



AN ADVERSARIALLY ROBUST RSHO DBiGRNN FRAMEWORK FOR DDOS INTRUSION DETECTION

^{1*}A. Immaculate, ² Dr. M Suganya

^{1*}Assistant Professor, Department of Computer Science, Kongunadu Arts and Science College, Coimbatore, India.

² Associate Professor & Head of a Department, Department of Information Technology, School of Computer studies, RVS College of Arts and Science, Sulur, India.

Abstract

DDoS-Distributed Denial-of-Service intruders constitute an important threat to network obtainability by exhausting network and computational sources via harmful traffic. Though machine learning(ml) and deep learning based Network Intrusion Detection Systems (NIDSs) can determine complex traffic design, high identification on clean data does not essentially imply robustness towards adversarial manipulation. This research implements an adversarially robust extension of the Refined Selfish Herd Optimization oriented Deep Bidirectional Gated Recurrent Neural Network (RSHO-DBiGRNN) prototype for DDoS intrusion detection. The prototype stays the core feature illustration, Hybrid Vulture–Dragonfly Optimization Algorithm (HVD-OA) feature selection, and RSHO-DBiGRNN elements while incorporating leakage-controlled, training-fitted preprocessing, multi-attack adversarial training, and robustness-aware configuration optimization. Network traffic dataset feature extracted using parallel Autoencoder (AE), Linear Discriminant Analysis (LDA), and Adaptive Entropy-Weighted Non-Negative Matrix Factorization (NMF) transition, succeeded by HVD-OA oriented feature selection. Adversarial samples are created using the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and DeepFool in the chosen feature space and incorporated into model training alongside clean samples. Moreover, the RSHO optimization objective is reformulated to select recurrent dropout configurations utilizing a connected clean and multi-attack adversarial validation criterion. Execution is evaluated under clean and adversarial states making use of accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), and receiver operating characteristic area under the curve (ROC-AUC). Threat specific execution degradation is also measured relative to clean test performance. The research looks into the contribution of combining multi-attack adversarial training with robustness-aware configuration chosen to DDoS finding resilience subjected to evaluated feature-space perturbations. Experimental output displays that the implemented model achieves 99.24% accuracy and 99.37% F1-score under clean test conditions. Under FGSM, PGD, and DeepFool attacks, it achieves F1-scores of 97.03%, 96.15%, and 96.97%, correspondingly.

Keywords: DDoS detection; network intrusion detection; adversarial examples; adversarial training; FGSM; PGD; DeepFool; DBiGRNN; RSHO; feature selection; robustness.

1. Introduction



Distributed Denial-of-Service (DDoS) attacks continue to be a crucial threat to networked systems because they can cause excessive communication and computational resources by creating large amounts of harmful traffic. Network intrusion detection systems (NIDSs) are thus broadly utilized to find unusual and harmful network events. With the growing volume and complication of network load, Deep learning (DL) based NIDSs has acquired focus due to their ability to understand distinct illustrations straight from high-dimensional load characteristics. Moreover, high identification performance on clean traffic does not essentially imply robustness in contrast to adversarial manipulation. An adversary may purposely change the input illustration when protecting the underlying traffic class, causing a trained detector to make an incorrect prediction. Accordingly, assessing an intrusion detection framework only on clean traffic gives an unfinished evaluation of its reliability in adversarial conditions. This generates a requirement to assess DDoS identification patterns not only based on their clean data accomplishment but also based on their ability to resist adversarial change.

Study on adversarial Machine Learning (ML) has thus expanded to cybersecurity and NIDS applications. Mohammadian et al. researched adversarial examples produced in network traffic applying CIC-IDS2017 and CIC-DDoS2019 and showed that the choice and connection of network characteristics influence the production of effective adversarial instances. Their outcomes imply that feature classes may behave differently through the attack conditions, highlighting the significance in the view of characteristics of network traffic illustration when assessing adversarial vulnerability [5]. This examination is specifically related to NIDS since network traffic characteristics are integrated and alterations to the input depiction can impact the action of the classifier. Moreover, researchers have reviewed the robustness of Deep Learning (DL) based intrusion identification structure under adversarial attacks. Sauka et al. analysed adversarial robust and understandable NIDS using adversarial training with FGSM, PGD and DeepFool and showed that opposed changes can impact the execution of Deep Learning (DL) oriented intrusion detection framework. Their review also demonstrated that adversarial training can give different levels of robustness based on the attack applied while training [9]. These outcomes highlight that assessing an IDS only under clean traffic may give an unfinished evaluation of its security action.

Though previous studies have analysed adversarial attacks, adversarial training and robust feature selection, an additional issue stays in optimization based deep intrusion detection. A framework configuration that executes well on clean verification data is not required for the configuration that gives the best action subjected to adversarial scenarios. Followingly, the optimization objective used for choosing model configurations can itself impact the robustness of the detector. A further research opportunity thus exists in making the configuration optimization of an existing DDoS detector robustness aware. In specific, the parameters chosen by a metaheuristic optimizer can be assessed not only based on clean verification but also based on their action under several adversarial scenarios.

More recent study has also regarded feature selection together with adversarial robustness. Vitorino et al. analysed reliable feature selection for cyber-attack identification by



connecting feature-selection methods with usual and adversarial training and followingly analysing the chosen feature group subjected to adversarial evasion state. Their research points out the connection within feature representation, feature selection and adversarial robustness [12]. Likewise, Holla et al. analysed adversarial threats to cloud oriented IDS and regarded adversarial training together with robust feature selection as defense mechanisms opposed adversarial manipulation [13]. These research shows the significance of jointly considering adversarial vulnerability and defensive mechanisms instead of depending entirely on clean-data identification execution.

Even though prior researchers have examined adversarial attacks, adversarial training, and robust feature selection, the choice of deep intrusion detection prototype configurations based connectedly on clean and multi-attack adversarial evaluation implementation stays a significant research path. A configuration that executes well on clean validation data may not create the best implementation subject to adversarial perturbations. Associatively, the optimization objective applied to choose prototype configurations can impact the robustness of the outcoming detector. This prompts the examination of a robustness-aware optimization technique that takes into account both clean and adversarial validation execution while choosing model configurations.

This study's path is related to the RSHO-DBiGRNN model, which links feature representation, HVD-OA oriented feature selection, deep bidirectional gated recurrent neural network learning, and RSHO oriented configuration optimization for attack detection. Constructing on the present model, this work examines an extension that integrates adversarial validation and mixed adversarial training while reformulating the RSHO objective to choose recurrent dropout configurations based on a linked clean and multi attack adversarial validation criterion. The target is to find whether this robustness-aware configuration-selection technique enhances the retention of detection execution subject to the validated FGSM, PGD, and DeepFool feature space perturbations.

In the implemented architecture, adversarial samples are produced employing the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and DeepFool in the chosen feature representation. Clean and adversarial representations are integrated into DBiGRNN training to reveal the detector to several changed states. More foremost, the RSHO optimization objective is revised so that recurrent dropout configurations are selected using both clean validation performance and multi-attack adversarial validation performance. Thus, the present study investigates whether integrating adversarial action into configuration optimization can give more robustness related to choosing the RSHO-DBiGRNN indicator. The primary beneficitation of the work involves adversarial robustness assessment, multi-attack adversarial training, Robustness aware RSHO, attack-specific robustness evaluation and Component-level analysis. This work is organized as follows. Section 2 considers related research on deep learning oriented NIDS, adversarial attacks, adversarial training and robust feature selection. Section 3 shows the threat model and problem formulation. Section 4 portrays the implemented adversarially robust RSHO-DBiGRNN model. Section 5 depicts



the adversarial learning strategy. Section 6 provides exploratory outcomes and review. Section 7 provides discussions and upcoming research paths.

2. Related Work

Study on deep learning-based network intrusion detection has progressively analysed the vulnerability of trained classifiers to adversarial manipulation. He et al. [1] demonstrated an extensive examination of adversarial machine learning for network intrusion detection systems, discussing attack strategies, defense mechanisms, and challenges connected with assessing model robustness. Their view emphasized that conventional classification execution does not alone establish resilience towards adversarial samples. Zhang et al. [2] overviewed adversarial attacks in contrast with deep learning-based NIDS and addressed defense mechanisms, emphasizing the requirement to regard the link between attack generation and model action. An existing study by Wang [3] also analysed deep learning-based intrusion detection in the existence of adversaries, providing for the improvement of study on adversarial vulnerability in network traffic classification.

Multiple research studies have examined how adversarial samples can impact intrusion detection decisions. Zhang et al. [4], in their Tiki-Taka study, analysed threats towards deep learning, relied on intrusion detection systems and researched defensive methods. Mohammadian et al. [5] investigated gradient oriented adversarial intrusions on deep learning relied on NIDS, additionally studying network intrusion datasets. Their study is related to the current research since gradient oriented perturbations can exhibit weaknesses that may not be apparent from clean-data assessment. Sharma and Chen [6] showed a systematic work of adversarial attacks contrast to network intrusion detection approaches, moreover illustrating the significance of evaluating detection models of various multiple threats conditions instead of depending completely on conventional test data.

Adversarial training has shown as a significant strategy for enhancing the resilience of intrusion detection prototypes. Abdelaty et al. [7] implemented GADoT, a GAN oriented adversarial training procedure for robust DDoS attack identification. Their work directly handles DDoS detection and examines the need of adversarially created traffic to strengthen the detector. Novaes et al. [8] also analyzes adversarial deep learning for DDoS detection and defense in software-defined networking conditions. These works give a relative base for integrating adversarial samples into the training system of a DDoS detector. On the other hand, the execution of an adversarial-training process relies on the attacks depicted while training and the states applied for following evaluation. Thus, a detector should be evaluated individually subject to clean traffic and the adversarial attack types chosen for the analysis.

Another study has examined adversarial robustness via various model structures and defensive mechanisms. Sauka et al. [9] examined adversarially robust and explainable network intrusion detection applying deep learning, in the view of FGSM, PGD, and DeepFool in their exploration. Their research is specifically related to the current study since it regards multiple attack mechanisms instead of restricting verification to a single perturbation technique. Yuan et al. [10] implemented a model to improve the adversarial



robustness of deep learning oriented intrusion detection techniques by connecting a deep learning detector with additional mechanisms for detecting adversarial instances. Though their defensive framework varies from the adversarial-training and optimization method analysed here, the work represents that robustness may need measures beyond maximizing clean-data classification accuracy.

Feature representation and feature selection are also related to the robustness of network intrusion detection. McCarthy et al. [11] examined feature vulnerability and robustness in regard with adversarial machine learning threats. Their study assists in analysing the connection within the chosen feature representation and the actions of a detector under perturbation. Vitorino et al. [12] researched dependable feature selection for adversarially robust cyber-attack detection, in regard with feature-selection method together with regular and adversarial training. This line of work is similar since the choice of data traffic features can influence both classification actions and the features available to an adversary. Holla et al. [13] moreover examined adversarial attacks to cloud intrusion detection in view of adversarial training and feature selection as defensive components. Jointly, these researches inspire assessing feature selection and adversarial defense as connected but distinct elements of a detection source.

Study on DDoS detection has also examined deep learning and optimization methods. Roopak et al. [14] analysed an intrusion detection system for DDoS attacks in IoT networks, whereas Batchu et al. [15] executed an optimization-driven deep learning structure for DDoS attack detection. These works provide conditions for linking studied traffic depictions with model optimization. In spite of this, conventional optimization for detection efficiency and optimization that explicitly accounts for adversarial robustness handles different objectives. A configuration that is efficient on clean validation data may not remain the same capability when assessed towards deliberately perturbed inputs. Based on the research taken into account here, the current study analyses an adversarially robust extension of the RSHO-DBiGRNN model. The implemented model connects FGSM, PGD, and DeepFool adversarial instances creation with adversarial training and a robustness-aware RSHO objective that views both as clean and adversarial validation capability. Distinct from a comparison depend only on clean-data accuracy, the assessment examines the inception and implemented model subject to clean and attack centered test states using accuracy, precision, recall, F1-score, Matthews correlation coefficient, ROC-AUC, and relative performance degradation. The aimed contribution is to assess whether this integration enhances the resilience of the chosen DDoS detector in contrast with the specific feature-space intrusion regarded in the experiments.

3. Threat Model and Problem Formulation

A. Threat Model

The implemented work regards a white-box adversarial setting in which the intruder has enough data of the target model to acquire the gradient knowledge needed for gradient oriented perturbation production. The intruder tries to alter the numerical load(input) illustration displayed to the DDoS classifier while protecting the true traffic label. The aim is



thus to create an adversarial illustration that can modify the classifier's forecast without altering the essential Normal or DDoS group designated to the source example. The adversarial analysis is done in feature space after HVD-OA choice; thus, the research validates illustration level evasion robustness instead of claiming packet-level realizability of the produced change.

Let $x \in R^m$ depicts the chosen feature representation of a network-flow instances, and let $y \in \{0,1\}$ depict its class label, where 0 indicates Normal traffic and 1 indicates DDoS traffic. An adversarial illustration is exhibited generally as

$$x^{adv} = x + \delta \quad (1)$$

Where δ implies the adversarial change. The perturbation is limited so that the outcome description stays within the allowed numerical domain initiated from the training data. Class labels are not altered, and omitted identifier fields are not brought into the intruder process.

B. Robust Optimization Problem

Let $d = [d_1, d_2]$ implies the two recurrent dropout parameters of the DBiGRNN. For every candidate configuration, the prototype is validated on clean and adversarial authorized data. The clean validation performance is stands for by $F1_{clean}$. The aggregated adversarial validation performance is specified as

$$F1_{adv} = \frac{F1_{FGSM} + F1_{PGD} + F1_{DeepFool}}{3} \quad (2)$$

The objective robustness-aware is further

$$F1_{robust} = 0.5F1_{clean} + 0.5F1_{adv} \quad (3)$$

The configuration of optimal recurrent dropout is selected according to

$$d^* = \arg \max_{d \in D} F1_{robust}(d) \quad (4)$$

Where, $D = \{(d_1, d_2) | 0.20 \leq d_1, d_2 \leq 0.50\}$

A present-study design choice is the equal weighting between clean and adversarial performance, instead of a total optimal weighting. Its goal is to protect the optimization from choosing a configuration exclusively based on clean-data action.

4. Implemented Adversarially Robust RSHO-DBiGRNN Prototype

A. Model Overview

The implemented model connects five principal steps: Leak-resistant preprocessing, approving feature illustration, HVD-OA variable selection, multi-intruder adversarial training, and DBiGRNN robustness-aware RSHO optimization. The overall processing sequence is shown in the figure 1.

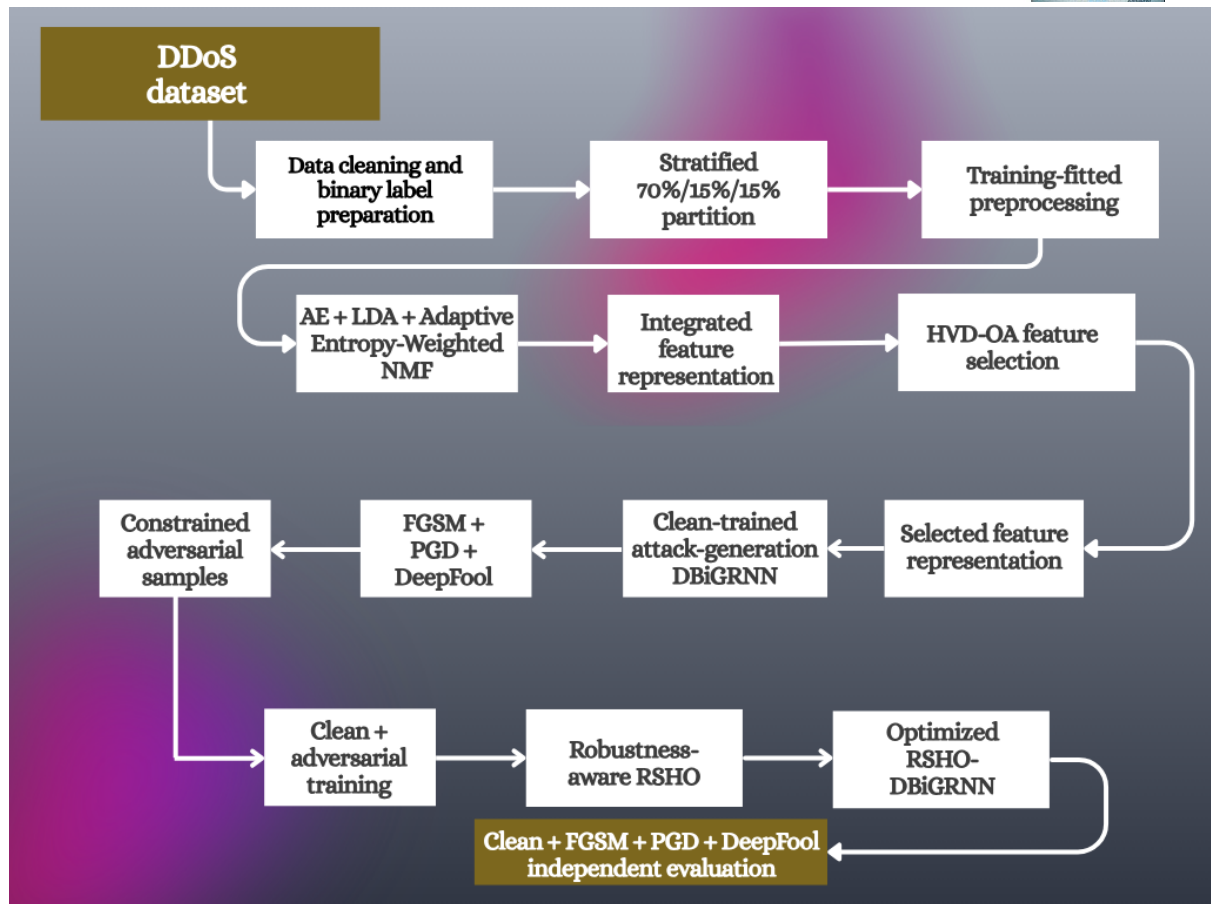


Fig 1. Overall Processing Sequence of the Proposed Work

B. Dataset and Data Preprocessing

The exploratory research employs the DDoS-Datasets benchmark acquired from Kaggle. The basic file is final_dataset.csv, containing 12,794,627 records and 85 columns. After omitting index-like, identifier, endpoint-address, timestamp and label fields, 79 numerical traffic attributes are obtainable as candidate prototype information(input). The fields Unnamed: 0, Flow ID, Src IP, Dst IP, Timestamp, and Label are omitted from the numerical prototype information(input). The objective of eliminating these variables is to stop the knowledge system from relying on record identity, endpoint identity, timestamp information or the target variable instead of the fundamental traffic features.

The target variable is illustrated as a binary classification problem:

$$y = \{0, \quad Normal \ 1, \quad DDoS\}$$

Non-finite values discovered in Flow Byts/s and Flow Pkts/s are modified to missing values before numerical imputation. Missing values are then changed using character based medians estimated exclusively from the training partition. The same training-obtained values are utilized to assess and test instances. The dataset is partitioned using stratified sampling into 70% training, 15% validation and 15% testing subsets. The training partition is used for knowledge preprocessing parameters, feature illustration, feature selection and adversarial



training. The validation division partition is retained for model configuration and robustness aware optimization, while the test division stays independent until final validation.

Eight zero-variance attributes were chosen over the preprocessing examination. Their exclusion is identified employing the training division to avoid data leakage. The determined attributes are Fwd URG Flags, Bwd URG Flags, Fwd Byts/b Avg, Fwd Pkts/b Avg, Fwd Blk Rate Avg, Bwd Byts/b Avg, Bwd Pkts/b Avg and Bwd Blk Rate Avg.

The maintained numerical variables are normalized using training fitted Min–Max normalization:

$$x'_i = \frac{x_i - x_i^{\min}}{x_i^{\max} - x_i^{\min}} \quad (5)$$

where $x_i^{\min}$ and $x_i^{\max}$ are the minimum and maximum values calculated from the training data. The output transformation is utilized to stably validate and test data. This stop data from the independent divisions from impacting the preprocessing stage.

C. Complementary Feature Representation

The normalized network traffic is changed via three complementary illustration mechanisms: AE, LDA and Adaptive Entropy-Weighted NMF. These changes are used **simultaneously to the same training-fitted normalized depiction**, instead of parallelly. Their results are followingly linked.

a. Autoencoder Representation

The Autoencoder absorbs a compact nonlinear depiction of the network traffic. For an input matrix X_{tr} , the encoder generates

$$Z_{AE} = f_{\theta_e}(X_{tr}) \quad (6)$$

The decoder rebuilds the input as

$$\hat{X}_{tr} = f_{\theta_d}(Z_{AE}) \quad (7)$$

The reconstruction objective is defined as

$$L_{AE} = ||X_{tr} - \hat{X}_{tr}||^2 \quad (8)$$

The AE latent dimension and NMF rank were configured as 32 in this work to obtain compact non linear representations from the normalized input.

b. LDA Representation

LDA delivers a differential illustration by requesting projection paths that maximize between group separation related in class variation .Since the problem has two classes, the maximum number of discriminant components is $C-1=1$, where $C=2$.

c. Adaptive Entropy-Weighted NMF

Adaptive Entropy-Weighted NMF generates a non-negative latent depiction when allocating feature weights according to their data submission. For every feature j , an entropy measure is measured from the normalized training depiction. A parallel data weight is then acquired, and the weighted non-negative matrix is factorized applying NMF. The NMF rank is configured as 32 in this research.

d. Integrated Feature Representation



The three depictions are linked as shown in equation 9.

$$Z = [Z_{AE} || Z_{LDA} || Z_{NMF}] \quad (9)$$

With 32 AE features, one LDA component, and 32 NMF components, the outcome combined depiction contains 65 (32 + 1 + 32) features. Therefore, according to the present preprocessing examination, the provisional 71-dimensional numerical depiction is converted into a 65-dimensional connected depiction before HVD-OA feature choice.

e. HVD-OA-Based Feature Selection

The connected depiction may have redundant or less useful dimensions. HVD-OA is thus used as a binary feature selection mechanism. Each candidate solution is presented as $S = [s_1, s_2, s_{65}]$, $s_j \in \{0, 1\}$, where $s_j = 1$ implies that the related feature is chosen and $s_j = 0$ implies that it is omitted. The fundamental HVD-OA procedure links the African Vulture Optimization Algorithm and Dragonfly Algorithm for binary feature selection. For every candidate feature group, classification efficiency is analysed utilizing stratified five-fold cross-validation in the training partition. Compatible with the basic methodology, **classification accuracy** is applied as the HVD-OA fitness measure:

$$F_{HVD}(s) = \frac{1}{5} \sum_{k=1}^5 Accuracy_k(s) \quad (10)$$

Nor the validation partition nor the test partition provides HVD-OA feature selection.

5. Adversarial Learning Strategy

Three complementary gradient-based intruders are taken to reveal the detector to different change based generation conduct. FGSM illustrates a single-step gradient-sign intrusion, PGD gives an iterative projected attack under a designated perturbation budget, and DeepFool seeks a comparatively small perturbation capable of passing over the classifier decision margin. Their joint use let robustness to be evaluated over a single attack-creation mechanism

A. Adversarial Threat Model

The implemented model regards an untargeted white-box avoidance setting. The adversary is to have an approach to the differentiable DBiGRNN attack-generation prototype and can apply its input gradients to create adversarial perturbations. The aim is to cause an originally accurate classified instance to be misclassified when protecting its original class label. The threat is thus assessed as an avoidance mechanism neither as a poisoning intruder or a marked class-conversion attack.

B. Attack-Generation DBiGRNN

Succeeding HVD-OA feature selection, the first DBiGRNN is trained entirely on spotless training instances. This model acts as a fixed attack-generation prototype. The attack-generation prototype uses Adam optimizer with a learning rate of 0.001, batch size of 32, binary cross-entropy loss, and a maximum of 100 training epochs. Early stopping with a patience of 10 epochs is applied to stop unnecessary training. The attack-generation prototype is not optimized via the implemented robustness-aware RSHO method. Its intention is to offer a fixed differentiable mark for making adversarial training and validation instances.



Splitting the attack-generation prototype from the RSHO candidate configurations restricts allowing the attack-generation method to rely on the final optimized configuration.

C. Fast Gradient Sign Method

FGSM creates an adversarial change from the sign of the gradient of the classification loss with regard to the input:

$$x_{FGSM} = x + \varepsilon \text{sign}(\nabla_x L(\theta, x, y)) \quad (11)$$

This research uses $\varepsilon = 0.10$. The perturbation is employed to the HVD-OA chosen feature depiction allocated to the DBiGRNN classifier, while the original class label is maintained.

D. Projected Gradient Descent

PGD uses repeated gradient oriented updates and projects the outcome depiction back into the allowed perturbation region. The iterative method can be demonstrated generally as

$$x^{t+1} = \Pi_{\beta_\varepsilon(x)} [x^t + \alpha \text{sign}(\nabla_{x^t} L(\theta, x^t, y))] \quad (12)$$

where α indicates the step size and Π depicts projection into the permitted perturbation region. The displayed configuration uses $\varepsilon = 0.10$, a step size of 0.01 with 10 iterations.

E. DeepFool

DeepFool iteratively estimates a perturbation that moves an input toward the decision boundary. Unlike FGSM and PGD, DeepFool does not use the same explicitly specified fixed perturbation in its basic formulation. The present study restricts DeepFool to a maximum of 50 iterations. The method terminates when the model forecast modifies or the iteration limit is reached. The created depiction is consequently subjected to the same domain limitations used to the chosen feature portrayal.

F. Perturbation Limitations

Adversarial perturbations are limited to the chosen numerical representation. The following restrictions are applied: Class labels are never changed(petrubed), Identifier fields and omitted fields are not changed(petrubed), Perturbations are employed only to the chosen numerical traffic depiction, Perturbed values are limited utilizing training generated feature levels and the same production principles are sustained through training and validation instances.

These limitations are specifically significant for network flow data because unlimited numerical perturbations may generate value outside the feature domain depiction by the training traffic.

G. Interpretation of Adversarial Training and Validation Instances

Adversarial instances are produced individually for training and validation partitions. The adversarial training set illustrated in the table 1.

Therefore, 50% of the training data stays clean and approximately 50% contains adversarial examples, with the adversarial section partitioned approximately equally among FGSM, PGD and DeepFool. Every adversarial element has both Normal and DDoS instances, and the original group labels are protected. The 50:50 clean-to-adversarial structure and about equal distribution within the three attack strategy, this research design selection meant to expose the model to both conventional and multiple adversarial traffic states. Adversarial



samples for the training and assessment partitions are created individually to conserve the divided boundaries. The test division is not utilized amidst adversarial training or RSHO optimization; adversarial test samples are derived only after the final RSHO-DBiGRNN configuration has been predetermined for separate assessment.

Table 1 :Adversarial Training Set

Training component	Proportion
Clean samples	50%
FGSM adversarial samples	16.67%
PGD adversarial samples	16.67%
DeepFool adversarial samples	16.67%
Total	100%

H. Process

Adversarial training reveals the detector to perturbed samples, but the prototype configuration chosen while optimization can still impact how the detector stabilizes clean and adversarial action. The implemented robustness-aware RSHO thus assesses every candidate recurrent dropout configuration using both clean and adversarial confirmation execution. The goal is to alter the fitness criterion used during its configuration selection stage.

The DBiGRNN has two recurrent dropout parameters: $d = [d1, d2]$, with the two parameters constrained to $0.20 \leq d1, d2 \leq 0.50$. The optimization apply a population of 20 candidate solutions over a maximum of 30 iterations as present-study optimization settings. The dropout interval of 0.20 - 0.50 succeeds the range disclosed in the underlying RSHO-DBiGRNN methodology.

For every candidate dropout configuration, the candidate prototype is validated using the clean validation set and the output F1-score, implied by $F1_{clean}(d)$, is used as the clean validation component of the robustness-aware objective. F1-score is chosen since it jointly displays precision and recall and thus accounts for the two false-positive and false-negative classification actions.

The candidate is also assessed by applying the FGSM, PGD and DeepFool validation sets. The aggregated adversarial validation score is

$$F1_{Adv}(d) = \frac{F1_{FGSM}(d) + F1_{PGD}(d) + F1_{DeepFool}(d)}{3} \quad (13)$$

The final fitness used by the implemented RSHO extension is

$$F_{Robust}(d) = 0.5F1_{clean}(d) + 0.5F1_{Adv}(d) \quad (14)$$

The first component depicts conventional detection action and the second illustrates multi-attack adversarial action. The optimal dropout pair is thus

$$d^* = \arg \arg \max_{d \in D} F_{Robust}(d) \quad (15)$$

The RSHO population-update mechanism stays according to the base RSHO technique. The modification initiated by this study is the replacement of a clean performance oriented configuration criterion with the implemented linked robustness aware fitness, where the search space stays constrained to the two recurrent dropout parameters.

6. Performance Evaluation

A. Metrics

The final detector is assessed using accuracy, precision, recall, F1-score, MCC and ROC-AUC. For binary classification,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

$$\text{Precision is } \text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

$$\text{Recall is } \text{Recall} = \frac{TP}{TP+FN} \quad (18)$$

$$\text{F1-Score is } F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

$$\text{MCC is calculated as } MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(FP+TP)(TP+FN)(TN+FP)(TN+FN)}} \quad (20)$$

ROC-AUC assesses the ranking capability of the forecasted chances through classification thresholds instead of depending on the fixed 0.50 classification threshold. For binary predictions, a DDoS probability threshold of 0.50 is used.

B. Adversarial Performance Degradation

To measure the consequence of each adversarial attack, the attack-specific metric is contrasted with the equivalent clean metric. For a metric M ,

$$D_a(M) = \frac{M_{Clean} - M_a}{M_{Clean}} \times 100 \quad (21)$$

where M_{Clean} is the clean-test value and M_a is the value under attack a.

The deterioration is measured individually for FGSM, PGD, and DeepFool. A lesser positive deterioration implies that a more proportion of clean-data performance is maintained under the following attack. Negative values, if acquired, must be explained as an increase relative to clean performance instead of automatically being addressed as a robustness improvement.

C. Evaluation protocol

The inception RSHO-DBiGRNN and implemented Robust RSHO-DBiGRNN are assessed subject to clean, FGSM, PGD, and DeepFool states. Accuracy, precision, recall, and F1-score are disclosed as percentages, while MCC and ROC-AUC are disclosed on a scale from 0 to 1. The contrast applies to a common held-out test set and steady assessment process. Adversarial perturbations are used to the chosen feature depiction; thus, the assessment examines feature-space robustness neither than packet-level realizability.

D. Clean-Test Performance

Table 2 gives the clean test reference for assessing adversarial execution. The inception RSHO-DBiGRNN has disclosed accuracy of 99.3% and F1-score of 99.80%, whereas the implemented Robust RSHO-DBiGRNN has an accuracy of 99.24% and F1-score of 99.37%. The implemented configuration also has a higher stated MCC (0.9843 versus 0.9744).

Table 2. Classification performances Clean data

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MC C	ROC-AUC
Baseline RSHO-DBiGRNN	99.3	99.98	99.63	99.80	0.9744	≈1.00
Proposed Robust RSHO-DBiGRNN	99.24	99.44	99.30	99.37	0.9843	0.99

E. Performance under Adversarial Attacks

Table 3 matches the prototype subject to FGSM, PGD, and DeepFool. The inception F1-scores are 89.51%, 83.94%, and 86.80%, correspondingly, contrast with 97.03%, 96.15%, and 96.97% for the implemented configuration. PGD generates the lowest F1-score for the two prototypes. The implemented configuration thus reveals higher F1-scores subject to all three attack states, indicating better retention of finding execution in the assessed feature space.

Table 3. Clean and adversarial test conditions Performance

Model	Test Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC	ROC-AUC
Baseline RSHO-DBiGRNN	Clean	99.35	99.98	99.63	99.80	0.9744	0.9991
Baseline RSHO-DBiGRNN	FGSM	88.02	94.24	85.24	89.51	0.7613	0.9412
Baseline RSHO-DBiGRNN	PGD	82.02	90.47	78.29	83.94	0.6463	0.8935
Baseline RSHO-DBiGRNN	DeepFool	84.37	87.96	85.67	86.80	0.6768	0.9168
Proposed Robust RSHO-DBiGRNN	Clean	99.24	99.44	99.30	99.37	0.9843	0.9987
Proposed Robust RSHO-DBiGRNN	FGSM	96.46	97.54	96.53	97.03	0.9265	0.9890
Proposed Robust RSHO-	PGD	95.41	96.83	95.48	96.15	0.9049	0.9811



DBiGRNN							
Proposed Robust RSHO-DBiGRNN	DeepFool	96.35	96.72	97.22	96.97	0.9239	0.9867

Table 4 discloses correlated degradation from each model's clean-test execution. For the inception of RSHO-DBiGRNN, the relevant F1-score degradation is 10.31%, 15.89%, and 13.03%, correspondingly. In contrast, the implemented Robust RSHO-DBiGRNN reveals relevant F1-score degradation of 2.35%, 3.24%, and 2.42%. These outcomes demonstrate smaller relevant F1-score losses for the implemented configuration subjected to the assessed intrude states. The degradation values correlate the absolute scores in Table-2 by measuring execution loss relevant to each prototype clean-test reference.

Table 4. Adversarial performance degradation

Model	Metric	FGSM	PGD	DeepFool
Baseline RSHO-DBiGRNN	Accuracy	11.40	17.44	15.08
Baseline RSHO-DBiGRNN	Precision	5.74	9.51	12.02
Baseline RSHO-DBiGRNN	Recall	14.44	21.42	14.01
Baseline RSHO-DBiGRNN	F1	10.31	15.89	13.03
Baseline RSHO-DBiGRNN	MCC	21.87	33.67	30.54
Baseline RSHO-DBiGRNN	ROC-AUC	5.80	10.57	8.24
Proposed Robust RSHO-DBiGRNN	Accuracy	2.80	3.86	2.91
Proposed Robust RSHO-DBiGRNN	Precision	1.91	2.62	2.74
Proposed Robust RSHO-DBiGRNN	Recall	2.79	3.85	2.09
Proposed Robust RSHO-DBiGRNN	F1	2.35	3.24	2.42
Proposed Robust RSHO-DBiGRNN	MCC	5.87	8.07	6.14
Proposed Robust RSHO-DBiGRNN	ROC – AUC	0.97	1.76	1.20

F. Effect of Robustness-Aware RSHO

Table 5 matches conventional RSHO with the implemented robustness-aware objective using validation execution. Conventional RSHO attains a clean F1-score of 99.44% and an average adversarial F1-score of 95.97%, while robustness-aware RSHO achieves 99.39% and 96.72%, respectively. The average adversarial F1-score enhances by 0.75 percentage points, and the robustness fitness improves from 0.9770 to 0.9805. The robustness-aware technique also generates higher F1-scores under FGSM, PGD, and DeepFool. The match between conventional RSHO and robustness-aware RSHO was executed using validation data. For each candidate dropout configuration, the clean validation F1-score and the F1-scores subject to FGSM, PGD, and DeepFool were validated. The mean of the three adversarial F1-scores was applied to indicate adversarial validation performance. The robustness-aware fitness was quantified as the equally weighted average of the clean validation F1-score and the mean adversarial validation F1-score, with the two scores



expressed as fractions. The chosen configuration was the one attaining the highest robustness-aware validation fitness. The individual test set was contained for final assessment and was not used for choosing the dropout parameters. These validation outcomes recommend that along with adversarial execution in the RSHO objective alters configuration option toward a more balanced clean adversarial standard.

Table 5. RSHO comparison of Conventional and robustness-aware

Optimization strategy	Clean Validation F1 (%)	FGSM F1 (%)	PGD F1 (%)	DeepFool F1 (%)	Average Adversarial F1 (%)	Robust Fitness
Conventional RSHO	99.44	96.48	95.11	96.31	95.97	0.9770
Robustness-aware RSHO	99.39	97.02	96.18	96.95	96.72	0.9805

G. Ablation Analysis

Table 6 analyzes the inputs of adversarial training and robustness-aware RSHO. The inception configuration has an average adversarial F1-score of 86.75%. Suggesting adversarial training improves the provided average to 96.13%, whereas the complete configuration integrating robustness-aware RSHO attains 96.72%. The match recommends that adversarial training generates the larger improvement, along with robustness-aware configuration choice contributing a moreover increase of 0.59 percentage points in average adversarial F1-score.

Table 6. Component level ablation analysis

Configuration	Average Adversarial F1 (%)	Robustness Summary	Clean F1 (%)	FGSM F1 (%)	PGD F1 (%)	DeepFool F1 (%)	MCC
Baseline RSHO-DBiGRNN	86.75	0.9328	99.80	89.51	83.94	86.80	0.9744
RSHO-DBiGRNN + Adversarial Training	96.13	0.9778	99.42	96.60	95.32	96.48	0.9855
Proposed Robust RSHO-DBiGRNN	96.72	0.9804	99.37	97.03	96.15	96.97	0.9843

Subject to adversarial states, its F1-scores were 89.51% for FGSM, 83.94% for PGD, and 86.80% for DeepFool. The outcome average adversarial F1-score was 86.75%, with a robustness outline of 0.9328. These values recommend a substantial variation between clean



and adversarial classification execution in the shortage of adversarial training. Parallely, RSHO-DBiGRNN derived from adversarial training, F1-scores enhanced to 96.60% subject to FGSM, 95.32% under PGD, and 96.48% under DeepFool, obtained an average adversarial F1-score of 96.13%. Its clean F1-score was 99.42%, and its robustness outline was 0.9778. Matched with incept RSHO-DBiGRNN, the average adversarial F1-score improved by 9.38 percentage points, whereas the robustness outline increased by 0.0450. This differentiation recommends that adversarial training is connected with a substantial enhancement in execution subject to the assessed perturbations. Adversarial training with robustness-aware RSHO acquired F1-scores of 97.03% under FGSM, 96.15% under PGD, and 96.97% under DeepFool, generating an average adversarial F1-score of 96.72%. Its clean F1-score was 99.37%, and its robustness outline was 0.9804. Relevant to adversarial RSHO-DBiGRNN, the average adversarial F1-score improved by 0.59 percentage points, where the robustness concise increased by 0.0026. The clean F1-score decreased by 0.05 percentage points, indicating a small variation in clean validation or test performance along with enhancement subject to all three attack states. On the whole, the ablation examination puts forward that blending adversarial training with robustness-aware RSHO may enhance the retention of detection execution constarined the valuation of FGSM, PGD, and DeepFool intrudes. The purpose of the whole model depends not only in integrating adversarial instances while training as well as in creating recurrent dropout choice sensitive to adversarial evaluation performance.

H. Comparative Analysis with Existing Adversarial Intrusion Detection Methods

Table 7 places the implemented prototype along with existing disclosed intrusion detection techniques assessed on NSL-KDD and CICIDS2018-related datasets. The cited studies state different responses to adversarial attacks and analyse defences in addition to adversarial-example training and robust feature selection. Such comparisons give context for this research, but absolute execution values must not be used to rank techniques directly because datasets, label definitions, feature depiction, and attack protocols differ. The implemented model approach distinction is the integration of mixed adversarial training with robustness-aware RSHO, which chooses recurrent dropout configurations using the two clean and multi-attack adversarial evaluation performance.

Table 7. Positioning of the Proposed Framework Against Existing Adversarial NIDS Studies

Study / Model	Dataset	Training / Defence	Evaluation condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Yuan et al. Baseline	NSL-KDD	Regular training	Clean	92.1	88.0	99.3	93.3
			FGSM	75.7	87.2	71.8	78.7
			BIM	67.3	84.0	59.2	69.4
			DeepFool	32.3	44.8	41.0	42.8



[10]		Adversarial -example defence	Clean	91.8	87.5	99.5	93.1
			FGSM	90.2	89.8	95.6	92.6
			BIM	86.2	85.9	89.2	87.5
			DeepFool	72.2	69.8	98.4	81.7
	CICIDS2018	Regular training	Clean	98.5	100.0	97.4	98.6
			FGSM	32.0	25.3	40.2	31.1
			BIM	22.5	24.7	28.5	26.5
			DeepFool	18.5	20.3	30.6	24.4
		Adversarial -example defence	Clean	98.1	100.0	96.7	98.3
			FGSM	95.0	97.5	94.4	95.9
			BIM	88.5	87.6	91.6	89.5
			DeepFool	71.5	54.0	92.9	68.3
Holla et al. Defended IDS [13]	CSE-CIC- IDS2018	Regular training	Clean	85.0	84.0	82.0	83.0
			FGSM	60.0	61.0	55.0	58.0
			PGD	55.0	54.0	50.0	52.0
			DeepFool	50.0	49.0	47.0	48.0
		Adversarial training + SHAP- based robust feature selection	Gradient- based attacks (reported aggregate)	88.0	87.0	85.0	86.0
Robust RSHO- DBiGRNN	DDoS	Robustness -aware RSHO + mixed adversarial training	Clean	99.24	99.44	99.30	99.37
			FGSM	96.46	97.54	96.53	97.03
			PGD	95.41	96.83	95.48	96.15
			DeepFool	96.35	96.72	97.22	96.97

7. Discussion

The output implies that combining multi-attack adversarial training with robustness-aware RSHO optimization may enhance the retention of DDoS detection execution subject to the evaluated feature space perturbations. The implemented framework reveals lower relative



F1-score degradation instead of inception under FGSM, PGD, and DeepFool. Through the assessed intruders, PGD generates the lowest F1-score for both configurations, highlighting the significance of assessing intrusion detection prototype in contrast to multiple intrusion mechanisms rather than depending on clean-test performance alone.

The ablation outcome recommends that adversarial training gives the larger enhancement in average adversarial F1-score, whereas robustness-aware RSHO supplies a further enhancement by choosing recurrent dropout configurations using both clean and multi-attack adversarial validation performance. This distinction depicts the principal organized assistance of the implemented extension. In contrast with existing studies on adversarial NIDS, adversarial training, and robust feature selection [10,13], this research analyzes configuration-level robustness optimization within an previous hybrid recurrent detection prototype. The assistance is not the initiation of established threats or adversarial training specifically, but their integration with an RSHO objective that regards clean and multi-attack validation execution. This research has constraints. The threats operate on chosen numerical features, so the output does not establish packet-level realizability. The fixed clean-trained attack-generation prototype also does not establish robustness in contrast with fully adaptive threats. Moreover, assessment is limited to binary Normal-versus-DDoS classification and the three designated gradient-based threats. Additionally work should analyze additional datasets, attack strategies, adaptive threat models, and repeated-run validations to examine the generalizability and stability of the implemented model.



References:

1. He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1), 538-566.
2. Zhang, C., Costa-Perez, X., & Patras, P. (2022). Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. *Ieee/Acm Transactions On Networking*, 30(3), 1294-1311.
3. Wang, Z. (2018). Deep learning-based intrusion detection with adversaries. *IEEE Access*, 6, 38367-38384.
4. Zhang, C., Costa-Pérez, X., & Patras, P. (2020, November). Tiki-taka: Attacking and defending deep learning-based intrusion detection systems. In *Proceedings of the 2020 ACM SIGSAC conference on cloud computing security workshop* (pp. 27-39).
5. Mohammadian, H., Ghorbani, A. A., & Lashkari, A. H. (2023). A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems. *Applied Soft Computing*, 137, 110173.
6. Sharma, S., & Chen, Z. (2024). A systematic study of adversarial attacks against network intrusion detection systems. *Electronics*, 13(24), 5030.
7. Abdelaty, M., Scott-Hayward, S., Doriguzzi-Corin, R., & Siracusa, D. (2021, October). Gadot: Gan-based adversarial training for robust ddos attack detection. In *2021 IEEE Conference on Communications and Network Security (CNS)* (pp. 119-127). IEEE.
8. Novaes, M. P., Carvalho, L. F., Lloret, J., & Proença Jr, M. L. (2021). Adversarial Deep Learning approach detection and defense against DDoS attacks in SDN environments. *Future Generation Computer Systems*, 125, 156-167.
9. Sauka, K., Shin, G. Y., Kim, D. W., & Han, M. M. (2022). Adversarial robust and explainable network intrusion detection systems based on deep learning. *Applied Sciences*, 12(13), 6451.
10. Yuan, X., Han, S., Huang, W., Ye, H., Kong, X., & Zhang, F. (2024). A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection system. *Computers & Security*, 137, 103644.
11. McCarthy, A., Andriotis, P., Ghadafi, E., & Legg, P. (2021, June). Feature vulnerability and robustness assessment against adversarial machine learning attacks. In *2021 international conference on cyber situational awareness, data analytics and assessment (CyberSA)* (pp. 1-8). IEEE.
12. Vitorino, J., Silva, M., Maia, E., & Praça, I. (2025). Reliable feature selection for adversarially robust cyber-attack detection. *Annals of Telecommunications*, 80(3), 341-355..



13. Holla, H., Polepalli, S. R., & Sasikumar, A. A. (2025). Adversarial threats to cloud IDS: robust defense with adversarial training and feature selection. *IEEE Access*, 13, 84992-85003.
14. Roopak, M., Tian, G. Y., & Chambers, J. (2020, January). An intrusion detection system against DDoS attacks in IoT networks. In *2020 10th annual computing and communication workshop and conference (CCWC)* (pp. 0562-0567). IEEE.
15. Batchu, R. K., Bikku, T., Thota, S., Seetha, H., & Ayoade, A. A. (2024). A novel optimization-driven deep learning framework for the detection of DDoS attacks. *Scientific Reports*, 14(1), 28024.