacks. For an input sample x with true label y, the adversarial example x' generated by FGSM is defined in Eq. (2):

$$x' = x + \in \cdot sign\left(\nabla_x J(\theta, x, y)\right) \tag{2}$$

where x is the original input sample, x' is the adversarially perturbed sample, $\in$ is the perturbation magnitude, $\mathrm{J}(\theta, x, y)$ is the loss function of model parameters θ on input x and

label y, $\nabla_x J(\theta, x, y)$ is the gradient of the loss with respect to the input, sign(·) returns the element-wise sign of the gradient.

$\in$ regulates the trade-off between attack subtlety and attack strength. Smaller values preserve closer proximity to the original sample but may produce weaker evasion effects, whereas larger values generally increase attack success at the cost of greater distortion. In this study, $\in = 0.1$ was used to serve as a moderate perturbation budget for testing whether relatively small changes in the standardized input space were sufficient to alter predictions and explanations.

To broaden robustness stress testing beyond a single-step attack, the evaluation protocol was extended through a multi-attack robustness sweep including both FGSM and Projected Gradient Descent (PGD). FGSM was evaluated at multiple perturbation budgets, specifically $\in \{0.01, 0.03, 0.05, 0.10\}$. In addition, representative PGD settings were included to provide a stronger iterative first-order robustness assessment. The broader sweep was executed on a stratified evaluation subset to preserve class representation while maintaining computational tractability.

For each attack configuration, the following performance indicators were recorded: clean accuracy, adversarial accuracy, weighted F1-score, macro F1-score, evasion rate, and, for PGD, attack success on initially correct samples. This extended robustness protocol was introduced to provide a more informative estimate of model vulnerability than single-step FGSM alone.

Adversarial evaluation in this study therefore comprised three linked components. First, clean and adversarial predictive outputs were compared to quantify performance degradation. Second, label changes induced by perturbation were used to estimate evasion effectiveness. Third, adversarial samples were passed into the explanation pipeline to determine whether adversarial perturbation induced systematic SHAP instability that could be exploited by a secondary detector.

3.4 SHAP-based explanation analysis

The proposed defense is based on the hypothesis that adversarial perturbations may alter not only the predicted class of a traffic sample but also the pattern of feature attributions used to justify that prediction. Explanation behavior was therefore treated as an additional observation space for adversarial analysis. Instead of relying solely on output labels or confidence scores, the framework computed SHAP explanations for matched clean and adversarial samples and measured the degree to which those explanations diverged.

3.4.1 Generation of clean and adversarial SHAP explanations

Let f denote the trained multiclass IDS model, and let x denote a preprocessed temporal input sequence. For each clean sample x, an adversarial counterpart x' was first generated through the attack pipeline. Because explanation analysis was performed over feature dimensions, the recurrent tensor was flattened into a two-dimensional feature representation before SHAP computation.

A background subset B drawn from clean training data was supplied to the explainer as the reference distribution. Using the prediction function of the trained IDS model and the background set B, SHAP attribution vectors were then computed for matched clean and adversarial samples. Let $\emptyset_i^{(c)}(x)$ denote the SHAP attribution of feature i for class c on sample x. The same procedure was repeated for the adversarial counterpart x', thereby enabling pairwise comparison between clean and adversarial explanation profiles.

3.4.2 Explanation drift based on mean absolute SHAP difference

The first explanation-based detection component was the explanation drift score, defined as the mean absolute difference between the clean and adversarial SHAP attribution vectors for the comparison class. The explanation drift for sample x is defined as

$$D(x) = \frac{1}{F} \sum_{j=1}^{F} \left| \emptyset_i^{(c)}(x) - \emptyset_i^{(c)}\left(x'\right) \right| \tag{3}$$

where F denotes the total feature dimension and ccc denotes the comparison class. Larger values of $D(x)$ indicate greater attribution displacement induced by adversarial perturbation.

3.4.3 Top-k feature overlap

The second explanation-based component measured whether the dominant explanatory features remained stable between clean and adversarial explanations. Let $S_k(x)$ denote the set of indices corresponding to the K features with the largest absolute SHAP values in the clean explanation, and let $S_k\left(x'\right)$ denote the corresponding set for the adversarial explanation. The normalized top-k feature overlap was defined as in Eq. (4):

$$O_k(x) = \frac{\left|S_k(x) \cap S_k\left(x'\right)\right|}{K} \tag{4}$$

By construction, $0 \leq O_k(x) \leq 1$. Values near 1 indicate that the dominant explanatory factors remain largely unchanged, whereas values near 0 indicate substantial instability in feature ranking. The explanation drift score and overlap score capture complementary aspects of explanation behavior: the

former reflects global attribution displacement, while the latter focuses on stability among the most influential explanatory features.

3.4.4 Two-stage SHAP evaluation strategy

Because Kernel SHAP is computationally expensive, explanation analysis was conducted in two stages. First, a small-sample Kernel SHAP branch was used on 100 matched clean and adversarial samples with a background set of 50 flattened training instances and *nsamples* = 100. This branch was used for detailed explanation comparison and figure *generation*.

Second, a larger GradientExplainer branch was introduced to evaluate explanation-driven detection on a substantially larger stratified subset. This branch was specifically included to reduce the dependence of detector evaluation on the small-sample Kernel SHAP analysis and to provide a more statistically robust estimate of detector performance.

3.5 Explanation-driven adversarial detection

3.5.1 Rule-based detector

The baseline explanation-driven detector combined the SHAP drift score and top-k feature overlap score in a rule-based formulation. Let $D(x)$ denote the drift score and $O_k(x)$ denote the overlap score for sample x A sample was flagged as adversarial when either the drift exceeded a selected threshold τ_D or the overlap fell below a predefined threshold τ_O, as shown in Eq. (5)

$$\hat{z}(x) = \begin{cases} 0, & \text{if } D(x) > \tau_D \text{ or } O_k(x) < \tau_O \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

where $\hat{z}(x) = 1$ denotes an adversarial sample and $\hat{z}(x) = 0$ denotes a non-adversarial sample. In the baseline detector, the overlap threshold was fixed at τ_O, while the drift threshold was selected from the empirical drift-score distribution.

3.5.2 Learned explanation-deviation detector

To reduce dependence on rigid fixed-threshold logic, the framework was extended with a learned secondary detector. A lightweight logistic regression classifier was trained on explanation-deviation features derived from matched clean and adversarial explanation pairs. These features included mean drift, drift dispersion, maximum attribution shift, L2 drift magnitude, top-k overlap, and the proportion of large attribution changes. The learned detector was introduced to provide a data-driven decision boundary over explanation instability features and to offer a more adaptable alternative to static thresholding in dynamic network environments.

3.6 Experimental Setup

All experiments were conducted in a Python-based machine learning environment using TensorFlow Keras for recurrent model development, scikit-learn for preprocessing and data partitioning, the Adversarial Robustness Toolbox for adversarial example generation, and SHAP for post hoc explanation analysis. A common experimental pipeline was maintained for both recurrent architectures to ensure that any observed differences were attributable to model design rather than to inconsistencies in preprocessing, training, or evaluation.

After preprocessing and temporal flow-window construction, the dataset was partitioned into stratified training and test subsets using an 80:20 split with a fixed random state of 42. The LSTM and GRU models were then trained using one-hot encoded labels, the Adam optimizer, categorical cross-entropy loss, batch size 64, validation split 0.1, and 10 training epochs. The LSTM model comprised a single LSTM layer with 64 units, followed by dropout of 0.3, a dense hidden layer with 32 ReLU units, and a softmax output layer. The GRU model used the same downstream structure, differing only in the substitution of the recurrent layer with a GRU layer of 64 units.

Following training, each model was wrapped as an ART Keras classifier. Clean performance was first evaluated on the held-out test set. FGSM at $\epsilon = 0.1$ was then used as the main adversarial branch for detailed explanation analysis. In parallel, a broader robustness sweep was executed across multiple FGSM and PGD settings in order to compare degradation trends under both single-step and iterative perturbation regimes.

For explanation-based detection, the small-sample Kernel SHAP branch used the first 100 matched clean and adversarial samples together with a background reference set of 50 flattened training samples and *nsamples* = 100. Detector labels on this subset were defined by whether the model prediction changed between the clean and adversarial versions of the same sample. The larger GradientExplainer branch evaluated the same explanation-deviation logic on a larger stratified subset drawn from the held-out test set. Under this unified protocol, the LSTM and GRU models were compared in terms of clean multiclass classification performance, adversarial susceptibility under multiple attack settings, and explanation-based detection effectiveness using both rule-based and learned detectors.

4 Results

4.1 Clean model performance

Both recurrent models achieved strong clean-label performance on the multiclass temporal flow-window intrusion detection task. The LSTM attained a clean accuracy of 0.9614 with a weighted F1-score of 0.9597, while the GRU attained a clean accuracy of 0.9615 with a weighted F1-score of 0.9598. These results indicate that both architectures learned the multiclass classification problem effectively under benign conditions and provided a strong baseline for adversarial robustness analysis. Although overall clean performance was high, the clean baseline should not be interpreted as uniformly strong across all classes. As discussed in the methodology and reflected by the model-comparison setup, minority attack categories remained inherently more difficult than the dominant classes, and this became more evident once adversarial perturbation was introduced. Table 1 is a summary of the two model's performance prior to adversarial perturbation.

4.2 Adversarial attack evaluation

Under the FGSM adversarial attack at $\epsilon = 0.1$, both models exhibited substantial degradation relative to their clean baselines. The LSTM declined from 0.9614 clean accuracy to 0.6094 adversarial accuracy, with its weighted F1-score falling from 0.9597 to 0.6290 and its evasion rate reaching 0.3738. The GRU declined from 0.9615 clean accuracy to 0.5130 adversarial accuracy, with its weighted F1-score falling from 0.9598 to 0.5690 and its evasion rate increasing to 0.4702. These results confirm that both recurrent architectures remained vulnerable under white-box evasion, despite their strong benign classification

Table 1 Comparative classification performance of GRU and LSTM models

Class	LSTM Precision	LSTM Recall	LSTM F1	GRU Precision	GRU Recall	GRU F1
backdoor	0.99	1.00	1.00	0.99	1.00	1.00
ddos	0.98	0.72	0.83	0.98	0.73	0.84
dos	1.00	0.92	0.96	1.00	0.92	0.96
injection	0.87	0.80	0.83	0.89	0.76	0.82
mitm	0.52	0.12	0.19	0.56	0.14	0.22
normal	0.97	1.00	0.98	0.97	1.00	0.98
password	0.84	0.88	0.86	0.81	0.91	0.85
ransomware	1.00	1.00	1.00	1.00	1.00	1.00
scanning	1.00	0.96	0.98	1.00	0.96	0.98
xss	0.97	0.91	0.94	0.97	0.92	0.95
accuracy			0.96			0.96
macro avg	0.91	0.83	0.86	0.92	0.83	0.86
weighted avg	0.96	0.96	0.96	0.96	0.96	0.96

Table 2 Comparative adversarial performance of GRU and LSTM models across attack classes under FGSM

Class	GRU Precision	GRU Recall	GRU F1	LSTM Precision	LSTM Recall	LSTM F1
backdoor	0.00	0.00	0.00	0.01	0.00	0.00
ddos	0.03	0.02	0.02	0.07	0.04	0.05
dos	0.70	0.87	0.78	0.94	0.88	0.91
injection	0.01	0.01	0.01	0.00	0.01	0.00
mitm	0.00	0.01	0.00	0.00	0.04	0.00
normal	0.94	0.71	0.81	0.92	0.85	0.88
password	0.05	0.09	0.07	0.07	0.12	0.09
ransomware	0.01	0.03	0.02	0.16	0.14	0.15
scanning	0.25	0.28	0.26	0.31	0.45	0.36
xss	0.14	0.15	0.15	0.01	0.01	0.01
accuracy			0.51			0.61
macro avg	0.21	0.22	0.21	0.25	0.25	0.25
weighted avg	0.65	0.51	0.57	0.65	0.61	0.63
evasion rate			47.02%			37.38%

performance. Table 2 summarizes the comparative adversarial performance for both models. Relative to the GRU, the LSTM retained higher adversarial accuracy and weighted F1-score under the main FGSM baseline, whereas the GRU exhibited a higher evasion rate. This indicates that the LSTM was comparatively more stable under the principal FGSM condition used for the explanation-analysis branch.

4.3 Broader robustness stress test across FGSM and PGD

To move beyond a single-step attack setting, the broader robustness sweep evaluated both recurrent models under multiple FGSM budgets and representative PGD conditions. The resulting trends showed that adversarial degradation increased progressively with perturbation strength and that iterative PGD produced stronger degradation than FGSM at comparable or nearby settings. This broader sweep is summarized in Table 3. The sweep configuration itself is reflected in the updated heavy notebook pipeline, which includes both FGSM and PGD entries in the broader robustness schedule.

For the GRU, adversarial accuracy decreased from 0.8978 at FGSM $\epsilon = 0.01$ to 0.7778 at $\epsilon = 0.03$, 0.7115 at $\epsilon = 0.05$, and 0.5165 at $\epsilon = 0.10$. The corresponding adversarial weighted F1-scores declined from 0.8959 to 0.7646, 0.7058, and 0.5731, while evasion rate increased from 0.0954 to 0.2117, 0.2774, and 0.4679. Under PGD, the GRU degraded further, reaching adversarial accuracy of 0.6583 and weighted F1-score of 0.6928 at $\epsilon = 0.03$, and 0.3520 and 0.4449, respectively, at $\epsilon = 0.10$. At the strongest PGD setting, attack success on initially correct samples reached 0.6429. These results show that the GRU remained particularly sensitive to stronger iterative perturbation.

For the LSTM, adversarial accuracy decreased from 0.8983 at FGSM $\epsilon = 0.01$ to 0.8045 at $\epsilon = 0.03$, 0.7575 at $\epsilon = 0.05$, and 0.6045 at $\epsilon = 0.10$. Its adversarial weighted F1-scores declined from 0.8919 to 0.7979, 0.7527, and 0.6271, while evasion rate increased from 0.0928 to 0.1915, 0.2418, and 0.3808. Under PGD, the LSTM declined to 0.6784 adversarial accuracy with weighted F1-score of 0.7073 at $\epsilon = 0.03$, and to 0.3277 adversarial accuracy with weighted F1-score of 0.4001 at $\epsilon = 0.10$. At the strongest PGD condition, attack success on initially correct samples reached 0.6664. As with the GRU, the iterative PGD setting exposed greater structural vulnerability than FGSM alone.

Taken together, the broader robustness sweep demonstrates that single-step FGSM underestimates adversarial fragility in both recurrent models. PGD produced stronger degradation than FGSM, especially at higher perturbation budgets. The robustness sweep table in the notebook captures the full set of attack-specific summary metrics.

It should be noted that the clean performance on the robustness-sweep subset remained stable for both models at approximately 0.96 accuracy and 0.96 weighted F1, and was omitted in the table above for compactness.

4.4 Class-level robustness under the main FGSM setting

Class-level degradation under FGSM was highly uneven across the ten traffic categories. In the GRU setting, the normal class remained comparatively more stable, with precision 0.79, recall 0.79, and F1-score 0.79. Backdoor and scanning retained partial detectability, with F1-scores of 0.67 and 0.53, respectively. By contrast, several classes collapsed almost completely, including ddos with F1-score

Table 3 Robustness sweep across FGSM and PGD settings for LSTM and GRU

Model	Attack	ϵ	Adversarial Accuracy	Adversarial Weighted F1	Adversarial Macro F1	Evasion Rate	Attack Success on Initially Correct
GRU	FGSM	0.01	0.8978	0.8959	0.7077	0.0954	
GRU	FGSM	0.03	0.7778	0.7646	0.4143	0.2117	
GRU	FGSM	0.05	0.7115	0.7058	0.3113	0.2774	
GRU	FGSM	0.10	0.5165	0.5731	0.2159	0.4679	
GRU	PGD	0.03	0.7483	0.7371	0.3551	0.2414	0.2323
GRU	PGD	0.10	0.3520	0.4351	0.1432	0.6362	0.6429
LSTM	FGSM	0.01	0.8983	0.8914	0.6990	0.0860	
LSTM	FGSM	0.03	0.7372	0.7377	0.3895	0.2470	
LSTM	FGSM	0.05	0.6611	0.6727	0.2974	0.3231	
LSTM	FGSM	0.10	0.6045	0.6260	0.2444	0.3795	
LSTM	PGD	0.03	0.6365	0.6606	0.3134	0.3478	0.3446
LSTM	PGD	0.10	0.3277	0.4099	0.0813	0.6573	0.6664

0.01, injection with F1-score 0.00, password with F1-score 0.00, and xss with F1-score 0.00. The dos class also degraded substantially, with F1-score 0.14. In the LSTM setting, the normal class again remained the most stable, with precision 0.78, recall 0.83, and F1-score 0.81, while backdoor retained partial performance with F1-score 0.51. However, dos, injection, password, ddos, and xss all degraded to near-zero F1-scores, and scanning fell to an F1-score of 0.07. In both models, the mitm class exhibited unusually high recall but extremely low precision, indicating that many perturbed samples from other classes were reassigned to this label rather than that the class itself remained robust (Table 4).

These results show that adversarial degradation was not uniform and disproportionately affected several minority attack classes. Accordingly, weighted summary metrics should be interpreted alongside class-level results.

Table 4 Small-sample SHAP detector performance for rule-based and learned detectors in the LSTM and GRU models

Model	Detector	ROC-AUC	F1-score
LSTM	Rule-based	0.8150	0.5778
LSTM	Learned	0.8843	0.7273
GRU	Rule-based	0.7523	0.6667
GRU	Learned	0.8705	0.7586

4.5 SHAP explanation behavior under clean and adversarial conditions

The SHAP analyses showed that adversarial perturbation induced measurable explanation instability in both recurrent architectures. This instability was evaluated using two complementary indicators: the SHAP drift score, which measures global attribution displacement between clean and adversarial explanations, and the top-k feature overlap score, which measures the stability of the most influential explanatory features.

The drift histograms for the GRU and LSTM are shown in Fig. 2 and Fig. 3, respectively. In both models, adversarial samples concentrated in higher-drift regions than normal samples, indicating that adversarial perturbation altered the local explanation structure of the classifier. However, the separation was only partial, and overlap remained between normal and adversarial distributions in both architectures. In the GRU setting, the adversarial distribution formed a tighter cluster around the mid-drift region, while the normal distribution was more dispersed and extended further into both lower and higher drift ranges. In the LSTM setting, a similar pattern was observed, with adversarial samples again concentrated in a narrower mid-drift band and normal samples spread more broadly across the drift axis. These results indicate that adversarial perturbation produced systematic explanation displacement, but not a perfectly separable drift signature.

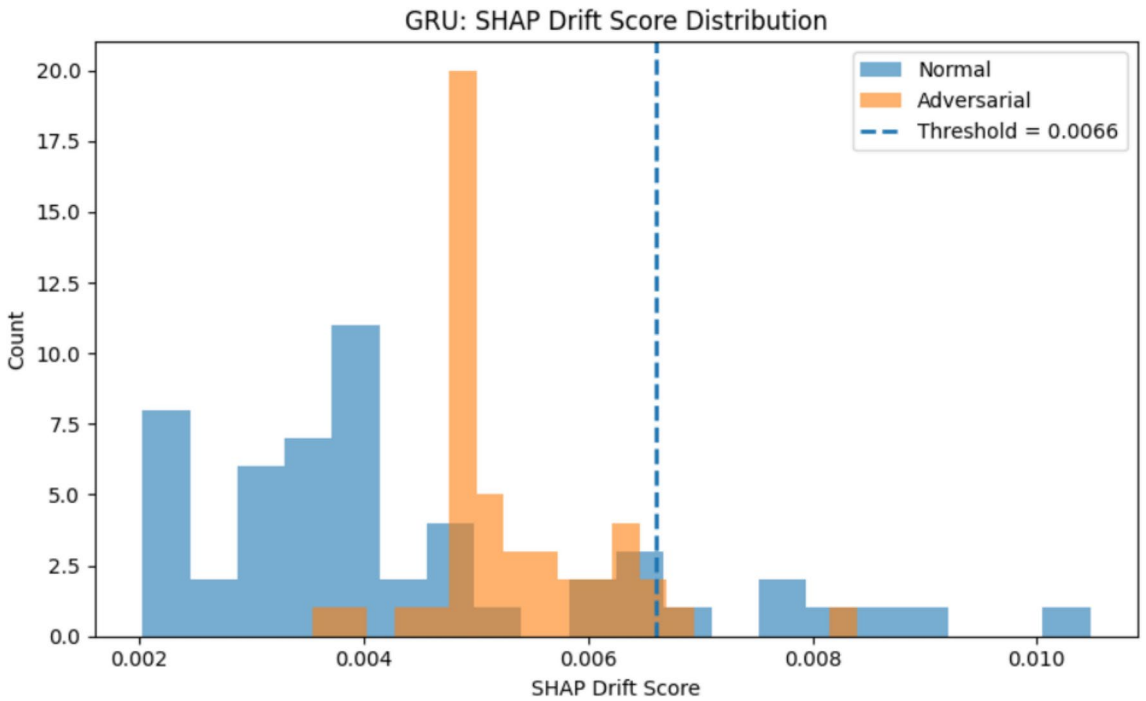

Fig. 2 Distribution of SHAP drift scores for normal and adversarial samples in the GRU-based analysis

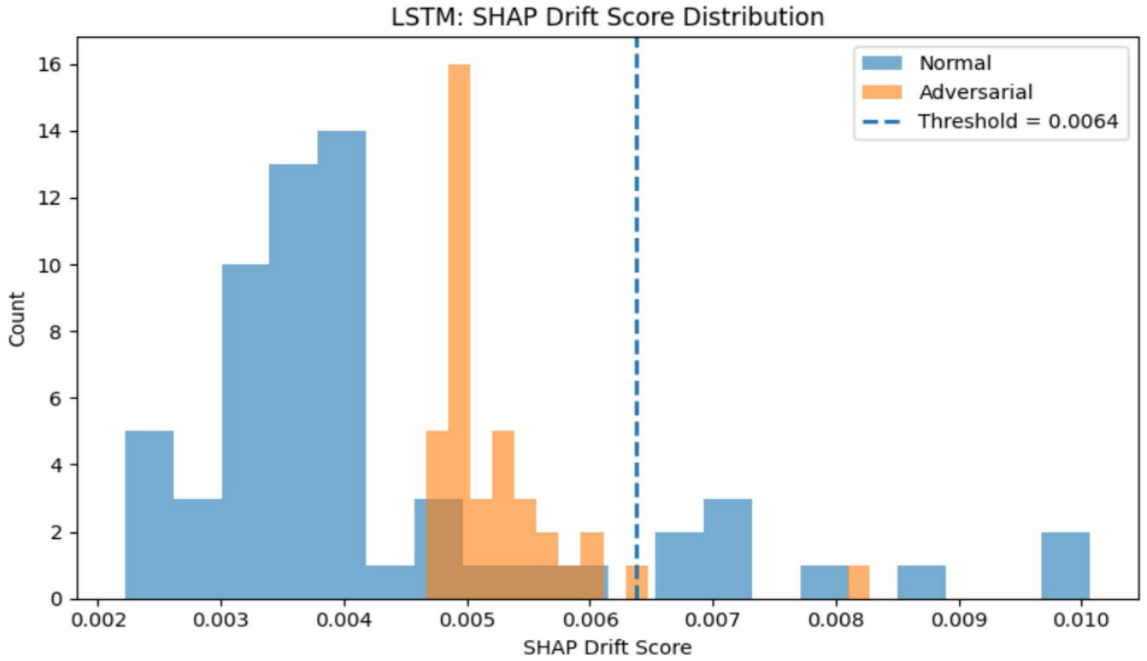

Fig. 3 Distribution of SHAP drift scores for normal and adversarial samples in the LSTM-based analysis

The top-k overlap distributions for the GRU and LSTM are shown in Fig. 4 and Fig. 5, respectively. In both models, the overlap scores were concentrated mainly in the lower ranges, especially between 0.0 and 0.4, indicating that adversarial perturbation frequently altered the ranking of the most influential features. The GRU showed a slightly stronger concentration of samples in the lowest overlap bands, whereas the LSTM retained a somewhat larger

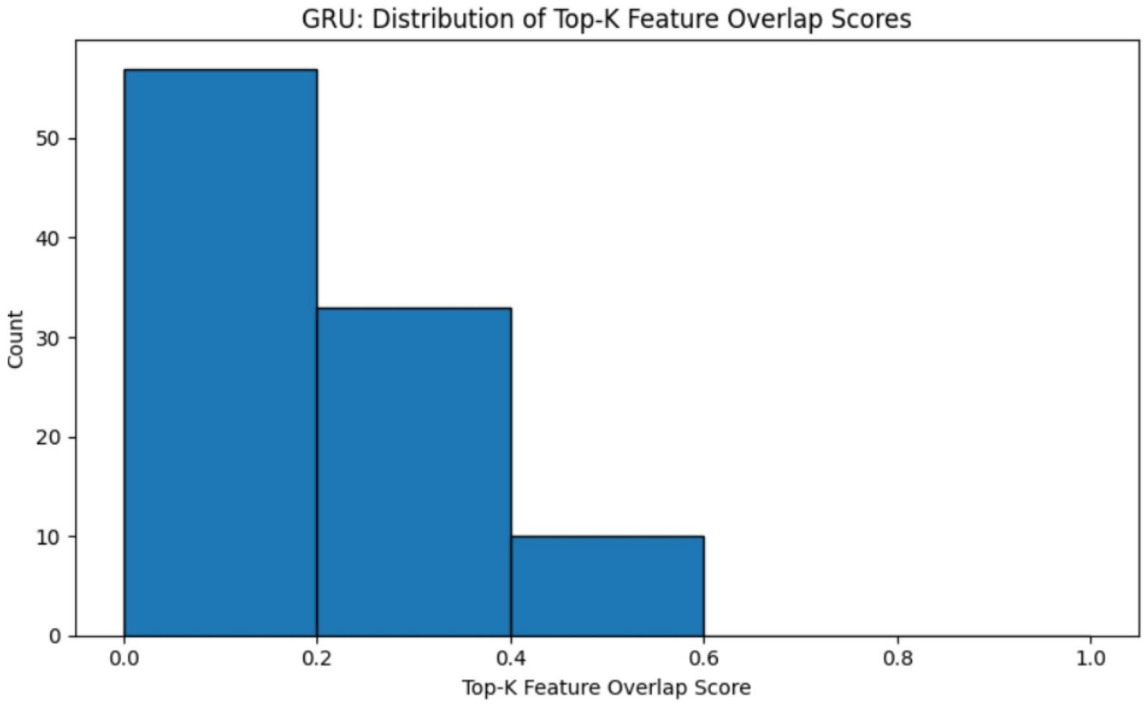

Fig. 4 Distribution of top-k feature overlap scores between clean and adversarial SHAP explanations for the GRU model

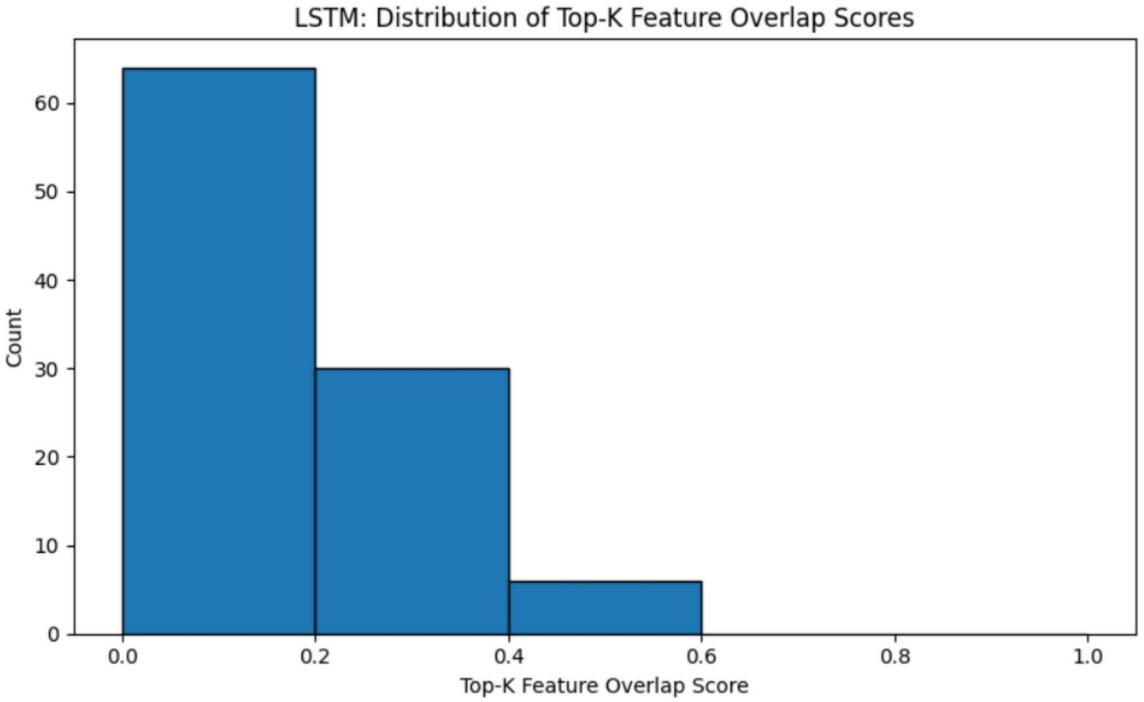

Fig. 5 Distribution of top-k feature overlap scores between clean and adversarial SHAP explanations for the LSTM model

proportion of samples at moderate overlap levels. This suggests that adversarial perturbation induced rank-level explanation instability in both architectures, with the GRU exhibiting somewhat stronger reordering of salient features and the LSTM preserving slightly more top-ranked explanatory consistency.

Taken together, the updated SHAP figures show that adversarial manipulation affected explanation behavior in two related but distinct ways: by shifting the overall attribution profile and by reordering the most influential features. These results support the use of both drift magnitude and top-k overlap as complementary explanation-based signals for adversarial monitoring.

4.6 Small-sample SHAP detector performance

The small-sample detector analysis, based on the Kernel SHAP branch, compared a rule-based baseline with a learned secondary detector trained on explanation-deviation features. For the LSTM, the clean accuracy and main FGSM adversarial accuracy entering this branch were 0.9614 and 0.6094, respectively, with clean and adversarial weighted F1-scores of 0.9597 and 0.6290 and evasion rate of 0.3738. The rule-based detector achieved ROC-AUC of 0.8150 and F1-score of 0.5778, while the learned detector improved performance to ROC-AUC of 0.8843 and F1-score of 0.7273. For the GRU, the corresponding clean and main FGSM adversarial accuracies were 0.9615 and 0.5130, with clean and adversarial weighted F1-scores of 0.9598 and 0.5690 and evasion rate of 0.4702. The rule-based detector achieved ROC-AUC of 0.7523 and F1-score of 0.6667, while the learned detector improved detector ROC-AUC to 0.8705 and F1-score to 0.7586.

These findings indicate that explanation-deviation features retained useful adversarial signal in both architectures, but that rigid thresholding did not fully exploit that signal. The learned detector provided a stronger and more flexible decision boundary than the fixed-threshold rule-based baseline.

4.7 Large-scale GradientExplainer detector performance

To reduce dependence on the 100-sample Kernel SHAP analysis, the detector was also evaluated on a larger 2,000-sample GradientExplainer branch. In this setting, the LSTM-based large-scale rule detector achieved ROC-AUC of 0.7047, accuracy of 0.7995, precision of 0.7318, recall of 0.7209, specificity of 0.8455, and F1-score of 0.7263. The GRU-based large-scale detector achieved ROC-AUC

Table 5 Large-scale GradientExplainer detector performance for the LSTM and GRU models

Model	ROC-AUC	Accuracy	Precision	Recall	Specificity	F1-score
LSTM	0.7047	0.7995	0.7318	0.7209	0.8455	0.7263
GRU	0.7323	0.7915	0.8212	0.7143	0.8607	0.7640

of 0.7323, accuracy of 0.7915, precision of 0.8212, recall of 0.7143, specificity of 0.8607, and F1-score of 0.7640. These larger-scale results are summarized in Table 5. Compared with the small-sample branch, the large-scale detector results were more conservative, but they still demonstrated meaningful separability between clean and adversarial samples. In this larger setting, the GRU exhibited slightly stronger detector performance than the LSTM in terms of ROC-AUC, precision, specificity, and F1-score, whereas the LSTM retained slightly higher recall balance. Overall, the large-scale evaluation confirms that explanation drift remained informative at a greater scale, while also showing that detector effectiveness was lower than would have been suggested by the earlier, smaller-sample analysis alone.

Overall, the results support four main conclusions. First, both recurrent models achieved strong clean multiclass performance, but both were vulnerable under adversarial perturbation. Second, iterative PGD produced stronger degradation than FGSM, confirming that FGSM alone provides only a partial picture of robustness. Third, explanation drift and top-k feature instability provided useful signals for adversarial discrimination, although the signal was not perfectly separable. Fourth, the learned secondary detector outperformed the rigid rule-based baseline in the small-sample branch, while the larger GradientExplainer evaluation showed that explanation-driven detection remained informative at broader scale, with the GRU showing a modest advantage in large-sample detector separability.

5 Discussion

5.1 Clean performance does not imply adversarial robustness

A central finding of this study is that strong clean-label classification performance does not guarantee adversarial robustness. Both recurrent models achieved high clean multiclass performance on the temporal flow-window intrusion detection task, yet both degraded substantially once adversarial perturbation was introduced. This confirms that conventional benchmark accuracy alone is insufficient as evidence of deployment readiness in adversarial environments. In practical terms, a model that performs strongly under benign traffic may still remain highly vulnerable when an attacker is able to manipulate feature-space inputs at inference time.

The updated results also show that the relationship between clean performance and adversarial resilience is not linear. Although the LSTM and GRU performed similarly under clean evaluation, their robustness profiles diverged under perturbation. Under the principal FGSM full-test-set setting, the LSTM retained higher adversarial accuracy and weighted F1-score than the GRU, indicating stronger baseline classifier robustness in that condition. However, this advantage did not transfer uniformly to all downstream analyses. In the larger explanation-based detector evaluation, the GRU yielded slightly stronger detector separability and specificity. This suggests that classifier robustness and explanation-drift separability should be interpreted as related but distinct properties rather than as interchangeable indicators of model quality.

5.2 Broader robustness stress testing reveals stronger vulnerability than FGSM alone

The broader FGSM and PGD sweep confirms that single-step FGSM alone understates vulnerability. As perturbation budgets increased, both recurrent models exhibited progressively stronger degradation, and the iterative PGD settings produced larger declines than FGSM at comparable budgets. This pattern strengthens the interpretation of the study by showing that the observed fragility is not limited to one attack configuration, but persists and intensifies under stronger first-order optimization pressure. This broader stress-testing range is important because it addresses a key limitation of the earlier experimental framing. FGSM remains useful as a baseline because it provides a simple and reproducible first-order attack condition, but the broader sweep shows that it offers only a partial view of structural robustness. The sharper performance losses under PGD indicate that the decision surfaces learned by both recurrent models remain vulnerable to stronger iterative perturbation, even when clean performance is high. From a security perspective, this means that IDS evaluation should not rely solely on one-step evasion tests when the intended deployment context includes adversarially capable threat actors.

5.3 LSTM and GRU exhibit different robustness trade-offs

The results do not support a single universal winner between the two recurrent architectures. Instead, they show a trade-off between classifier resilience and explanation-based

detector separability. The LSTM performed better under the main FGSM baseline, retaining higher adversarial accuracy and lower evasion than the GRU. This indicates that, under the principal single-step perturbation condition used for the main explanation branch, the LSTM produced a more stable classification response.

By contrast, the GRU demonstrated a modest advantage in the larger-scale explanation-driven detector evaluation. Its larger-sample rule-based detector achieved slightly stronger ROC-AUC, specificity, and F1-score than the corresponding LSTM detector, indicating that adversarial perturbations in the GRU setting produced explanation shifts that were somewhat easier to distinguish from normal explanation behavior. This difference suggests that a model may degrade more under direct evasion while still exhibiting more discriminative explanation instability. Accordingly, the selection of a recurrent architecture for adversarially aware IDS deployment should not be based only on clean performance or only on adversarial accuracy, but should also consider the stability and detectability of its explanation behavior.

5.4 Class-level adversarial vulnerability is highly uneven

Adversarial degradation was not uniform across the ten traffic classes. In both recurrent models, the normal class remained comparatively more stable than several minority attack classes, while multiple underrepresented categories degraded sharply under perturbation. In particular, classes such as ddos, injection, password, and xss collapsed to near-zero F1-scores in the main FGSM evaluation, and this broader pattern of fragility remained visible in the multi-attack sweep. These findings show that adversarial vulnerability is not a single global property of the classifier, but is instead distributed unevenly across local class regions in feature space. This class imbalance has important implications for interpretation. Weighted averages remained relatively higher than macro-level performance because the majority normal class retained partial stability, thereby sustaining aggregate metrics even when several attack categories became highly fragile. For IDS evaluation, this means that weighted F1-score alone can mask substantial operational risk. A model may still appear strong at aggregate level while becoming ineffective on precisely those minority attack classes that are most important for security monitoring. The evidence therefore reinforces the need to report both weighted and macro-level metrics and to interpret class-wise degradation directly.

5.5 Explanation drift remains a useful but imperfect adversarial signal

The SHAP analyses indicate that adversarial perturbation affected not only final predictions but also the explanatory structure underlying those predictions. In both architectures, perturbation redistributed attribution magnitude across the temporal flow-window representation, indicating that adversarial manipulation altered the local decision rationale rather than merely the output label. The use of true sliding flow-window sequences makes these explanation changes more interpretable in a temporally structured intrusion detection setting. The explanation-deviation results also show that drift and top-k feature overlap capture complementary aspects of explanation instability. Drift summarizes global attribution displacement, whereas top-k overlap measures whether the dominant explanatory features remain stable after perturbation. The results suggest that adversarial perturbation can simultaneously increase overall attribution displacement and reorder salient explanatory factors. This is important because an explanation-based detector built on only one of these signals would likely miss part of the adversarial effect.

At the same time, explanation drift should not be interpreted as a perfectly separable signature of attack. The drift and overlap histograms show that adversarial and normal samples occupy partially overlapping regions, even though adversarial samples tend to shift toward higher drift and lower overlap. This means that explanation instability is best understood as a probabilistic monitoring signal rather than as a deterministic marker of manipulation. Its value lies in augmenting output-level monitoring, not in replacing it entirely.

5.6 Learned explanation-based detection improves over rigid thresholding

The learned detector was introduced to reduce reliance on rigid thresholding, and the results support that design choice. In the small-sample Kernel SHAP branch, the learned secondary detector consistently improved on the baseline rule-based detector in both models. This indicates that the explanation-deviation features contain useful adversarial information that is not fully captured by a fixed threshold over drift and overlap alone. A lightweight learned decision boundary can combine multiple deviation statistics more flexibly than static cutoffs and can therefore exploit interactions among drift magnitude, variance, overlap, and other explanation-based signals. The rule-based detector remains valuable as an interpretable baseline, but the learned detector provides stronger evidence that explanation-driven adversarial monitoring is better formulated as a data-driven

discrimination problem rather than purely as a hand-tuned thresholding task. In operational terms, this suggests that explanation-based monitoring layers in IDS and SOC settings should be designed to support recalibration and adaptation rather than rigid decision logic.

5.7 Larger-scale validation tempers small-sample optimism

The larger GradientExplainer branch was included to reduce dependence on the 100-sample Kernel SHAP analysis, and the detector findings show why that addition was necessary. The small-sample branch provided encouraging initial evidence, especially for the learned detector, but the larger-scale evaluation produced more conservative detector estimates. Explanation-driven detection remained informative at broader scale, yet the larger-sample performance was lower than might have been inferred from the smaller, more focused SHAP subset alone. This moderation is methodologically important. It indicates that the earlier small-sample detector evidence should be interpreted as exploratory rather than definitive. By contrast, the larger GradientExplainer branch provides a more credible estimate of detector behavior under broader test-set variation. The fact that explanation drift remained informative under this more demanding setting strengthens the study, but the drop relative to the smaller branch also cautions against overclaiming detector effectiveness. In other words, the explanation-based defense remains promising, but it should be presented as a meaningful auxiliary layer rather than as a complete solution.

5.8 Operational implications for IDS and SOC environments

The findings have direct implications for practical IDS and SOC deployment. First, they show that recurrent IDS models with strong conventional performance can still be substantially vulnerable under adversarial pressure. Second, they show that explanation-based monitoring can provide an additional signal when predictive output alone is insufficient to identify manipulated inputs. Third, they suggest that model selection in security pipelines should not be guided solely by clean predictive accuracy, but should also account for adversarial degradation patterns and the stability of explanations under attack.

From an operational perspective, the LSTM may be preferable when the priority is stronger classifier robustness under the principal FGSM condition, while the GRU may be preferable when the priority is larger-scale explanation-drift separability and stronger false-positive control in the detector layer. This means that architecture selection depends on the intended balance between classifier resilience and explanation-driven monitoring. In a SOC workflow, explanation-based detection is therefore best viewed as a supportive monitoring layer that can help prioritize suspicious inputs for escalation, rather than as a standalone substitute for robust base-model performance.

5.9 Limitations and scope of the proposed method

The present study should be interpreted within the limits of its experimental scope. First, the analysis was conducted on a single dataset, which may restrict the generalizability of the observed robustness and explanation-drift patterns to other traffic distributions, attack taxonomies, and class balances. Second, although adversarial evaluation was broadened beyond FGSM through representative PGD settings, the attack space remains incomplete and does not yet include transfer-based attacks, stronger optimization-based variants beyond the selected settings, or fully adaptive explanation-aware adversaries.

Third, this study does not evaluate robustness on native packet traces or richer session-level temporal constructions. Accordingly, the findings should be interpreted as robustness results for temporally structured flow windows rather than for fully packet-level sequence models. Fourth, exact Kernel SHAP analysis remained computationally expensive and therefore had to be restricted to a small subset, even though this limitation was partly mitigated by the addition of the larger GradientExplainer branch. Finally, the explanation-based detectors remain dependent on the chosen attribution method, feature set, and calibration strategy, all of which may influence operating characteristics.

5.10 Comparison with the state of the art

Prior work has shown, separately, that recurrent models can perform strongly for intrusion detection, that explainable AI can improve interpretability and analyst trust in IDS, and that adversarial perturbations can substantially degrade IDS reliability; however, these strands remain only partially integrated in the literature. Recurrent IDS studies typically emphasize clean classification performance, XAI-based IDS studies focus mainly on transparency, and adversarial IDS studies often stop at predictive degradation under attack, while protocol-specific work in industrial IoT and Controller Area Networks further shows that IDS design is strongly shaped by communication context. As summarized in Table 6, earlier work has therefore made important contributions in recurrent intrusion detection, explainable IDS design, adversarial robustness analysis, and protocol-aware IDS development, but usually in isolation. By contrast, the present study integrates these elements within a common temporal flow-window pipeline and evaluates LSTM and GRU under matched clean, adversarial, and

Table 6 Comparison of representative IDS studies with the present work

Study	Focus	Robustness	XAI / defense	Main gap relative to this study
[35]	Explainable IDS	No	XAI for interpretation	Uses explainability for transparency, not adversarial monitoring
[36]	XAI for IDS survey	Mixed	Survey of XAI methods	Does not implement explanation-drift detection
[37]	Adversarial IDS robustness	Yes, FGSM/BIM/PGD	No XAI-based detector	Studies stronger attacks, but not SHAP-based monitoring
[38]	Adversarial NIDS evaluation	Yes, multi-attack	No explanation detector	Broad robustness study, but no recurrent LSTM vs GRU explanation analysis
[39]	Deep learning IDS for Controller Area Networks	Not the main focus	No explanation-based detector	Useful for protocol-specific IDS context, but does not evaluate SHAP drift or explanation-driven adversarial detection
[40]	Deep learning IDS for industrial IoT sensing systems	No	No explanation layer	Demonstrates DL-based IDS in industrial IoT, but does not study adversarial robustness or explanation-instability monitoring
Protocol-specific explainable IDS studies	Explainable IDS in IoT / protocol domains	Limited	Explainable ML / SHAP style interpretation	Focus on interpretability, not adversarial explanation drift
Present study	Recurrent IDS, adversarial robustness, explanation-based monitoring	Yes, FGSM plus PGD	SHAP drift, top-k overlap, rule-based and learned detector	Integrates recurrent comparison, robustness, explanation instability, and larger-scale validation

explanation-driven detection conditions, thereby extending prior work through broader FGSM and PGD stress testing, SHAP-based explanation-instability analysis, and both rule-based and learned explanation-driven monitoring.

Existing recurrent IDS studies show that LSTM and GRU can achieve strong clean intrusion-detection performance, but they rarely test whether that performance remains reliable under adversarial pressure. Likewise, explainable IDS research demonstrates that SHAP and related methods can improve interpretability, yet typically treats explanation as a post hoc aid rather than as a measurable signal of manipulation, while adversarial IDS studies usually stop at predictive degradation without examining whether explanations themselves can reveal attack-induced instability. In addition, protocol-specific IDS research in domains such as industrial IoT and Controller Area Networks highlights that IDS design and evaluation are strongly shaped by communication context, but these studies generally focus on detection capability rather than on adversarial robustness and explanation-instability monitoring. Against this background, the present study extends prior work by providing a controlled LSTM versus GRU comparison under a unified temporal flow-window pipeline, broadening robustness evaluation through a main FGSM branch and a wider FGSM plus PGD sweep, and introducing explanation-instability monitoring through SHAP drift, top-k feature overlap, a rule-based detector, and a lightweight learned detector with larger-sample GradientExplainer validation. The comparison should therefore be interpreted as a methodological positioning exercise rather than a claim of absolute superiority, since the reviewed studies differ in datasets, protocols, feature representations, and evaluation settings.

5.11 Discussion summary

Overall, this study shows that adversarial perturbation affects both predictive performance and explanation structure in recurrent intrusion detection systems, and that this instability can be exploited as a meaningful signal for adversarial monitoring. The broader FGSM and PGD sweep demonstrates that model fragility is greater than FGSM alone would suggest, while the SHAP-based analyses show that adversarial manipulation induces measurable attribution drift and top-k feature instability. At the same time, explanation drift is not perfectly separable and should therefore be treated as an auxiliary monitoring signal rather than a complete defense. The results further show that a learned detector improves on rigid thresholding and that larger-scale validation provides a more realistic estimate of detector performance than the earlier small-sample SHAP branch alone. Finally, the comparative analysis indicates that LSTM and

GRU exhibit different trade-offs rather than a single universal advantage: the LSTM retained stronger classifier robustness under the principal FGSM setting, whereas the GRU showed a modest advantage in larger-scale explanation-based detector separability. Taken together, these findings support the central argument of the study that explanation drift is not merely a by-product of adversarial failure, but a practically useful indicator of model instability under attack when interpreted within a broader multi-attack robustness framework.

6 Conclusion

This study examined adversarial robustness in recurrent intrusion detection systems by comparing LSTM and GRU models in a multiclass setting and by evaluating SHAP-based explanation instability as a secondary monitoring signal. Using true sliding flow-window sequences, both models achieved strong clean performance but degraded substantially under adversarial perturbation, showing that benign-condition accuracy alone is not a sufficient indicator of robustness. The broader robustness sweep further showed that vulnerability increased with perturbation strength and that iterative PGD exposed stronger fragility than FGSM alone. At class level, degradation was highly uneven, with normal traffic remaining comparatively more stable while several minority attack categories deteriorated sharply under attack.

The study also showed that adversarial perturbation affected not only predicted labels but also the explanatory structure underlying those predictions. SHAP explanation drift therefore provided a useful auxiliary signal for adversarial monitoring, although it was not perfectly separable and should not be interpreted as a standalone defense. A learned secondary detector outperformed the rigid rule-based baseline, indicating that explanation-driven monitoring is more effective when formulated as a lightweight learned discrimination task rather than as fixed thresholding alone. The comparative results further revealed no single universal winner between the two recurrent architectures: the LSTM retained stronger adversarial predictive performance under the main FGSM setting, whereas the GRU showed a modest advantage in larger-scale explanation-based detector separability and false-positive control.

Relative to the state of the art, the contribution of this study lies not in claiming universal superiority over all IDS models, but in integrating several strands that are often treated separately in prior work. Existing recurrent IDS studies typically emphasize clean classification performance, XAI-based IDS studies focus mainly on interpretability, and adversarial IDS studies often stop at predictive degradation under attack. By contrast, the present study combines a controlled LSTM versus GRU comparison, broader FGSM and PGD robustness evaluation, SHAP-based explanation-instability analysis, and both rule-based and learned explanation-driven detection within a unified temporal flow-window framework. In this sense, the study extends prior work by positioning explanation drift not only as an interpretive aid, but also as a practical monitoring signal for adversarial instability.

Overall, the study supports three main conclusions. First, recurrent IDS models can achieve strong clean multiclass performance while still remaining highly vulnerable under adversarial perturbation. Second, broader attack-family evaluation is necessary because stronger iterative attacks reveal more severe fragility than FGSM alone. Third, explanation drift is a practically useful indicator of model instability under attack, especially when incorporated into a learned secondary detector. The study remains limited by its single-dataset setting and incomplete attack space, and future work should extend the framework across additional datasets, richer temporal representations, protocol-specific environments, and adaptive explanation-aware adversarial settings.

Acknowledgements This work was supported in part by the South African National Research Foundation under Grants numbers 141951, CHN231229201835, 137951, AJCR230704126719.

Author contributions EMM, SY, and ZW proposed the adopted framework and obtained the datasets for the research. SY's and ZW's rich experience was instrumental in improving our work. EMM did the literature survey of the paper and wrote the main manuscript. All authors contributed to the editing and proofreading. All authors read and approved the final manuscript.

Funding Open access funding provided by University of Johannesburg.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interest The authors declare no competing interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Buczak, A.L., Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun. Surv. Tutor. **18**(2), 1153–1176 (2016). https://doi.org/10.1109/COMST.2015.2494502
2. Sommer R, Paxson V (2010) Outside the closed world: On using machine learning for network intrusion detection. In: *2010 IEEE Symposium on Security and Privacy*, pp 305–316. IEEE. https://doi.org/10.1109/SP.2010.25.
3. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv*. https://doi.org/10.48550/arXiv.1406.1078.
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
5. Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: *3rd International Conference on Learning Representations (ICLR 2015)*. https://doi.org/10.48550/arXiv.1412.6572.
6. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: *2nd International Conference on Learning Representations (ICLR 2014)*. https://doi.org/10.48550/arXiv.1312.6199.
7. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Adv. Neural. Inf. Process. Syst. **30**, 4765–4774 (2017)
8. García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: techniques, systems and challenges. Comput. Secur. **28**(1–2), 18–28 (2009). https://doi.org/10.1016/j.cose.2008.08.003
9. He, K., Kim, D.D., Asghar, M.R.: Adversarial machine learning for network intrusion detection systems: a comprehensive survey. IEEE Commun. Surv. Tutor. **25**(1), 538–566 (2023). https://doi.org/10.1109/COMST.2022.3233793
10. Mohale, V.Z., Obagbuwa, I.C.: A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Front. Artif. Intell. **8**, 1526221 (2025). https://doi.org/10.3389/frai.2025.1526221
11. Althubiti SA, Nick W, Mason J, Yuan X, Esterline A (2018) Applying long short-term memory recurrent neural network for intrusion detection. In: *2018 IEEE SoutheastCon*, pp 1–5. IEEE. https://doi.org/10.1109/SECON.2018.8478898.
12. Han, J., Pak, W.: Hierarchical LSTM-based network intrusion detection system using hybrid classification. Appl. Sci. **13**(5), 3089 (2023). https://doi.org/10.3390/app13053089
13. Kasongo, S.M., Sun, Y.: A deep learning technique for intrusion detection system using a recurrent neural networks based framework. Comput. Commun. **199**, 113–125 (2023). https://doi.org/10.1016/j.comcom.2022.12.010
14. Yin, C., Zhu, Y., Fei, J., He, X.: A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access **5**, 21954–21961 (2017). https://doi.org/10.1109/ACCESS.2017.2762418
15. Shoab, M., Alsbatin, L.: GRU enabled intrusion detection system for IoT environment with swarm optimization and Gaussian random forest classification. Comput. Mater. Contin. **81**(1), 625–642 (2024). https://doi.org/10.32604/cmc.2024.053721
16. Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., Colajanni, M.: Modeling realistic adversarial attacks against network intrusion detection systems. Digit Threat Res Pract **3**(3), 24 (2022). https://doi.org/10.1145/3469659
17. Jmila, H., Ibn Khedher, M.: Adversarial machine learning for network intrusion detection: A comparative study. Comput. Netw. **214**, 109073 (2022). https://doi.org/10.1016/j.comnet.2022.109073
18. Catillo M, Del Vecchio A, Pecchia A, Villano U (2023) A case study with CICIDS2017 on the robustness of machine learning against adversarial attacks in intrusion detection. In: *Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES '23)*. https://doi.org/10.1145/3600160.3605031.
19. Pantelakis V, Bountakas P, Farao A, Xenakis C (2023) Adversarial machine learning attacks on multiclass classification of IoT network traffic. In: *Proceedings of the 18th International Conference on Availability, Reliability and Security (ARES '23)*. https://doi.org/10.1145/3600160.3605086.
20. 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. **12**(13), 6451 (2022). https://doi.org/10.3390/app12136451
21. Wali, S., Farrukh, Y.A., Khan, I.: Explainable AI and random forest based reliable intrusion detection system. Comput. Secur. **157**, 104542 (2025). https://doi.org/10.1016/j.cose.2025.104542
22. Khan, N., Ahmad, K., Al Tamimi, A., Alani, M.M., Bermak, A., Khalil, I.: Explainable AI-based intrusion detection systems for Industry 5.0 and adversarial XAI: a systematic review. Information **16**(12), 1036 (2025). https://doi.org/10.3390/info16121036
23. Sharma, A., Rani, S., Shabaz, M.: A comprehensive review of explainable AI in cybersecurity: decoding the black box. ICT Express **11**(6), 1200–1219 (2025). https://doi.org/10.1016/j.icte.2025.10.004
24. Gaspar, D., Silva, P., Silva, C.: Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. IEEE Access **12**, 30164–30175 (2024). https://doi.org/10.1109/ACCESS.2024.3368377
25. Hermosilla, P., Berríos, S., Allende-Cid, H.: Explainable AI for forensic analysis: a comparative study of SHAP and LIME in intrusion detection models. Appl. Sci. **15**(13), 7329 (2025). https://doi.org/10.3390/app15137329
26. Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., Zomaya, A.Y.: An explainable deep learning-enabled intrusion detection framework in IoT networks. Inf. Sci. **639**, 119000 (2023). https://doi.org/10.1016/j.ins.2023.119000
27. Al, S., Sağıroğlu, Ş: Explainable artificial intelligence models in intrusion detection systems. Eng. Appl. Artif. Intell. **144**, 110145 (2025). https://doi.org/10.1016/j.engappai.2025.110145
28. Fidel G, Bitton R, Shabtai A (2020) When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. In: *2020 International Joint Conference on Neural Networks (IJCNN)*, pp 1–8. IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207637.
29. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020) Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, pp 180–186. https://doi.org/10.1145/3375627.3375830.
30. Vadillo, J., Santana, R., Lozano, J.A.: Adversarial attacks in explainable machine learning: a survey of threats against models and humans. WIREs Data Min. Knowl. Discov. **15**(1), e1567 (2025). https://doi.org/10.1002/widm.1567
31. Pawlicki, M., Pawlicka, A., Kozik, R., Choraś, M.: The survey on the dual nature of xAI challenges in intrusion detection and their potential for AI innovation. Artif. Intell. Rev. **57**, 330 (2024). https://doi.org/10.1007/s10462-024-10972-3

32. Yang, W., Wei, Y., Wei, H., et al.: Survey on explainable AI: from approaches, limitations and applications aspects. Hum. Cent. Intell. Syst. **3**(3), 161–188 (2023). https://doi.org/10.1007/s44230-023-00038-y
33. Booij, T.M., Chiscop, I., Meeuwissen, E., Moustafa, N., Den Hartog, F.T.H.: ToN_IoT: the role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets. IEEE Internet Things J. **9**(1), 485–496 (2022). https://doi.org/10.1109/JIOT.2021.3085194
34. Moustafa, N.: A new distributed architecture for evaluating AI-based security systems at the edge: network TON_IoT datasets. Sustain. Cities Soc. **72**, 102994 (2021). https://doi.org/10.1016/j.scs.2021.102994
35. Patil, S., Varadarajan, V., Mazhar, S.M., Sahibzada, A.: Explainable artificial intelligence for intrusion detection system. Electronics **11**(19), 3079 (2022). https://doi.org/10.3390/electronics11193079
36. Neupane, S., HaddadPajouh, H., Wang, M., & Choo, K.-K. R. (2022). Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities. *IEEE Access*. Advance online publication. https://arxiv.org/abs/2207.06236.
37. Debicha, I., Debatty, T., Dricot, J.-M., & Mees, W. (2021). Adversarial training for deep learning-based intrusion detection systems. In *Proceedings of the Sixteenth International Conference on Systems (ICONS 2021)*. https://doi.org/10.48550/arXiv.2104.09852
38. Sharma, S., Chen, Z.: A systematic study of adversarial attacks against network intrusion detection systems. Electronics **13**(24), 5030 (2024). https://doi.org/10.3390/electronics13245030
39. ALTAN, G., HABEŞOĞLU, E., ABASIKELEŞ TURGUT, I. (2025). Deep Learning for Intrusion Detection on Controller Area Networks. In: ALTAN, G., ABASIKELEŞ TURGUT, I. (eds) Deep Learning in Ad-Hoc Wireless Networks. Studies in Big Data, vol 172. Springer, Cham. https://doi.org/10.1007/978-3-031-86075-1_5
40. Altan, G.: SecureDeepNet-IoT: a deep learning application for invasion detection in industrial Internet of Things sensing systems. Trans. Emerg. Telecommun. Technol. **32**, e4228 (2021). https://doi.org/10.1002/ett.4228

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.