Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

基于深度学习的NIDS对抗样本检测与防御技术

硕士研究生　袁梦佳

2025 年 12 月 14 日

问题回溯

- 相关内容
  - 2024.11.17 郑俊怡 《文本分类硬标签黑盒模型的对抗样本生成方法研究》
  - 2023.10.22 邵思源 《面向NIDS的流量对抗样本检测》
  - 2023.05.29 程瑶 《单词级文本对抗攻击》
  - 2021.07.26 杨若晗 《特定安全攻防场景中的对抗样本生成方法》
  - 2021.08.31 王琛 《特定安全领域中的对抗样本防御方法》

内容提要

- 预期收获
- 内涵解析与研究目标
- 研究背景与研究意义
- 研究历史与现状
- 知识基础
- 算法原理
  - GDA-FS
  - NIDS-DA
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 熟悉网络入侵检测系统的基本概念
  - 了解对抗攻击与防御技术
  - 了解对抗样本检测技术的发展历史与研究方向
  - 理解网络流量对抗样本检测技术的原理与思路

内涵解析与研究目标

- 研究目的
  - 以网络入侵检测系统为对象，研究对抗样本检测与防御方法
  - 结合深度学习、编码器、特征压缩、对抗训练等技术
  - 实现对抗样本检测准确率的提升，以及入侵检测模型鲁棒性的提升

- 内涵解析
  - 入侵检测系统：通过实时分析网络流量或系统活动来识别和预警恶意攻击行为
  - 基于深度学习的网络入侵检测系统（Network Intrusion Detection System，NIDS）
    - 以深度学习为基础的网络流量入侵检测系统，用于区分良性/恶意流量
  - 对抗样本是一种经过精心设计、微小扰动的输入数据，能诱导模型做出错误预测

研究背景与研究意义

- 研究背景
  - 面对日益复杂的网络攻击，传统NIDS逐渐暴露出检测能力的局限性
    - 对预定义规则的依赖和对未知攻击模式的泛化能力不足
    - 在面对零日攻击、对抗攻击等高级威胁时，存在显著的检测盲区
  - 基于深度学习的NIDS 性能出色，但其模型本身存在对抗脆弱性的固有缺陷
    - 在高维特征空间中构建的决策边界复杂，但在局部往往是高度线性的，使得决策行为在微小扰动下变得不稳定
    - 模型决策依赖于其数据所呈现的统计分布
    - 攻击者可借助模型梯度，施加微小且合规的扰动欺骗模型，诱导模型做出错误判决
- 研究意义
  - 加固安全防线，提升 NIDS 的鲁棒性与可靠性
  - 提高深度学习模型在复杂多变网络环境中的可靠性和信任度

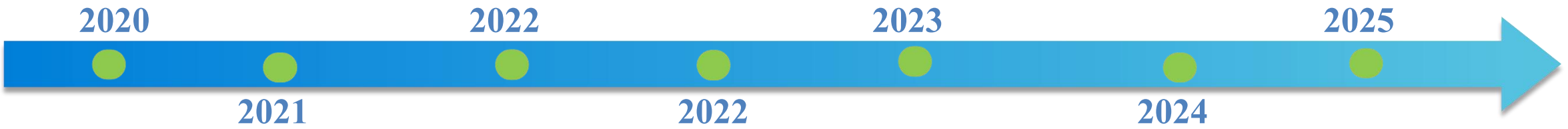
研究历史和现状

Peng等人提出了一种基于GAN的对抗样本检测技术，GAN的生成器网络学习干净数据样本的潜在空间分布，通过监测重构误差和鉴别器对可疑数据样本的输出之间的不匹配来检测对抗样本。

Khettaf 等人提出一种双层检测机制来防御对抗样本攻击：通过统计检验快速识别可能包含对抗扰动的可疑网络流量并标记；再利用一个基于标注数据训练的辅助分类器对这些可疑流量进行分析，实现对对抗样本的精准提取。

Debicha等人提出了一种基于迁移学习的多对抗检测方法（TAD），通过设计串行/并行架构的多个对抗检测器并结合决策融合规则，在检测未知攻击时优于单检测器，尤其在并行IDS中显著提升了对抗样本检准确率。

Kumar等人提出了NIDS-DA，一种基于深度自编码器（DAE）的新型对抗样本检测方法，专门针对网络入侵检测系统（NIDS）中仅修改非功能特征以保留恶意功能的对抗攻击。相比传统方法，具有更高的准确性和实用性。

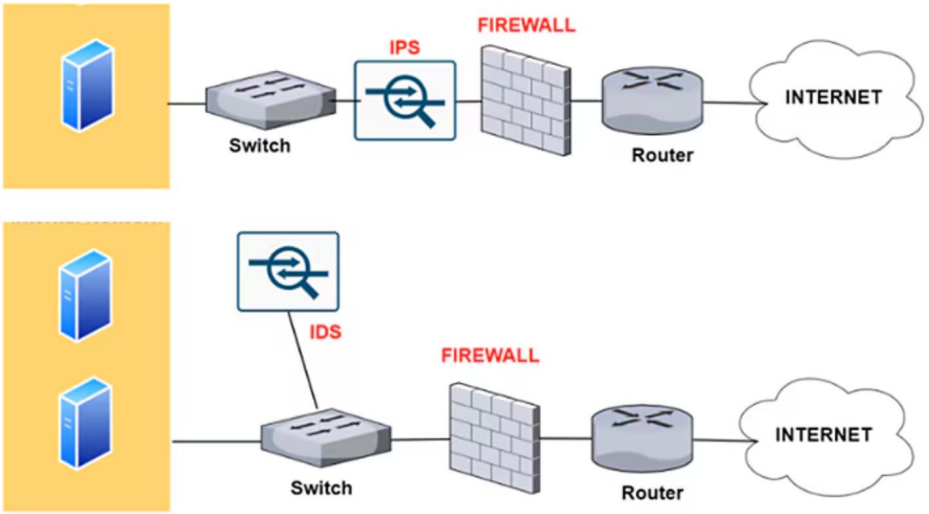


Alhajjar等人提出了一种基于进化计算的方法，通过整合三种前沿技术——粒子群优化算法、遗传算法和生成对抗网络，构建了多层次对抗样本生成框架。

Wang等人提出了MANDA，一种基于流形和决策边界对抗样本检测方法，通过分析对抗样本与原始数据流形的不一致性及决策边界邻近性检测对抗样本，保护入侵检测系统免受对抗攻击。

Roshan 等人提出了一种针对基于优化的白盒对抗攻击（C&W 攻击）的两阶段防御策略GDA-FS，通过训练阶段的高斯数据增强（GDA）和 测试阶段的特征压缩（FS），显著提升网络入侵检测系统的鲁棒性。

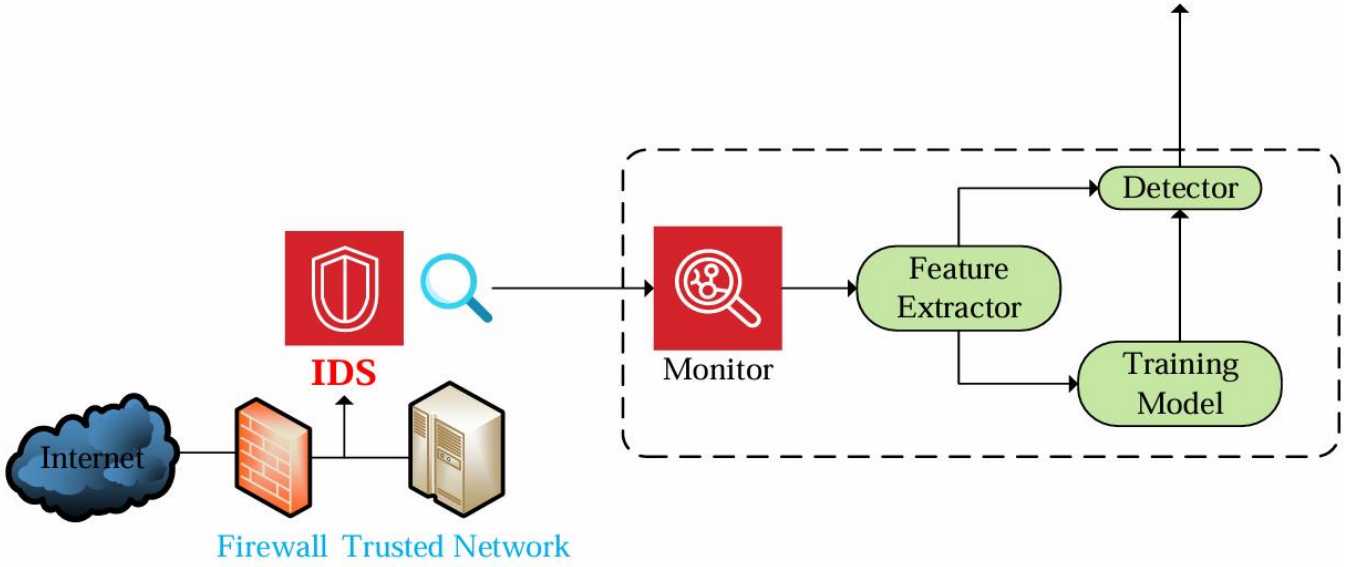
- 入侵检测系统（Intrusion Detection System, IDS）
  - 定义
    - 一种安全设备或软件应用程序
    - 通过持续监控网络流量或系统活动，识别恶意行为、可疑活动或安全策略违规行为，并发出警报
    - 被动系统，主要负责检测和报告，不直接阻止流量或采取防御措施
  - 分类



	类型	部署位置	核心原理
基于部署位置	基于网络 NIDS	网络关键节点	捕获并分析网络数据包
	基于主机 HIDS	单个主机	分析系统日志、文件完整性等主机内部活动

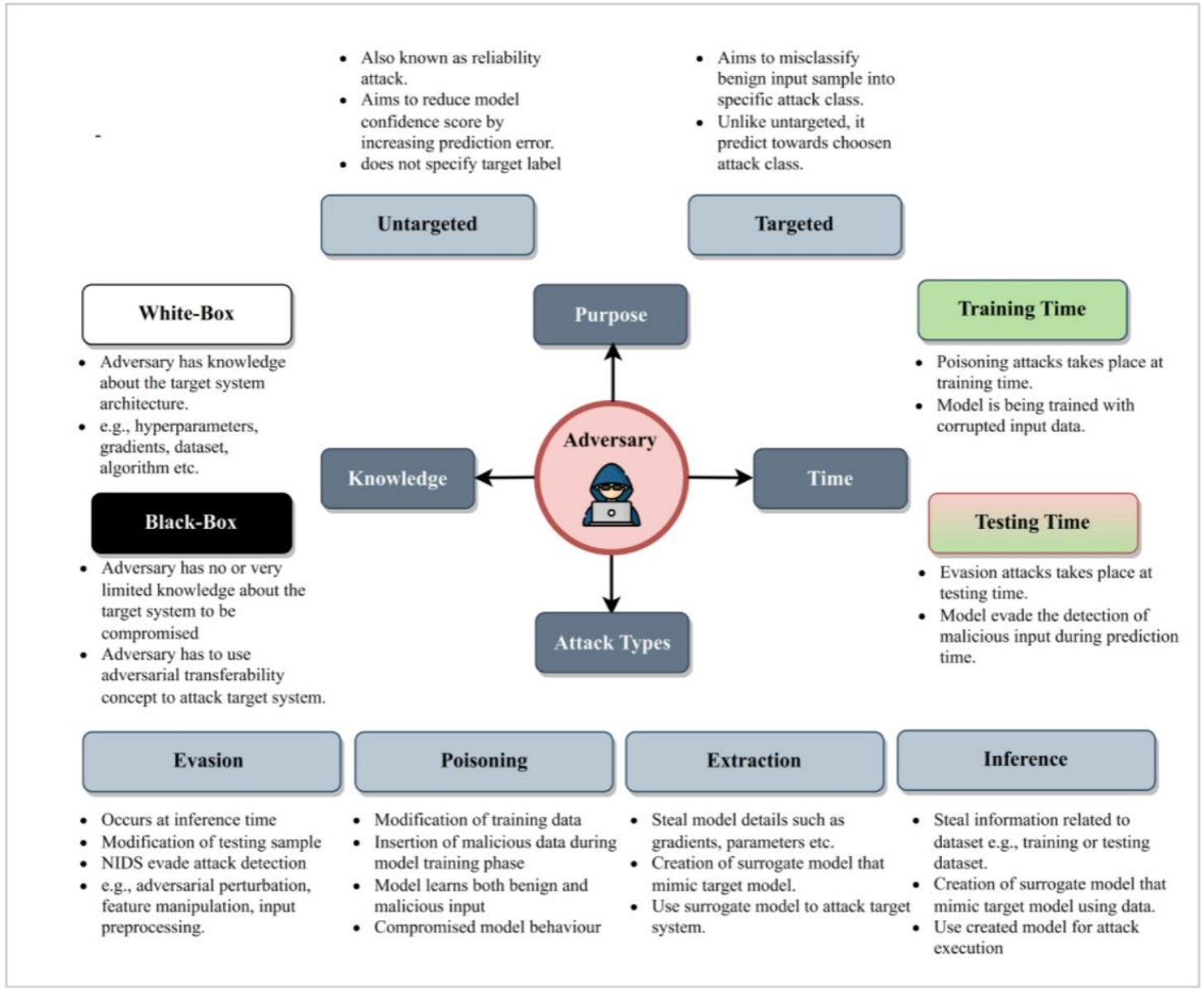
	类型	核心原理
基于检测技术	基于特征	与已知攻击特征库进行模式匹配
	基于异常	建立正常行为基线，识别偏离基线的活动
	混合型	结合前两种方法

- 网络入侵检测系统（NIDS）
  - 部署在防火墙之后、可信网络之前的一个关键网络节点上
  - 工作流程
    - 数据来源：Internet的外部流量经过Firewall的初步过滤后，一份副本会被发送到NIDS
    - 数据捕获：监控器实时抓取并复制流经该节点的所有数据包
    - 特征提取：特征提取器从原始数据中提取关键特征
    - 入侵判定：检测器将特征与训练模型/规则库进行比对分析



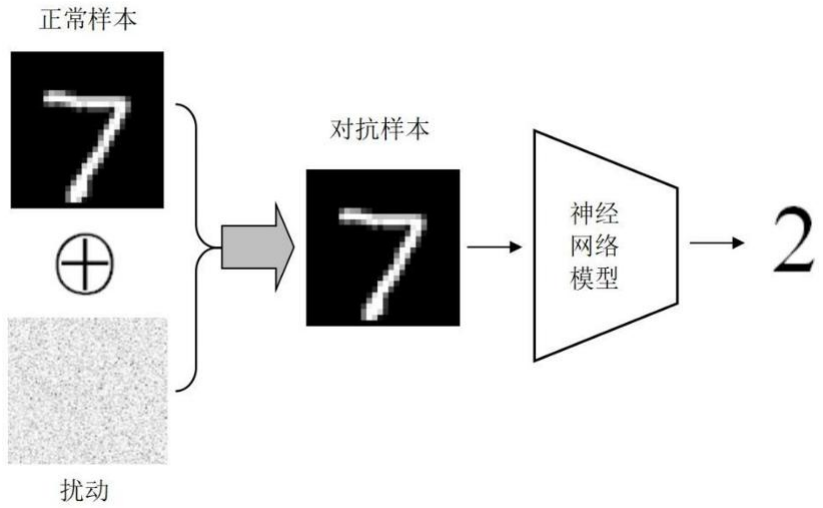
知识基础

- 对抗攻击
  - 攻击目的
    - 有目标攻击
    - 无目标攻击
  - 攻击时机
    - 训练阶段
    - 测试阶段
  - 攻击知识
    - 白盒攻击
    - 黑盒攻击
  - 攻击类型
    - 逃避攻击、提取攻击
    - 投毒攻击、推理攻击



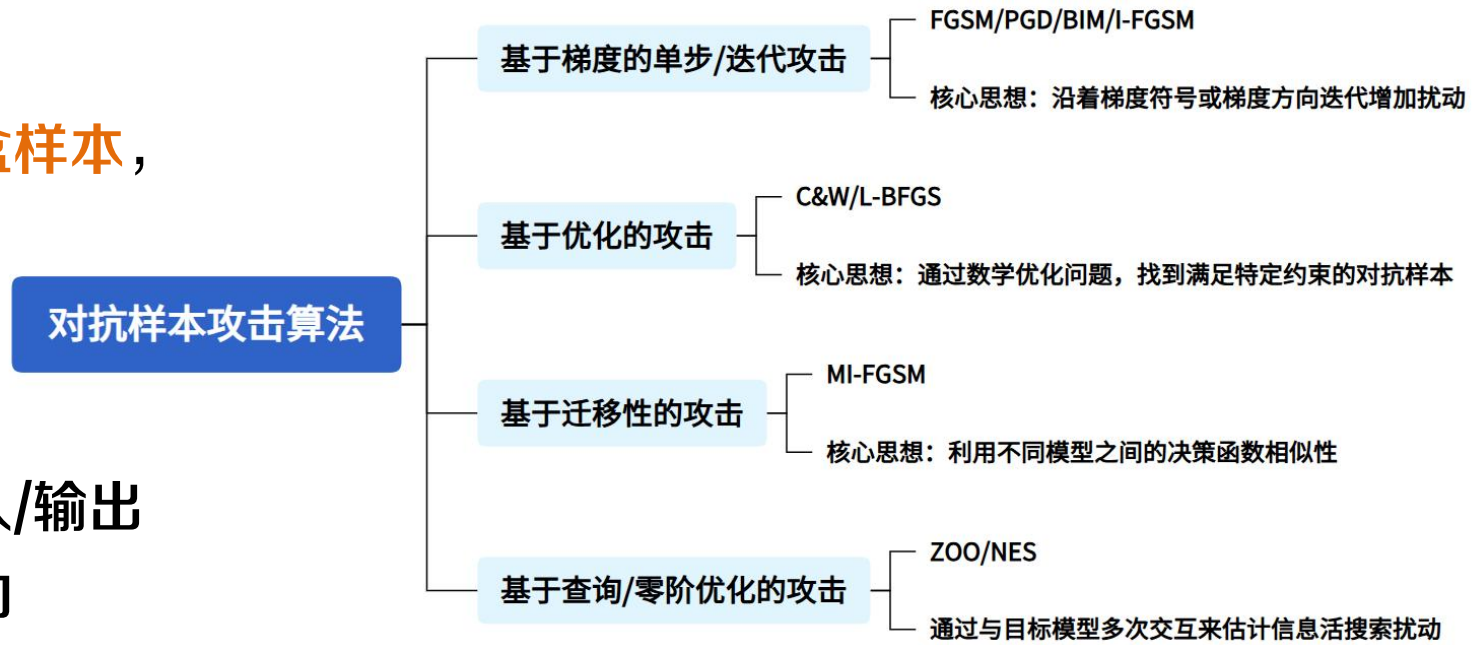
知识基础

- 对抗样本
  - 在原始样本中加入一些人为设计的、难以察觉的微小扰动而生成的样本
  - 目的：让机器学习或深度学习模型产生错误分类或者预测



- 对抗样本攻击算法
  - 基于梯度的单步/迭代攻击：FGSM/PGD/BIM/I-FGSM
    - 沿着梯度符号或梯度方向迭代增加扰动
    - 目标是最大化损失函数
  - 基于优化的攻击：C&W/L-BFGS
    - 通过数学优化问题，找到满足特定约束（最小扰动范数）的对抗样本
    - 最小化一个包含扰动范数和分类错误的组合损失函数

- 对抗样本攻击算法
  - 基于迁移性的攻击：MI-FGSM
    - 利用不同模型之间的决策函数相似性
    - 在入侵检测模型上生成白盒样本，用于攻击黑盒目标模型
  - 基于查询/零阶优化的攻击：ZOO/NES
    - 通过与目标模型的多次输入/输出交互来估计信息或搜索扰动
    - 使用有限差分或演化算法在黑盒设置下近似梯度或直接搜索



知识基础

- 对抗样本检测技术
  - 基于特征学习的对抗样本检测
    - 利用对抗样本与原始样本的不同特征来进行对抗样本检测
  - 基于分布统计的对抗样本检测
    - 通过分析模型输出层的概率分布特征，识别对抗样本与正常样本的统计差异
  - 基于中间输出的对抗样本检测
    - 利用对抗样本在深度神经网络中间层激活状态与正常样本的显著差异
- 对抗样本防御技术
  - 对抗训练
    - 对抗样本与原始样本混合训练模型，使其学习到更加平滑、稳健的边界
  - 特征去噪
    - 在对抗样本输入模型之前进行去噪处理，将攻击者千方百计添加到原始样本上的轻微干扰去除

Boosting robustness of network intrusion detection systems: A novel two phase defense strategy against untargeted white-box optimization adversarial attack

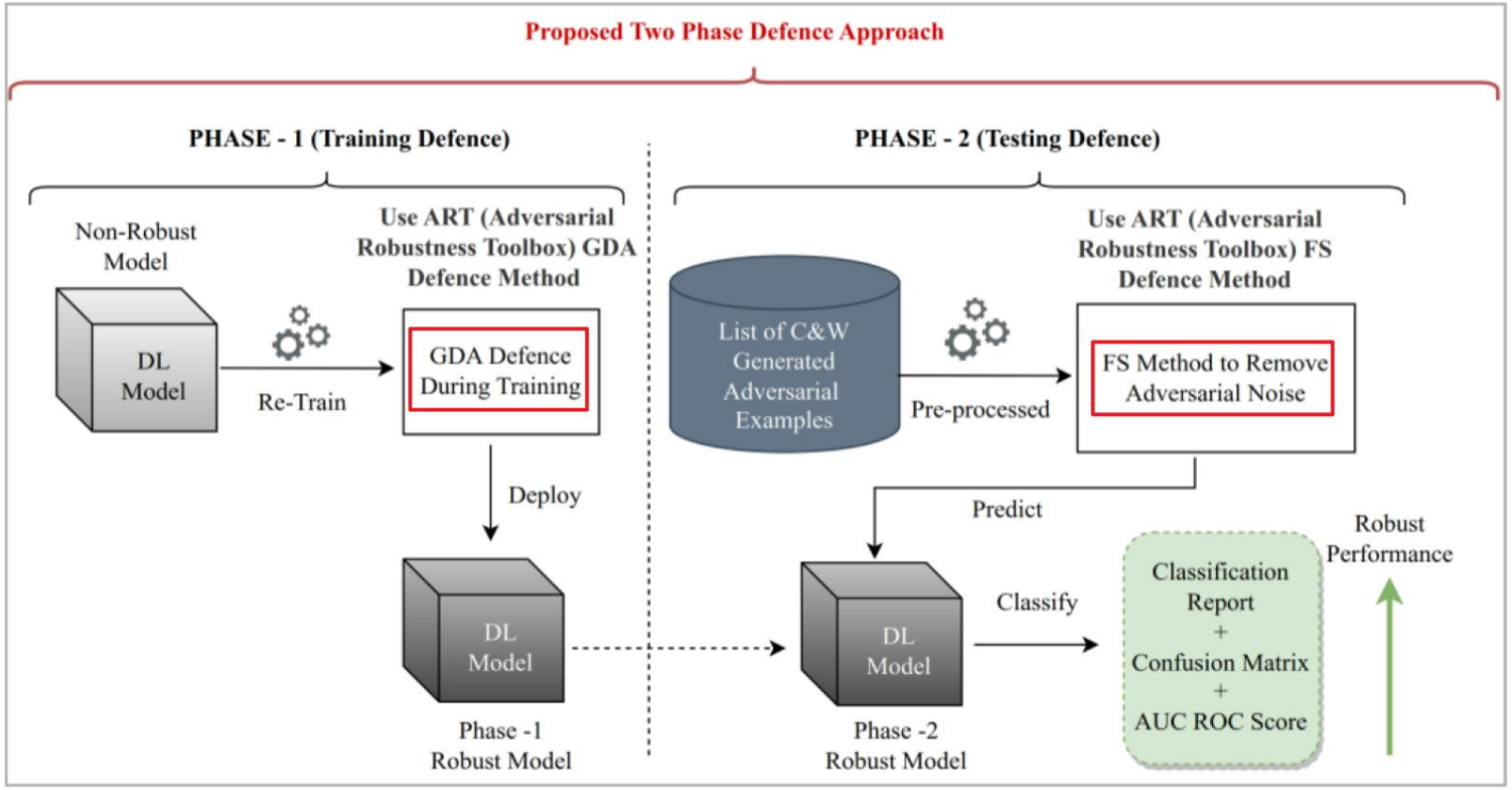
算法原理 TIPO

T	目标	提升NIDS对C&W对抗攻击的鲁棒性
I	输入	1组网络流量样本(良性/恶意/对抗)
P	处理	1.输入样本进行数据清洗、特征选择与特征归一化 2.GDA防御：向训练数据注入高斯噪声，通过对抗训练生成鲁棒模型 3.FS防御：对对抗样本特征压缩，过滤微小扰动后再输入鲁棒模型分类
O	输出	1个对对抗样本具有鲁棒性的模型

P	问题	1.NIDS模型面对C&W优化型白盒对抗攻击时表现脆弱，防御能力失效 2.大多数现有研究仅关注单一阶段（训练/测试）的防御，防护效果有限
C	条件	攻击假设：白盒攻击（攻击者已知模型参数和梯度）
D	难点	GDA可能降低模型对正常样本的敏感度，FS过度压缩会损失特征信息
L	水平	ESWA 2024（中科院一区）

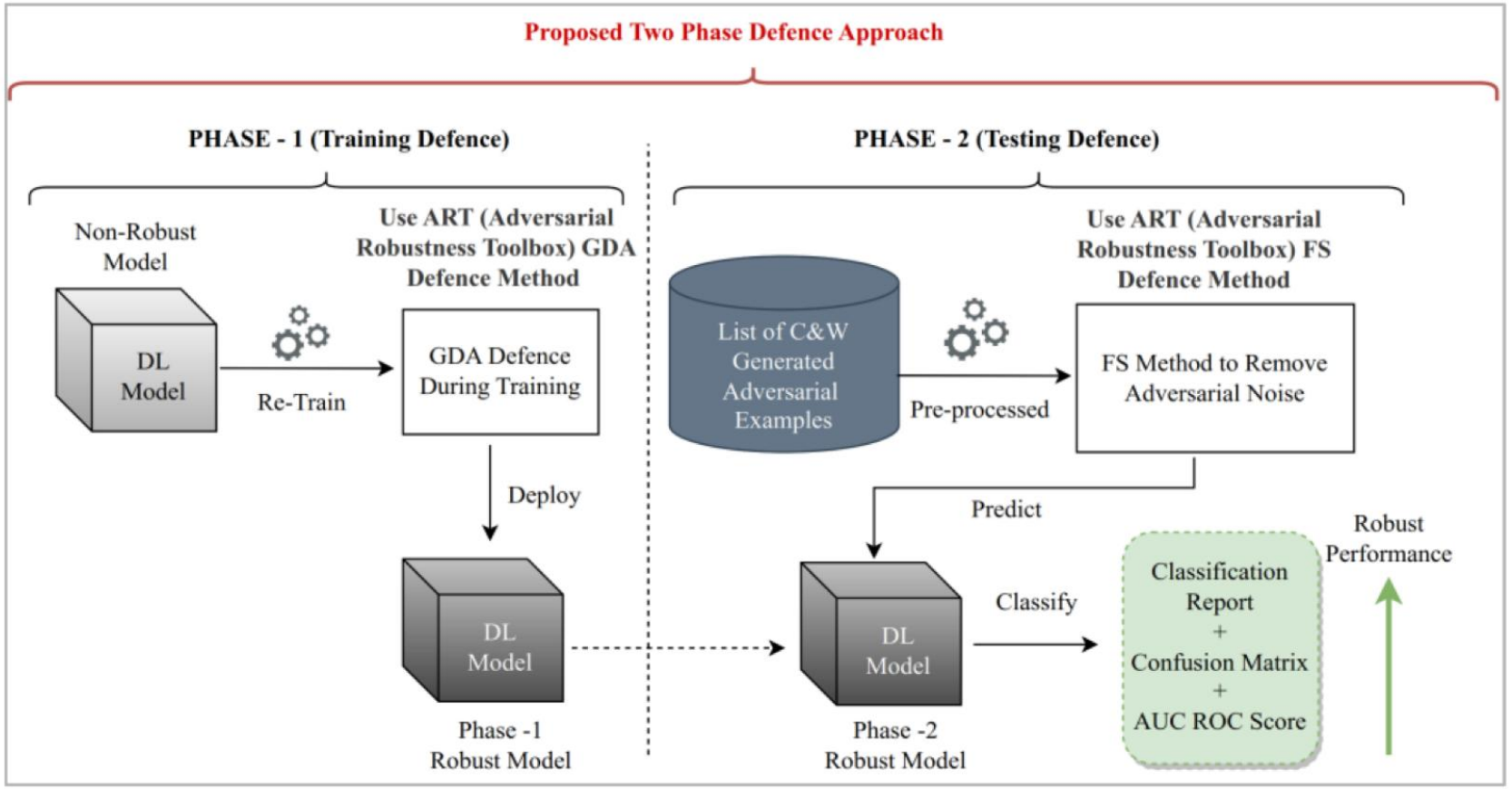
算法原理 创新说明

- GDA-FS
  - 以往工作多仅关注单阶段防御，本文首次在NIDS中验证两阶段联合防御的有效性
  - 训练阶段采用高斯数据增强（GDA）进行对抗训练；测试阶段应用特征压缩（FS）过滤对抗噪声



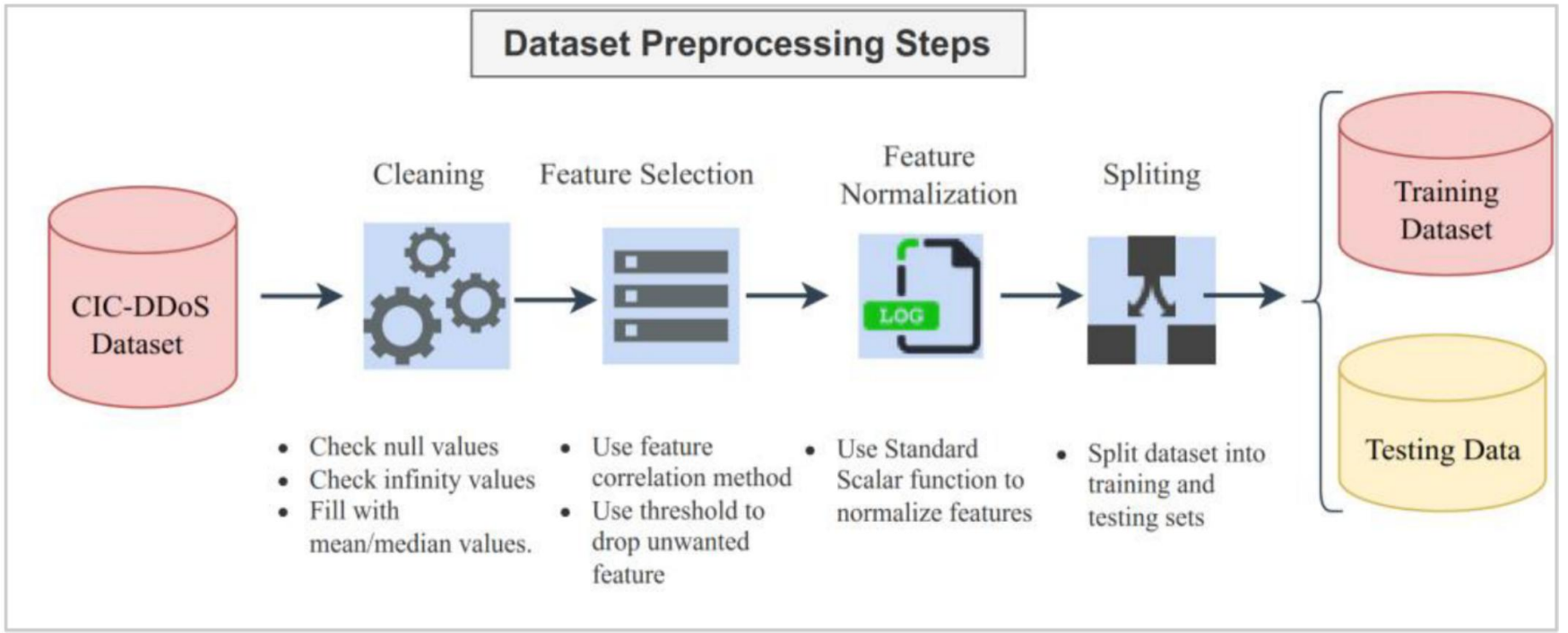
算法原理 GDA-FS

- GDA-FS
  - 前期准备
    - 数据预处理
    - 构建NIDS
    - C&W生成对抗样本
  - 核心步骤
    - 训练阶段防御GDA
    - 测试阶段防御FS



算法原理 数据预处理

- 数据预处理
  - 数据清洗：检查空值，检查无穷值，填充均值/中位数
  - 特征选择：计算特征之间相关性，删除相关值大于0.9的特征
  - 特征归一化
  - 数据集划分



算法原理 构建NIDS

- 基于深度学习的NIDS模型
  - 输入：良性/恶意流量数据样本
  - 模型结构
    - 输入层
    - 隐藏层：四层全连接层（ReLU）
    - 输出层
  - 超参数优化
    - 随机搜索
  - 损失函数
    - 二元交叉熵（Binary Cross-Entropy）

Hyperparameter	Values
Number of Layers	One Input + Four Hidden + One Output
Number of Neurons in Each Layer	40, 30, 15
Hidden Layer Activation Function	ReLU
Output Layer Activation Function	Sigmoid
Batch Size	4048
Learning Rate	0.0001
Number of Epochs	50
Optimizer	Adam
Loss Function	Binary_Crossentropy
Validation Split Parameter	0.25

算法原理 C&W生成对抗样本

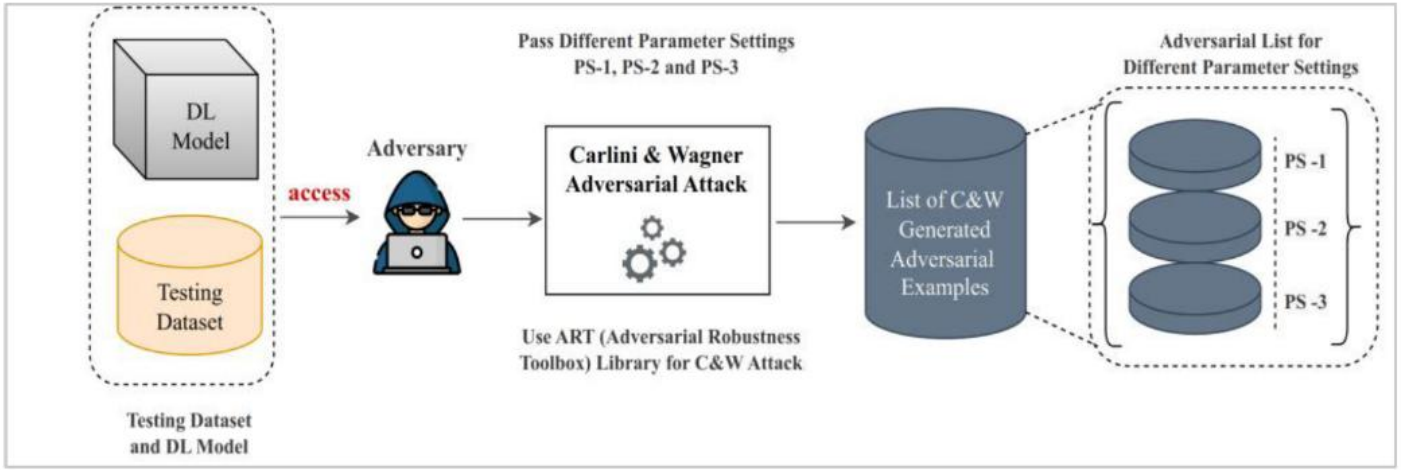
- C&W算法生成对抗样本
  - 攻击类型：基于优化的白盒攻击，攻击者已知模型结构和参数；无目标攻击
  - 带约束的优化方法
    - 最小化扰动σ

$$OptimizationFunction(OF) : minimize\, \sigma \,||\, \sigma \,||p \;+\; C^{*}F(x \;+\; \sigma)$$

    - 攻击损失函数——最大化模型的错误分类置信度

$$ObjectiveFunction\, f(x') \;=\; max(max\,\{H\,(x')i \;:\; i \neq\; t\} \;-\; H(x')t,\; -c)$$

  - 参数配置
    - PS-1（学习率0.0002）
    - PS-2（学习率0.0003）
    - PS-3（学习率0.0004）

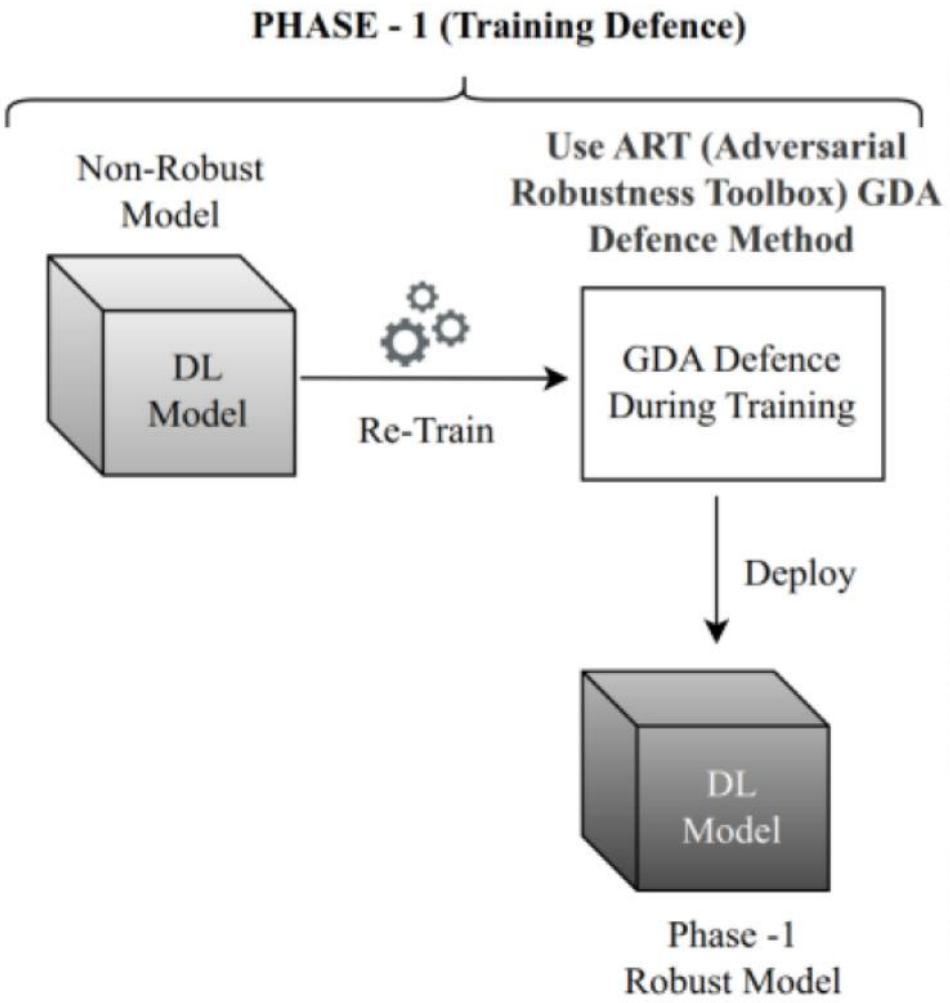


算法原理 训练阶段防御GDA

- **Gaussian data augmentation(GDA)**
  - 将高斯噪声显式纳入损失函数，约束模型在输入局部邻域内的预测稳定性
  - 目标：最小化模型在输入高斯邻域内的期望损失

$$\text{Minimum}\,\theta\,\mathrm{E}(\mathrm{x},\mathrm{y})\sim\mathrm{D}[\mathrm{E}\Delta\mathrm{x}\sim\mathrm{N}(0,\sigma 2)[\mathrm{J}(\theta,\mathrm{x}+\Delta\mathrm{x},\mathrm{y})]]$$

    - 对原始数据x添加一个微小的随机扰动Δx，扰动服从正态（高斯）分布，得到$x+\Delta x$
    - 训练模型参数θ，使模型不仅对原始数据x分类正确，还要对带噪声数据$x+\Delta x$ 分类正确
  - GDA和传统高斯数据增强的区别
    - 前者是目标层面约束
    - 后者是数据层面约束



算法原理 测试阶段防御FS

- **Feature squeezing(FS)**
  - 原始特征向量x映射为压缩向量$\hat{x}$ 使得$|\hat{x}-x|\leq\epsilon$

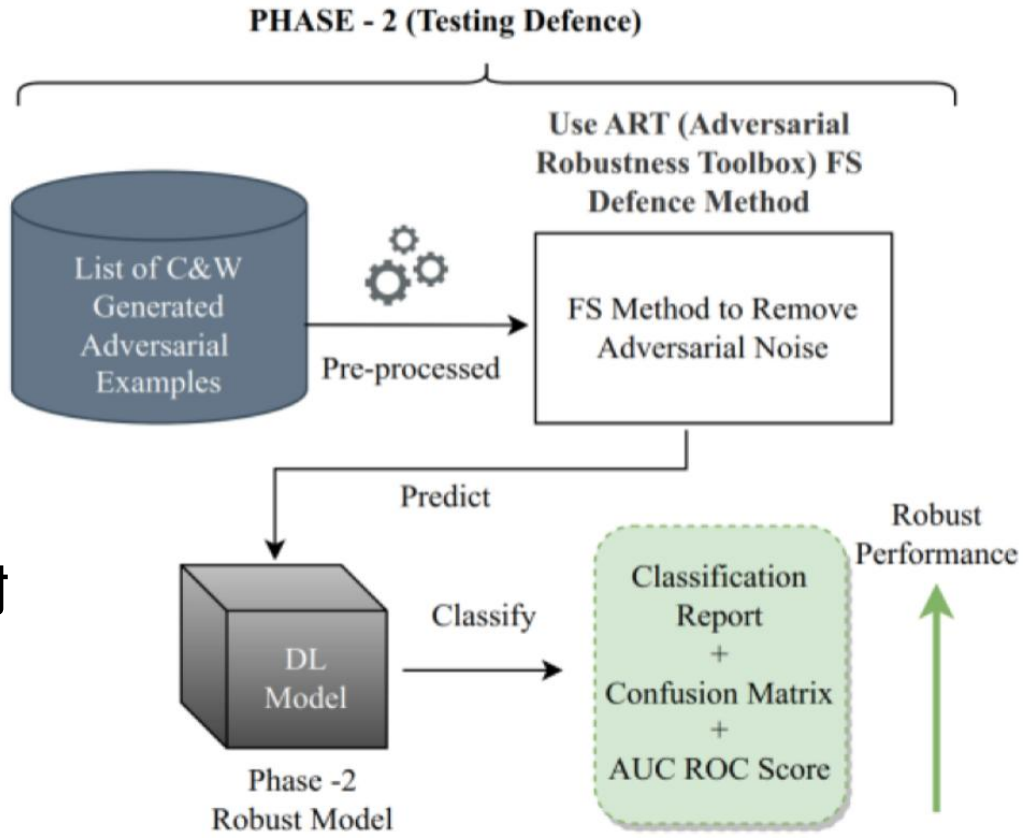
    削弱对抗样本的影响

$$|\text{Squeeze}(\mathbf{x}') - \text{Squeeze}(\mathbf{x})| \approx 0$$

- 两种特征压缩技术
  - 特征量化
    - 降低数值特征的精度来实现压缩
    - 对于网络流量中的每一个数值特征，将其映射到一个预定义的、较低数量的离散值中

$$\hat{x}_i = \text{round}\left(\frac{x_i \cdot N}{R}\right) \cdot \frac{R}{N}$$

  - 特征平滑/滤波
    - 对相邻特征值进行局部平滑
    - 方法：滑动平均、一维高斯滤波器



- 数据资源—CIC-DDoS-2019数据集

良性样本	恶意样本	训练数据集	测试数据集
107,764	119,384	131,745 (良性 62,391, 恶意 69,354)	95,403 (良性 45,373, 恶意 50,030)

- 评价指标
  - 准确率
  - 召回率
  - 精确率
  - F1-Score
  - AUC-ROC

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

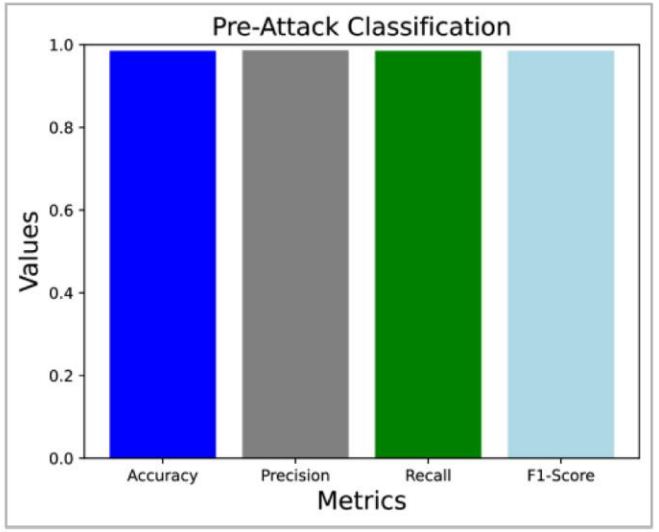
$$Precision = \frac{TP}{TP + FP}$$

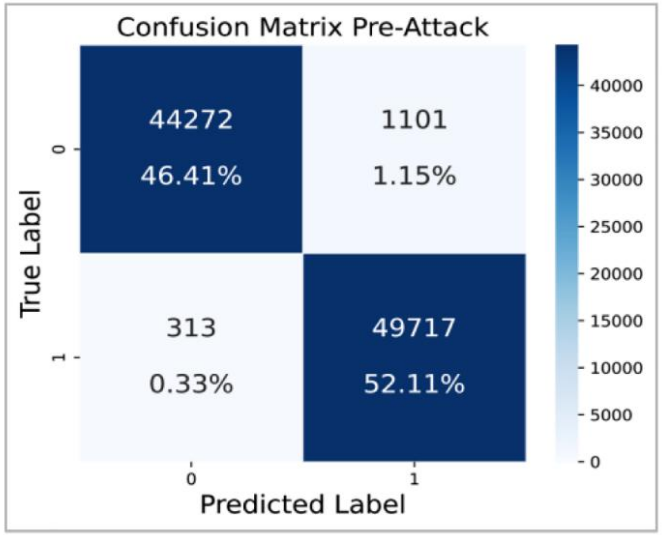
$$F1\text{-}score = 2 \cdot \frac{Precision * Recall}{Precision + Recall}$$

- 对抗攻击指标
  - Evasion Rate（规避率）

实验结果 对抗攻击测试

- NIDS未受对抗攻击时的性能
  - 对正常网络流量和攻击流量的基础分类能力优秀
    - 在干净数据集上的分类准确率达到 98.5%
    - 其他关键指标同样表现优异
  - 模型对正常和恶意活动的区分能力极强
    - AUC-ROC分数为 0.9847
  - 具有较高的检测可靠性
    - 误报率（1.15%）和漏报率（0.33%）均比较低



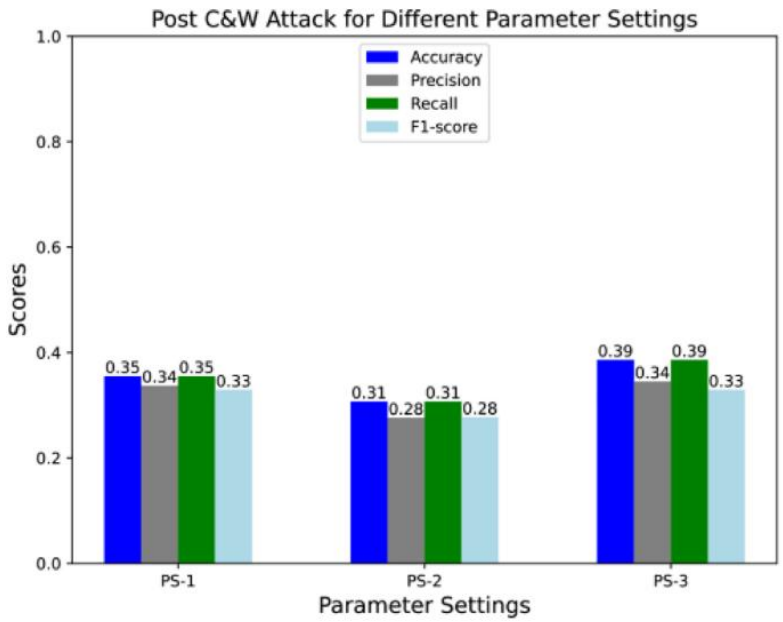


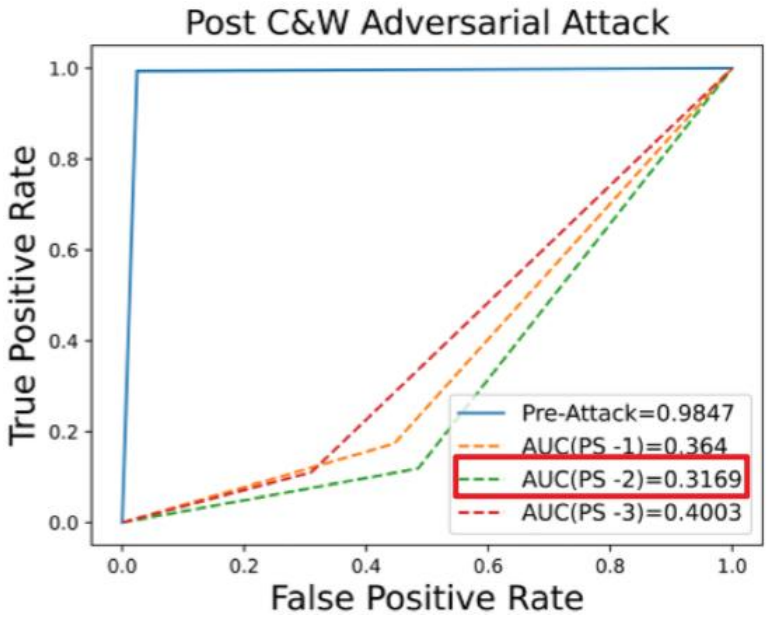
Name	Classification Report				Confusion Matrix				Area under the ROC
Pre-Attack Evaluation	A	P	R	F	TN	FP	FN	TP	AUC-ROC
(Weighted Average)	0.985	0.986	0.985	0.985	44,272	1101	313	49,717	0.9847

A – Accuracy, P – Precision, R – Recall, F-F1Score, TN – True Negative, TP – True Positive, FP - False Positive, FN – False Negative.

实验结果 对抗攻击测试

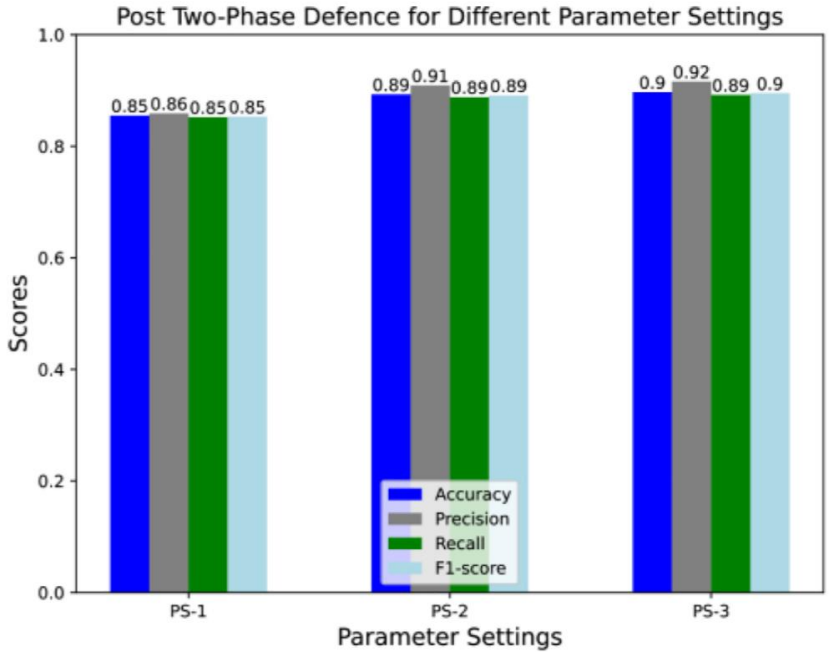
- NIDS受到对抗攻击后的性能
  - 模型性能显著下降，对白盒攻击极度脆弱
    - 原始模型准确率显著下降98.5%→30.7%，C&W攻击能有效欺骗模型
    - 模型区分正常与恶意流量的能力几乎失效（AUC-ROC从0.9847→0.3169~0.4003）
  - 攻击强度与参数配置相关
    - PS-2攻击效果最强；PS-3攻击效果略弱，可能因过大的扰动被部分检测到

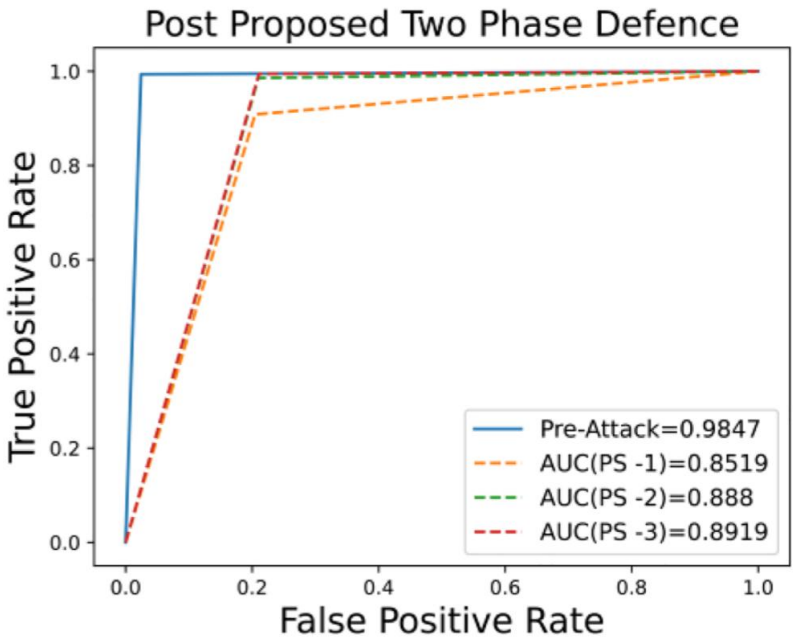




实验结果 鲁棒性测试

- 评估GDA-FS的鲁棒性
  - 显著提升模型的鲁棒性能
    - 模型准确率分别达到85.5%、89.3%和89.7%，较攻击后性能（PS-1仅35.5%）显著提升
  - 对不同攻击强度的对抗样本均有较高鲁棒性
    - PS-1至PS-3的AUC-ROC分别为0.8519、0.888和0.8919

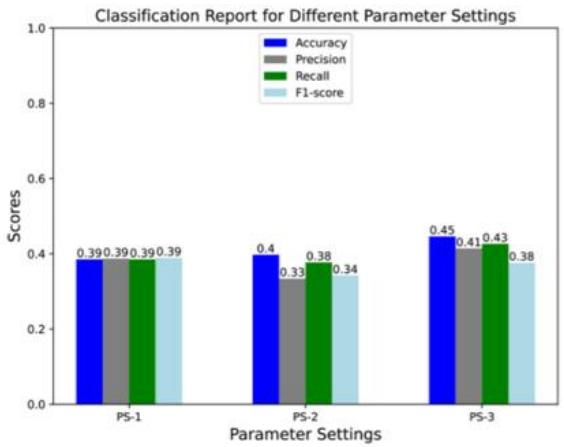


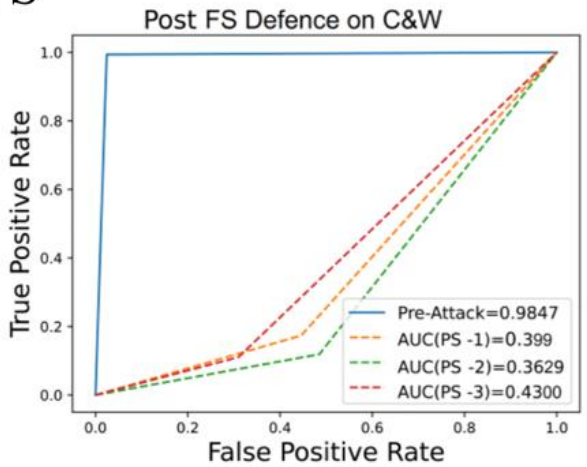


实验结果 消融实验

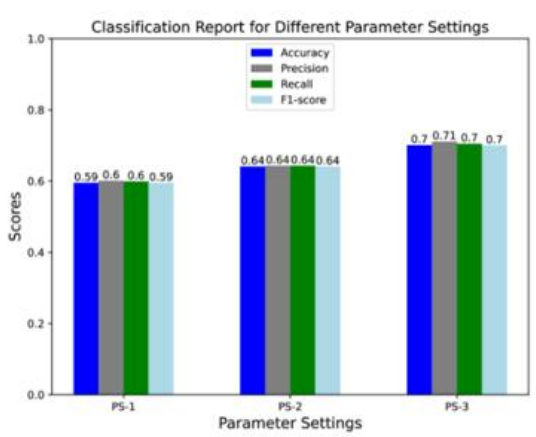
- GDA/FS单阶段防御
  - 单独使用任一种防御均无法达到两阶段组合的效果
  - GDA优于FS
    - GDA单阶段防御在训练阶段增强模型鲁棒性，效果明显优于仅用FS的测试阶段处理
    - FS虽能压缩特征噪声，但无法有效纠正对抗扰动导致的误分类

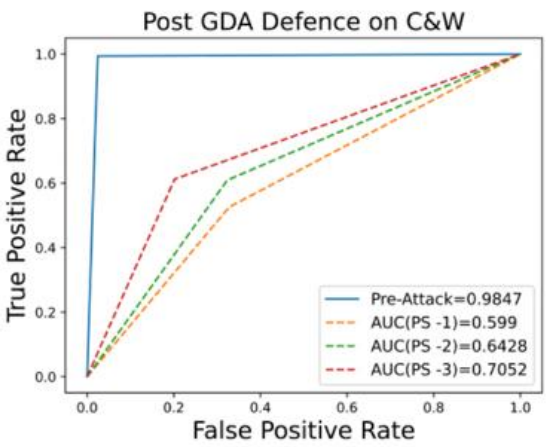
FS





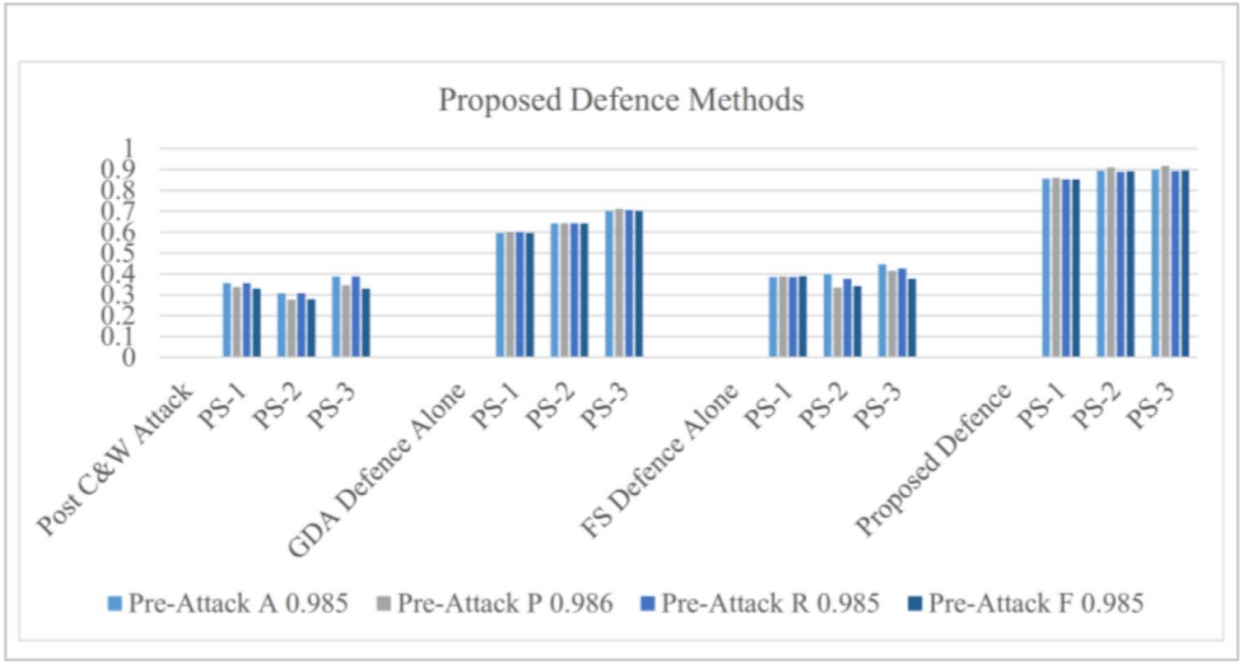
GDA





实验结果 实验总结

- GDA-FS两阶段防御策略的显著优势
  - GDA-FS策略使模型在C&W攻击下的性能，接近无攻击时的基准水平
- 单阶段防御的局限性
  - GDA：虽能部分提升鲁棒性，但无法完全抵抗对抗攻击
  - FS：效果较弱（PS-3准确率44.6%），单独依赖特征压缩不足以应对对抗扰动
- 对抗攻击的严重威胁
  - C&W攻击使模型性能断崖式下降
  - 深度学习模型在对抗样本下的脆弱性，凸显防御必要性



NIDS-DA: Detecting functionally preserved adversarial examples for network intrusion detection system using deep autoencoders

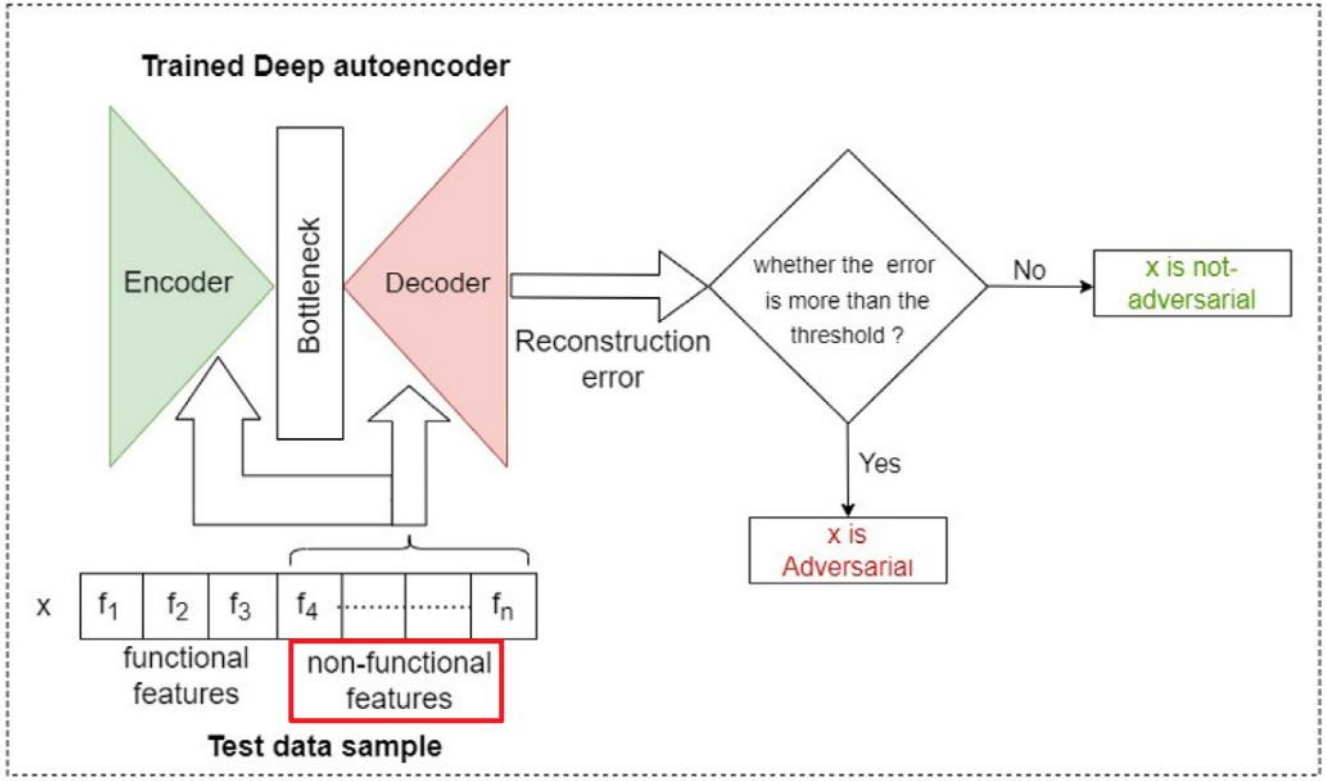
算法原理 TIPO

T	目标	检测功能保留型网络流量对抗样本
I	输入	1组网络流量样本（良性/恶意/对抗）
P	处理	1.输入样本进行归一化、特征选择（Chi-2检验） 2.替代分类器（模拟NIDS）区分良性/恶意样本 3.被分类为“良性”的样本送入DAE检测 4.DAE计算重构误差并与阈值比较，判定是否为对抗样本
O	输出	1组网络流量样本标签（是否为对抗样本）

P	问题	功能保留型对抗样本仅修改不影响恶意功能的特征来欺骗NIDS，传统基于全特征或监督学习的检测方法难以有效区分，且误报率高
C	条件	1.对抗样本需保持功能性特征不变，仅对非功能性特征添加扰动 2.攻击者不知道模型的参数、配置和梯度信息，但可以访问用于训练目标模型的相同训练数据
D	难点	如何区分网络流量的功能性特征与非功能性特征
L	水平	ESWA 2025（中科院一区）

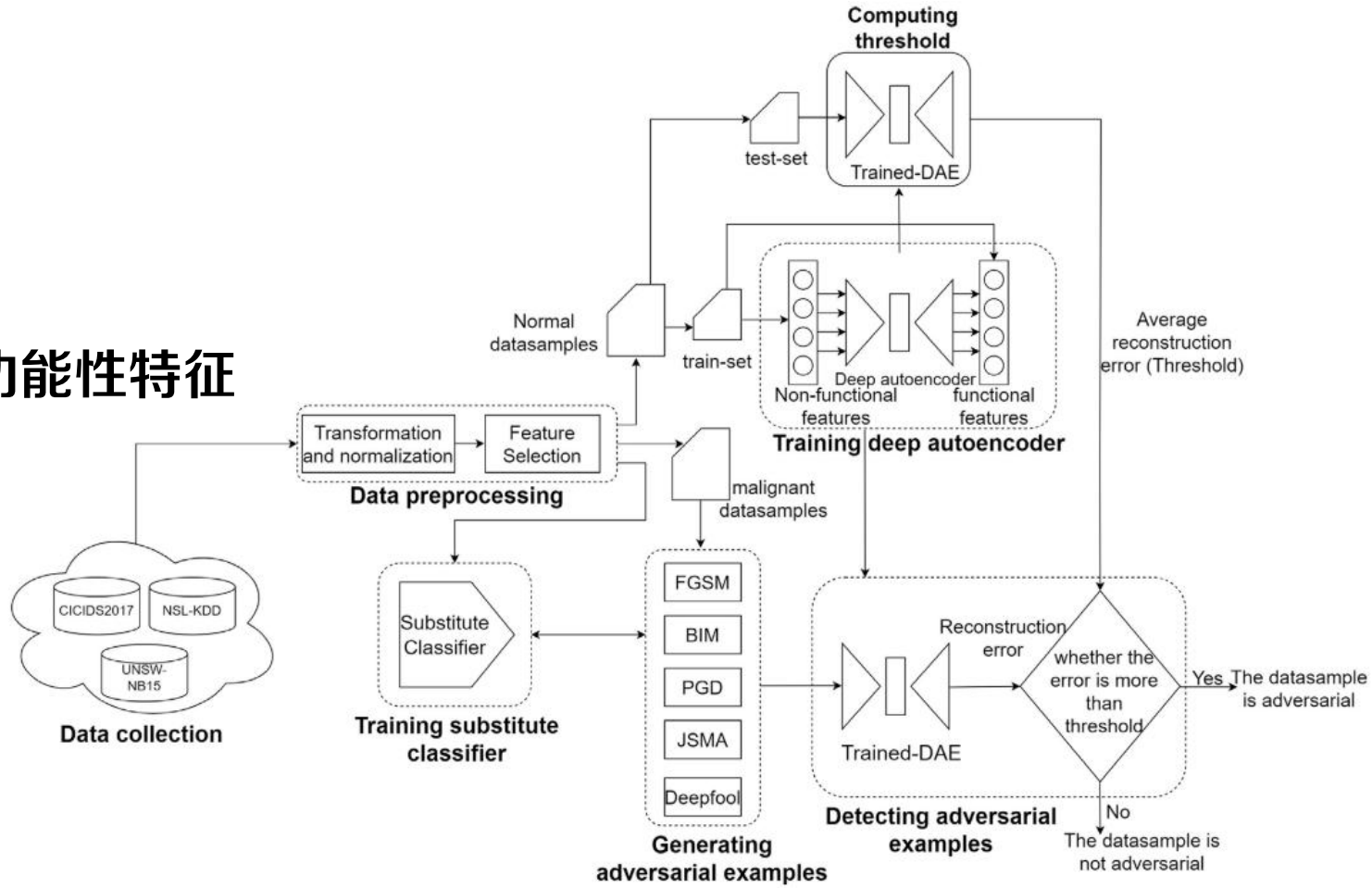
算法原理 创新说明

- **NIDS-DA**
  - 攻击者通过保留功能性特征，仅对非功能性特征添加扰动来生成对抗样本
  - 提出利用深度自编码器（DAE）学习非功能性特征的内在模式与分布，对功能保留型对抗样本进行检测



算法原理 NIDS-DA

- NIDS-DA
  - 数据预处理
    - 最小-最大归一化
    - Chi-2方法进行特征选择
  - 区分样本功能性特征与非功能性特征
  - 替代分类器（NIDS）训练
  - 生成对抗样本
  - 训练深度自编码器
  - 对抗样本检测



算法原理 功能性特征与非功能性特征

- 功能性特征
  - 对特定攻击类型实现其恶意目标必不可少的特征，修改这些特征会导致攻击失效
    - DoS攻击：时间相关特征和协议类型
    - R2L攻击：内容相关特征

Functional features of various attacks in NSL-KDD dataset.

Attack	Intrinsic	Content	Time-based features	Host-based features
Probe	✓		✓	✓
Dos	✓		✓	
U2R	✓	✓		
R2L	✓	✓		

- 非功能性特征
  - 与攻击目标无关的特征，修改后不影响攻击有效性
    - DoS攻击：登录失败次数或DNS查询长度
- 区分方法
  - 基于统计技术的方法　　特征与标签之间的相关性
  - 基于行为和分布模式的方法
  - 基于降维技术的方法
    - PCA降维对特征分组，功能特征出现在方差大的组中

算法原理 替代分类器训练与对抗样本生成

- 训练替代分类器
  - 基于深度神经网络（DNN）
  - 输入：良性/恶意流量
  - 目标：区分良性/恶意流量
- 生成对抗样本—以DoS为例
  - 输入：恶意流量数据样本
  - 攻击算法：FGSM/BIM/PGD/JSMA/Deepfool
  - 生成过程
    - 固定功能性特征
    - 使用攻击算法计算非功能性特征的扰动方向
    - 替代分类器将其分为“良性”，迭代完成
    - 将样本加入对抗样本集合

Algorithm 1: Generating adversarial examples with constrained perturbation

Input : Dataset D of Dos datasamples and trained substitute classifier f

Output: Adversarial examples

$NF \leftarrow \{f_1, f_2, ...f_k\}$; /* Non functional features in D */

$Adv \leftarrow \{\}$

Initialize perturbation parameter δ

$Adv - Gen - Algo \leftarrow \{FGSM, BIM, PGD, JSMA, Deepfool\}$

for $\forall x$ *in* D **do**

  $x^* \leftarrow Adv - Gen - Algo(x, f, NF, \delta)$; /* function call returns adversarial example x* by perturbing only non-functional features NF */

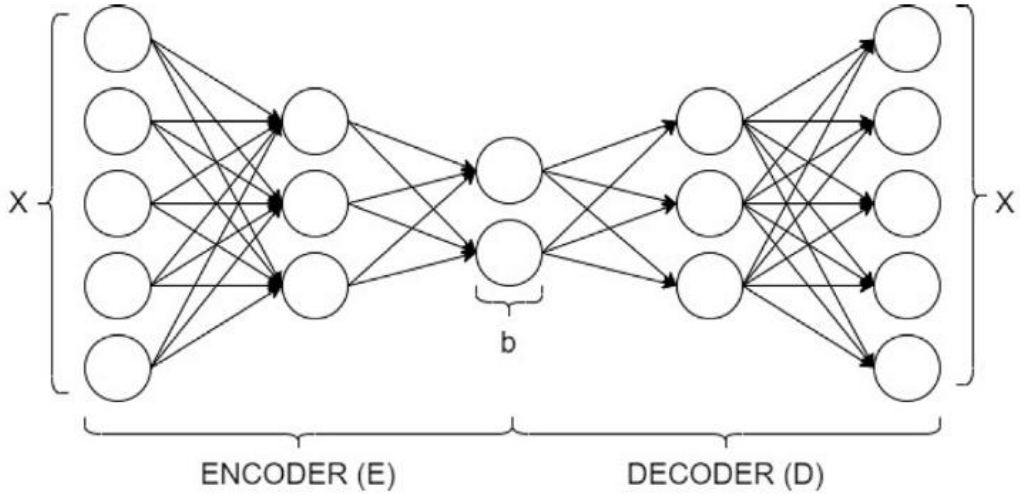
  $Adv \leftarrow Adv \cup x^*$

end

return Adv

算法原理 深度自编码器

- DAE结构
  - 输入：正常样本（良性）的非功能性特征
  - 编码器-解码器对称架构
    - 编码器E：输入数据x映射为低维、稠密的潜在表示b
    - 瓶颈层：编码过程的输出，解码过程的输入（b）
    - 解码器D：从潜在表示 b中重构出原始数据x
  - 输出：即解码器的输出，模型对原始输入的重构数据，目标是尽可能使重构数据接近原始的数据 x



算法原理 深度自编码器

- DAE训练
  - 训练目标：最小化输入数据x与重构数据x'之间的差异

$$Reconstruction\ error = \underset{E,D}{argmin}\ \|x - x'\|^2$$

  - 训练过程
    - 初始化：随机初始化编码器和解码器的权重与偏置，设定训练参数
    - 利用DAE重构数据，计算损失
    - 反向传播更新模型参数
  - 阈值计算
    - 将测试集正常样本输入训练好的DAE
    - 计算每个样本的重构误差（RMSE）
    - 计算所有测试样本的平均误差，作为检测阈值

Algorithm 2: Training Deep autoencoder

Input : Dataset D of normal datasamples
Output: Trained DAE and threshold value

$D_{train}, D_{test} \leftarrow train\_test\_split(D)$
$D_{train\_non\_func} \leftarrow D_{train}$ {non functional features}
$D_{test\_non\_func} \leftarrow D_{test}$ {non functional features}
Initialize learning_rate, epochs, batch_size and nos. of layers
Initialize weights and biases of Encoder E and Decoder D
while $epochs \geq 0$ **do**
    $encoded\_output \leftarrow E(batch, D_{train\_non\_func})$
    $decoded\_output \leftarrow D(batch, encoded\_output)$
    $loss \leftarrow compute\_loss(batch, decoded\_output)$
    compute_gradient(loss)
    update_weight_and_biases(learning_rate)
    $epochs \leftarrow epochs - 1$
end
; /* Using trained DAE for computing threshold */
$aggregate\_error \leftarrow 0$
for $\forall x$ *in* $D_{test\_non\_func}$ **do**
    $R.E \leftarrow RMSE(x, D(E(x))$
    $aggregate\_error \leftarrow R.E + aggregate\_error$
end
$threshold \leftarrow \frac{aggregate\_error}{size(D_{test\_non\_func})}$
return E, D, $threshold$

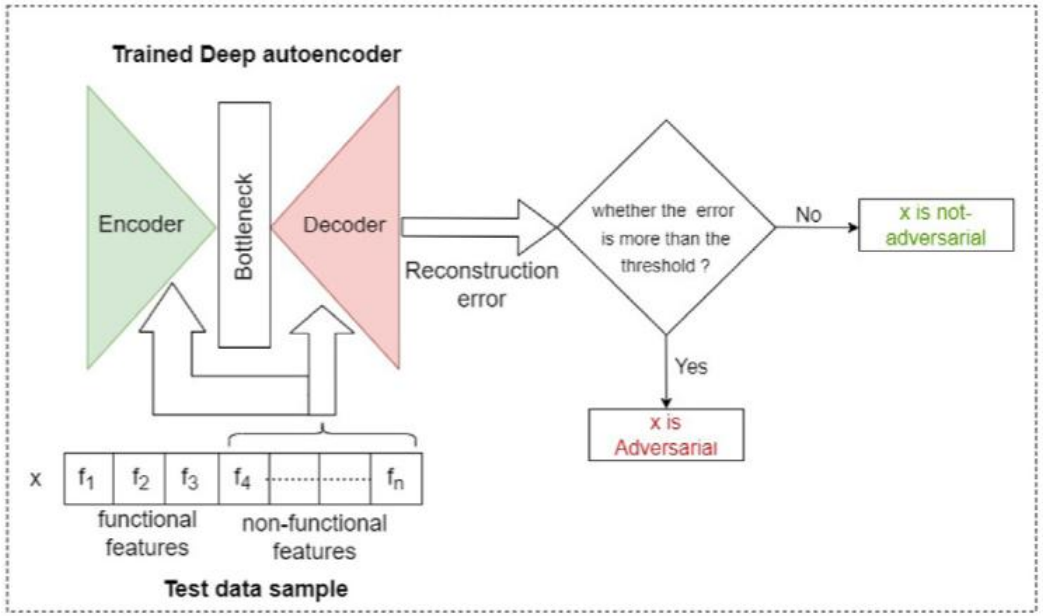
算法原理 对抗样本检测

- DAE检测对抗样本
  - 输入：待检测样本（分类器判别为“正常的”）
  - 检测过程
    - 提取样本非功能性特征输入DAE
    - 计算重构误差RE
    - RE 大于阈值$threshold$ 判定为对抗样本；否则为正常样本

Algorithm 3: Detecting adversarial examples

Input : Test sample x, trained DAE and threshold
Output: Whether x is adversarial or not

$RE \leftarrow RMSE(x, D(E(x)))$
if $RE > threshold$ **then**
  return $True$
else
  return $False$
end



实验流程 实验设置

- 数据资源

数据集	特征数	正样本数	负样本数	样本总数
UNSW-NB15	43	37,000	4,089	41,089
CICIDS2017	78	97,686	128,025	225,711
NSL-KDD	42	67,342	45,927	113,269

- 评价指标
  - 准确率
  - 召回率
  - 精确率
  - F1-Score

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad Recall = \frac{TP}{TP + FN}$$

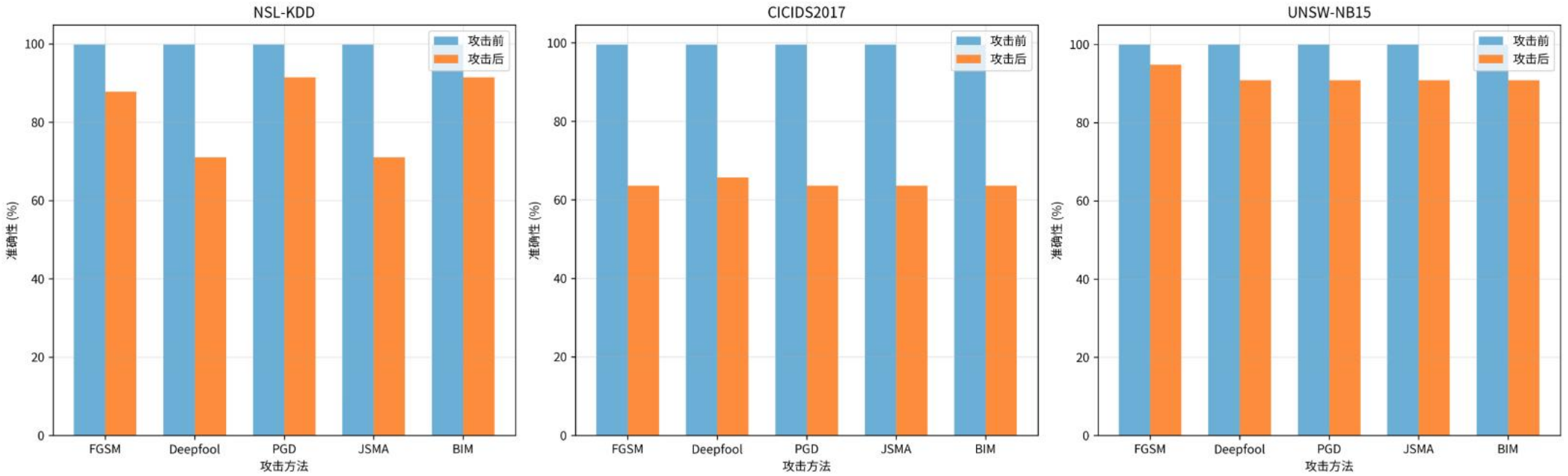
$$Precision = \frac{TP}{TP + FP} \qquad F1\text{-}score = 2 \cdot \frac{Precision * Recall}{Precision + Recall}$$

- 对抗攻击方法
  - FGSM、BIM、PGD、JSMA、DeepFool

实验结果 对抗攻击测试

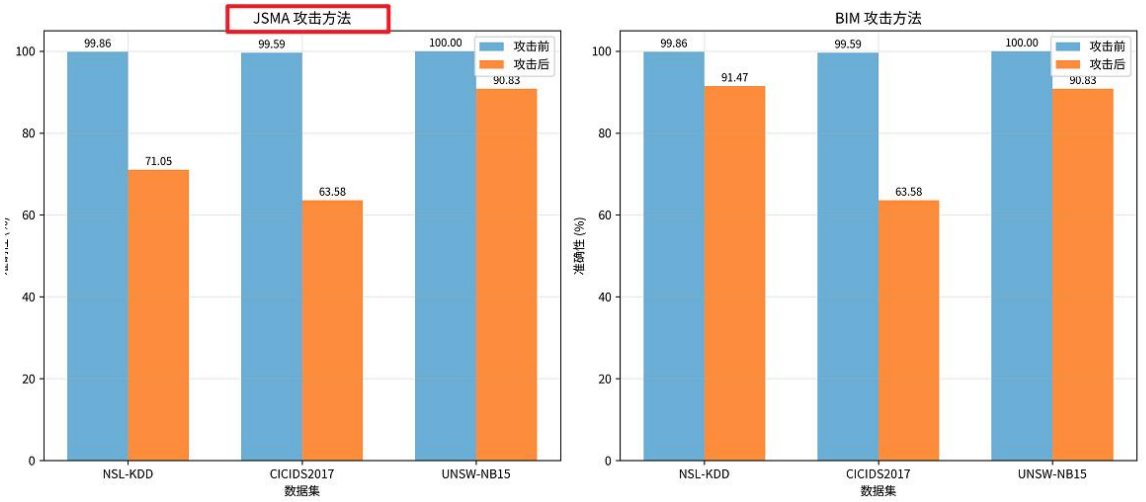
- 评估替代分类器在对抗攻击下的性能
  - 数据集：混合数据集（正常/恶意/对抗）
  - 所有数据集在攻击前都具有很高的准确性
  - 攻击后的准确率显著降低，CICIDS2017数据集受影响最大
  - 替代分类器将相当大比例的对抗样本错误地分类为正常，假阳率激增

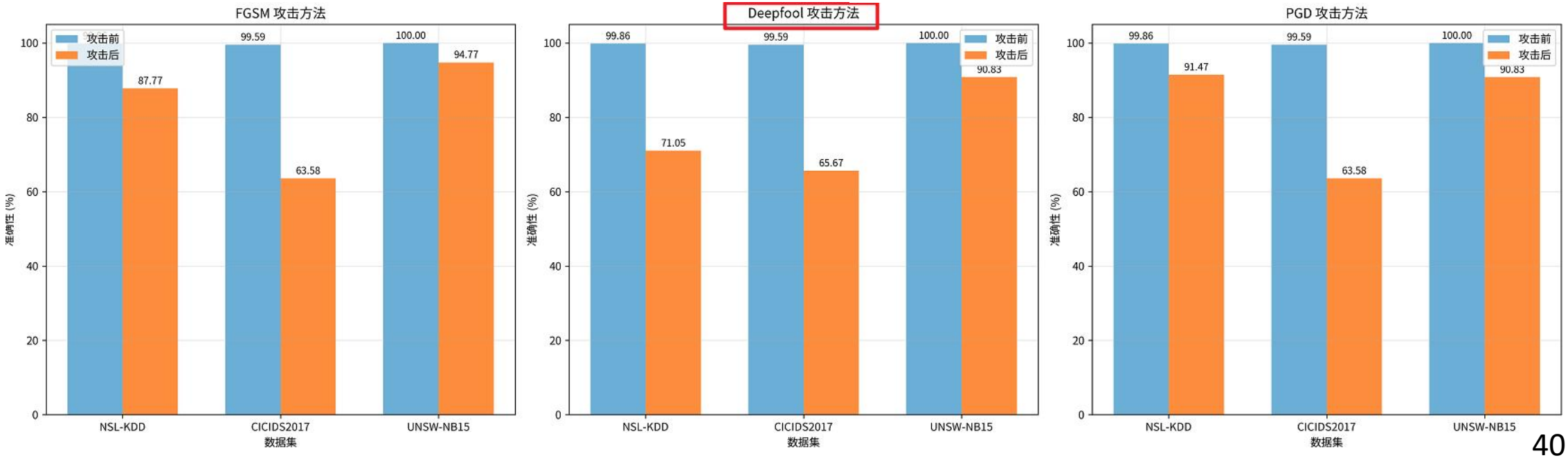
Dataset	Normal data samples	Dos samples	Adversarial examples
UNSW-NB15	11 083	1244	1244
CICIDS2017	9773	12 799	12 708
NSL-KDD	13 465	9189	3121



实验结果 对抗攻击测试

- 不同对抗攻击对替代分类器准确性的影响
  - Deepfool和JSMA攻击方法使分类器准确性下降的最多
  - UNSW-NB15数据集训练出的分类器相对于其他两个数据集表现出更强的鲁棒性

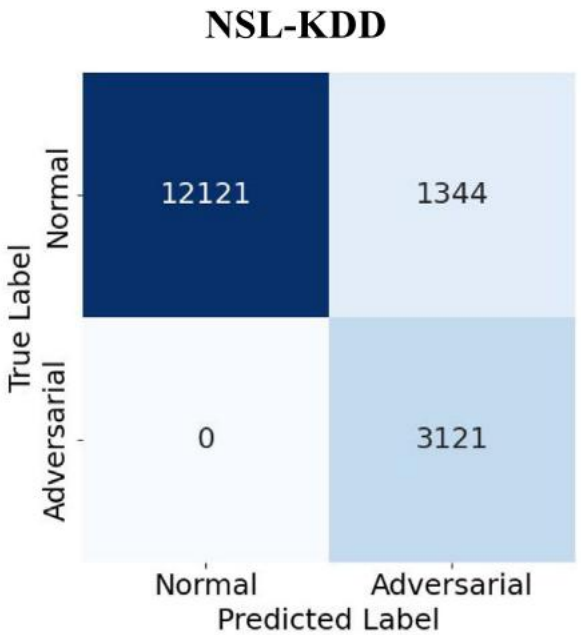


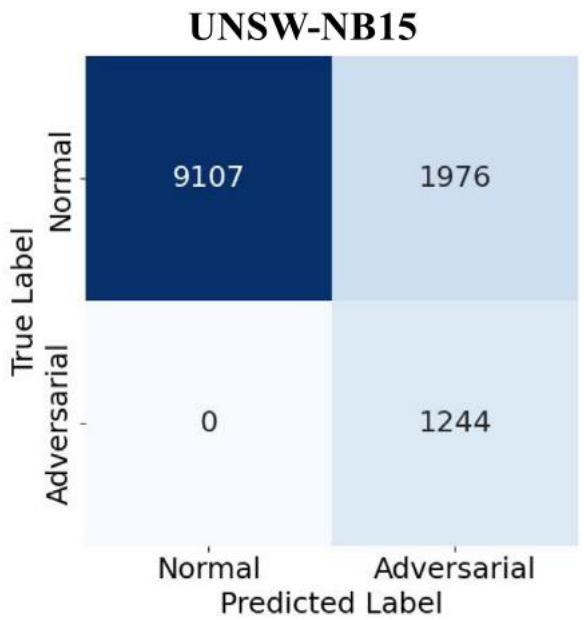


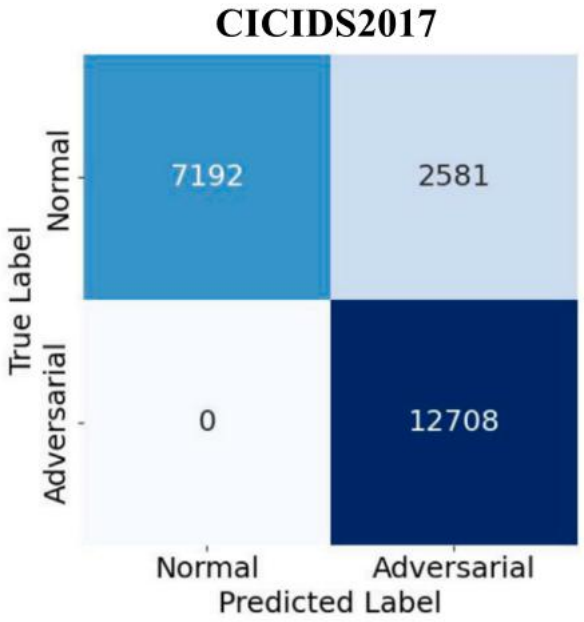
实验结果 对抗性测试

- 针对混合数据集评估DAE的对抗性检测能力
  - 在三个数据集上分别训练DAE，得到阈值
  - DAE成功地检测到所有对抗样本
  - 在识别攻击方面具有较高的可靠性，跨数据集表现稳定

Dataset	Threshold value
UNSW-NB15	0.0025
CICIDS2017	0.0000413
NSL-KDD	0.00109







实验结果 对比实验

- NIDS-DA方法在所有数据集上均具有较高的检测精度
  - 具有较强的鲁棒性和泛化能力
- 所提方法的性能显著优于近年来的先进方法
  - 与Alslman等人（2024）使用去噪自编码器检测相比，提升约15%
  - 与Roshan等人的对抗训练、高置信度等复杂防御策略相比，提升约11%
  - 与Debicha等人使用基于迁移学习的多个对抗样本检测器检测相比，提升18%

Detection method	Datasets		
	CICIDS2017	UNSW-NB15	NSL-KDD
Our Work	**99.92**	**99.92**	**99.97**
Alslman et al. (2024)	-	85.00	–
Roshan et al. (2024)	98.72	-	–
Debicha, Bauwens, et al. (2023)	88.67	-	81.85

实验结果 消融实验

- NF-DAE方法与FF-DAE方法相比
  - NF-DAE：即本文方法，只在非功能特征上训练的DAE
  - FF-DAE：在完整特征集上训练的DAE
  - 阈值分析
    - FF-DAE需拟合更多特征，重构误差分布更分散，阈值更高
  - 实验结果分析
    - NF-DAE方法在所有数据集上，均显著降低了假阳性数量
    - 误报率更低

数据集	FF-DAE 阈值	NF-DAE 阈值
UNSW-NB15	0.01568636	0.0025
CICIDS2017	0.000565036	0.0000413
NSL-KDD	0.00253260	0.00109

	NSL-KDD	UNSW-NB15	CICIDS2017
Nos. of normal data samples	13 465	11 083	9773
Nos. of false positives reported by FF-DAE	**1709**	**5137**	**2581**
Nos. of false positives reported by NF-DAE	1344	1976	1268

特点总结与未来展望

特点总结与未来展望

- 算法创新
  - NIDS-DA：首次通过深度自编码器（DAE）检测仅修改非功能特征的功能保留型对抗样本
  - GDA-FS：将GDA与FS结合，在训练与测试两阶段协同防御，突破单阶段防御的局限性
- 算法优势
  - NIDS-DA：与FF-DAE相比显著降低误报率，提高了NIDS的检测准确性
  - GDA-FS：GDA增强了模型的鲁棒性，FS的引入有效减少了对抗噪声的影响
- 未来展望
  - 模型轻量化：使用预训练模型或设计更轻量级的特征提取器提高检测实时性
  - 智能特征工程：自动化识别出最关键的特征子集，而不是依赖手动划分
  - 结合可解释AI（XAI）技术，分析对抗样本的生成机制，优化防御策略

参考文献

[1] Roshan M K, Zafar A. Boosting robustness of network intrusion detection systems: A novel two phase defense strategy against untargeted white-box optimization adversarial attack[J]. Expert Systems with Applications, 2024, 249: 123567.

[2] Kumar V, Kumar K, Singh M, et al. NIDS-DA: Detecting functionally preserved adversarial examples for network intrusion detection system using deep autoencoders[J]. Expert Systems with Applications, 2025, 270: 126513.

道德经

道可道，非常道。名可名，非常名。无名天地之始。有名万物之母。故常无欲以观其妙。常有欲以观其徼。此两者同出而异名，同谓之玄。玄之又玄，众妙之门。

谢谢！