AMDS：攻击感知多阶段防御系统：采用两阶段自适应权重学习的网络入侵检测

Oluseyi Olukola
计算科学与计算机工程学院
南密西西比大学
美国密西西比州哈蒂斯堡
oluseyi.olukola@usm.edu

Nick Rahimi
计算科学与计算机工程学院
南密西西比大学
美国密西西比州哈蒂斯堡
nick.rahimi@usm.edu

arXiv:2603.00859v1 [cs.CR] 1 Mar 2026

摘要——**基于机器学习的网络入侵检测系统容易受到对抗攻击的影响，在基于梯度和分布偏移的威胁模型下，其分类性能都会下降。现有防御通常采用统一的检测策略，可能无法兼顾异质攻击特征。本文提出一种攻击感知多阶段防御框架，通过对集成分歧、预测不确定性和分布异常信号进行加权组合，学习针对特定攻击的检测策略。对七种对抗攻击类型的实证分析揭示了各自不同的检测特征，从而支持构建两阶段自适应检测机制。在基准入侵检测数据集上的实验评估表明，该系统的受试者工作特征曲线下面积达到 94.2%，与对抗训练集成模型相比，分类准确率提高了 4.5 个百分点，F1分数提高了 9.0 个百分点。在具备完整架构知识的自适应白盒攻击下，该系统似乎仍能保持 94.4% 的准确率，攻击成功率为 4.2%；不过，这项评估仅涵盖两种自适应变体，不能构成正式的鲁棒性保证。跨数据集验证进一步表明，防御效果取决于基线分类器能力，并且可能随特征维度而变化。这些结果表明，将针对特定攻击的优化与多信号整合相结合，可以为提升基于机器学习的入侵检测系统的对抗鲁棒性提供一种实用方法。**

索引词—**网络入侵检测、对抗机器学习、集成学习、自适应防御、针对特定攻击的优化**

I. 引言

网络入侵检测系统（NIDS）已成为现代网络安全基础设施的基石，能够在日益复杂的威胁环境中实时检测恶意网络流量。机器学习（ML）和深度学习（DL）技术的采用显著提升了 NIDS 检测已知攻击和零日攻击的能力 [1] ，因为这些技术能够学习复杂的流量模式，超越静态基于特征码的系统所能捕捉的范围 [2]–[4]。这一进展涵盖了多种架构范式，包括用于网络流分析的图神经网络 [5]、基于 Transformer 的序列建模 [6], 、用于大流量攻击检测的深度学习方法 [7], , 以及联邦学习，以支持保护隐私的分布式部署 [8]。尽管取得了这些进展，基于机器学习的 NIDS 仍容易受到对抗攻击的影响；这类攻击通过精心设计的网络流量扰动发起，虽然人类分析人员难以察觉，却能显著降低检测性能 [9]–[12]。

近期研究指出，针对网络入侵检测系统的对抗攻击可分为多种类别 [13], [14]。快速梯度符号法（FGSM）和 PGD 等基于梯度的方法会扰动输入特征以逃避检测，同时保持最小失真 [9], [15], [16]。基于决策的黑盒攻击利用模型输出空间，而不需要梯度信息 [17]–[19]。包括特征注入和流量变形在内的分布偏移攻击会改变网络流的统计属性，以绕过检测阈值 [10], [20], [21]。实证研究表明，基于梯度的攻击对网络入侵检测系统模型的攻击成功率可达 80–95%，而分布偏移攻击在高达 72% 的情况下可能逃过检测 [12], [22], [23]。

针对这些威胁，研究者提出了多种防御范式 [24]。对抗训练在模型训练期间加入对抗样本，能够提高模型对已知攻击的鲁棒性，但往往会牺牲干净样本上的准确率，且无法泛化到未见过的攻击类型 [9], [25], [26]。集成学习结合多个分类器，以利用各模型的互补优势 [27]–[30]；然而，现有集成方法将模型分歧视为噪声，并通过投票机制将其平均消除，因而可能忽略分歧所传递的信号——它可能表明存在对抗操纵。预处理和检测门控，包括基于扩散的输入净化和不确定性感知异常检测，在识别对抗样本方面展现出潜力 [31]–[33], ，但这些方法通常依赖单一检测信号，容易被自适应对手绕过 [34]。

尽管已有进展，现有防御似乎仍存在四个反复出现的局限 [11], [12]：集成方法将分歧视为噪声，而非信号 [27]；当前方法倾向于对异质的攻击类型采用统一的检测策略 [32], [35]；特定攻击防御假设已预先了解攻击类型 [36]；并且系统通常只能提供检测或分类，而不能同时提供二者。为解决这些局限，我们提出AMDS（攻击感知多重

阶段防御系统），将分歧触发的自适应防御、特定攻击权重学习、两阶段检测细化、攻击自适应模型加权和双输出架构整合到统一的防御流程中。

具体而言，本文为网络入侵检测中的对抗防御做出了五项贡献。

1) 我们发现，单一检测信号——分歧、熵或异常——似乎都无法为对抗检测提供足够的判别能力（单项曲线下面积：0.138–0.549）；而通过学习权重组合三种互补信号，可达到 97.1% 的曲线下面积。这一发现表明，采用优化权重进行多信号整合，或许是实现有效对抗防御的必要基础。

2) 通过对六种攻击类型（快速梯度符号法、PGD-L_∞、C&W、同时扰动随机逼近、注入、变形）进行实证分析，我们针对每种攻击类别学习了经优化的专用检测权重；此外，还将 PGD-L_2 作为弱有效性案例进行评估（攻击成功率为 16.6%），以考察在对抗信号极弱时检测器的表现。我们的分析表明，基于梯度的攻击往往可通过分歧主导策略检测（平均 $\beta = 0.452$）；而分布偏移攻击则呈现异质模式（变形：异常主导 $\gamma = 0.533$, 曲线下面积=0.988；注入：熵主导 $\alpha = 0.433$，曲线下面积=0.896）。基于这些模式，我们提出将通用推断与类别级细化相结合的两阶段自适应检测方法，其曲线下面积达到 94.2%（比仅通用方法高+36.3 个百分点，比仅攻击特定方法高 +17.2 个百分点）。类别推断在所评估的攻击中准确率达到 100%（7/7 均正确分类），但未经进一步验证，不应将这一结果外推到更广泛的攻击分类体系。

3）我们提出了攻击自适应模型加权方法，根据推断出的攻击类别动态调整集成模型的权重。在所评估的攻击中，该方法使注入攻击的性能提升了 +30.0 个百分点，分布攻击的平均性能提升了 +17.9 个百分点，总体 F1 分数提升了 +11.0 个百分点。这些提升在分布偏移攻击中尤为显著，而包括对抗训练在内的基线防御对此类攻击几乎无能为力。

4）我们通过双输出架构提供威胁情报，同时实现对抗检测（二分类）和攻击分类（多类别），在干净数据上达到 96.3% 的分类准确率，总体准确率达到 84.8%。据我们所知，现有系统只能提供一个维度的威胁情报；AMDS 旨在同时提供这两者。

5）我们针对经过对抗训练的基线（使用三个随机种子进行 PGD训练）开展了系统评估，发现 AMDS 的准确率和 F1分数分别比对抗训练高出 +4.5 个百分点和 +9.0 个百分点；而与标准集成相比，对抗训练本身带来的收益似乎微乎其微（准确率仅提高 -0.6 个百分点）。在完全了解 AMDS架构的白盒自适应攻击下，该系统仍保持 94.4 % 的准确率，攻击成功率为 4.2%；不过，此评估仅涵盖两种自适应变体，并不构成正式的鲁棒性保证。基于 UNSW-NB15 的跨数据集评估提示了一项部署考量：AMDS 的效果似乎取决于基线模型能力，攻击强度会随特征维度增加而增大，其中 $\sqrt{d}$ 表示攻击强度，d 表示特征维度。

本文其余部分组织如下：第二节回顾相关工作。第三节介绍 AMDS架构。第四节阐述实验方法。第五节报告实验结果，包括组件验证、基线比较、跨数据集泛化和鲁棒性分析。第六节总结全文，并讨论局限性与未来研究方向。

II. 相关工作

针对基于机器学习的网络入侵检测系统（NIDS）的对抗攻击包括 FGSM、PGD 和 C&W 等基于梯度的扰动 [9], [15],[17],基于迁移的黑盒攻击 [18], , 以及在改变特征分布的同时保持协议有效性的问题空间操纵 [10], [37], [38]。包括注入和变形策略在内的分布偏移攻击 [20], [21], 仍然尤其棘手，因为通过对基于梯度的样本进行对抗训练所获得的鲁棒性可能无法有效泛化到这些威胁 [39], [40]。实证研究表明，在白盒和黑盒设置下，针对 NIDS 模型的攻击成功率都很高 [12], [22]。

防御策略大致可分为基于集成的鲁棒性方法和基于检测的方法。集成对抗训练旨在降低不同模型间的可迁移性 [25], , 而 TRADES 则对干净样本准确率与鲁棒性之间的权衡进行建模 [41]。在 NIDS 领域，融合深度学习 [28] 和注意力机制 [29] 的集成架构能够增强对基于梯度的攻击的抵御能力，但通常通过投票或取平均值来汇总预测结果，无形中抑制了分歧，而非将其用作对抗信号。

基于检测的防御旨在利用特征不一致性 [32],重构误差 [42], 内在维度 [35], 不确定性估计 [22], 或净化机制 [31]等单一信号识别对抗样本输入。尽管这些方法在特定威胁模型下有效，但往往依赖固定阈值，且可能容易受到自适应对手的攻击 [34]。动态防御选择框架 [36]引入了自适应路由，但依赖启发式攻击推断，而非学习式优化。

尽管取得了显著进展，现有文献中仍存在三个反复出现的局限。第一，集成不一致通常被视为噪声，而非有价值的对抗信号。第二，大多数防御采用统一策略应对异构攻击类型，因此，与基于梯度的攻击相比，在应对分布偏移威胁时可能鲁棒性较弱。第三，现有系统通常只能提供对抗检测或攻击分类，无法同时实现两者。所提出的框架旨在

通过学习式攻击特定权重优化、覆盖异构威胁的自适应多阶段检测，以及用于集成威胁情报的双输出架构来解决这些局限。

III. 方法

图1展示了 AMDS架构，该架构集成了四个组件：对抗感知集成（第三节）、用于效率感知分流的级联路由器（第三节-A1）、两阶段自适应检测器（第三节-A1），以及攻击自适应模型加权（第三节-A4）。以下各小节将详细介绍每个组件，随后介绍训练和推理算法。

对抗感知集成

对抗感知集成旨在通过模型多样性同时提供基础预测和对抗信号。

A. 集成方法

1) 模型选择与分歧：我们构建了一个由六个分类器组成的集成 $\mathcal{M} = \{M_1, M_2, \ldots, M_6\}$，涵盖基于树的方法、提升法、线性方法和神经网络范式，以最大限度地提高架构多样性（详细配置见第四节）。每个模型 M_i 都会生成一个类别预测 $\hat{}y_i \in \{0, 1, \ldots, C-1\}$，其中 $C = 7$ 个类别，并输出一个概率分布 $\mathbf{p}_i = [p_i^{(0)}, p_i^{(1)}, \ldots, p_i^{(C-1)}]$。

集成预测通过均匀投票整合各个模型的输出：

$$\hat{y}_{\text{ens}} = \arg\max_c \sum_{i=1}^{6} \mathbb{1}[\hat{y}_i = c] \tag{1}$$

模型分歧作为一种对抗信号，计算方式为各类别预测方差的平均值：

$$D(\mathbf{x}) = \frac{1}{C} \sum_{c=0}^{C-1} \text{Var}(\{p_1^{(c)}, p_2^{(c)}, \ldots, p_6^{(c)}\}) \tag{2}$$

其直觉是，对抗扰动可能对不同模型产生不均衡的影响，导致相较于干净输入出现更大的分歧。

级联路由器

级联路由器实现了一个三阶段处理流水线，旨在仅在必要时通过逐渐复杂的检测机制对样本进行路由。

在阶段 1（快速集成筛查）中，预测置信度高（$\max(\mathbf{p}_{\text{ens}}) > 0.85$）且分歧低（$D(\mathbf{x}) < 0.15$）的样本会立即分类，无需进一步处理。在阶段 2（ADS 计算）中，其余样本将触发多信号可疑度评分；若 ADS分数低于检测阈值，则该样本被分类为干净样本。在第 3 阶段（深度验证）中，高度可疑的样本将接受综合分析，包括两阶段自适应检测、攻击自适应加权和置信度校准，然后再进行最终分类。这种分阶段方法旨在避免对大多数流量进行耗费昂贵的 $O(d^2)$ 马氏距离计算；第五节报告了实证路由分布。

攻击驱动的可疑度评分 *(ADS)*

ADS 计算器通过学习得到的权重组合熵、分歧和异常这三种互补信号，以检测对抗性操纵。

2) 多信号检测：我们计算三个检测信号，旨在捕捉对抗行为的不同方面。

熵使用香农熵 [43] 根据集成概率衡量预测不确定性：

$$H(\mathbf{x}) = -\sum_{c=0}^{C-1} p_{\text{ens}}^{(c)} \log p_{\text{ens}}^{(c)} \tag{3}$$

分歧衡量跨模型预测方差，如公式 2 所定义。

异常分数使用马氏距离 [44] 衡量特征空间偏差，该距离基于良性训练分布$\mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$：

$$A(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{4}$$

3) 加权组合与学习：ADS 通过学习得到的权重组合归一化信号 $H,\ D,\ A \in [0, 1]$：

$$\text{ADS}(\mathbf{x}; \mathbf{w}) = \alpha \cdot \hat{H}(\mathbf{x}) + \beta \cdot \hat{D}(\mathbf{x}) + \gamma \cdot \hat{A}(\mathbf{x}) \tag{5}$$

其中 $\mathbf{w} = [\alpha, \beta, \gamma]$，且 $\alpha+\beta+\gamma = 1$。我们学习三组权重。通用权重 $\mathbf{w}^{\text{gen}}$ 经过优化，可对所有攻击类型进行综合对抗检测：

$$\mathbf{w}_{\text{gen}} = \arg\max_{\mathbf{w}} \text{AUC}(\text{ADS}(\mathcal{D}_{\text{clean}}; \mathbf{w}), \text{ADS}(\mathcal{D}_{\text{adv}}; \mathbf{w})) \tag{6}$$

攻击特定权重$\mathbf{w}_k$ 针对每种攻击类型k ，使用限制于 $\mathcal{D}_k$的同一目标进行优化。类别级权重$\mathbf{w}^{\text{gradient}}$ 和$\mathbf{w}_{\text{dist}}$ 随后通过对每个攻击类别（分别为基于梯度的和分布偏移）内的攻击特定权重取平均得到。

4) 两阶段自适应检测：两阶段检测将通用推断与类别特定细化相结合（算法1）。在阶段 1中，系统使用$\mathbf{w}^{\text{gen}}$ 计算通用可疑度评分，并同时根据归一化异常分数推断攻击类别：$\hat{}A(\mathbf{x}) > \tau_{\text{anomaly}} = 0.50$ 的样本被归类为分布偏移攻击，其余样本被归类为基于梯度的攻击。在阶段 2中，系统使用类别特定权重$\mathbf{w}_{\hat{c}}$重新计算可疑度评分，从而产生细化后的检测决策。其依据是，类别级细化可通过利用攻击族内的共享结构，弥补较弱的单个攻击特定权重（在第五节中评估）。

攻击自适应模型加权

攻击自适应加权会根据推断出的攻击类别调整集成模型权重，旨在利用各模型针对不同威胁类型的特定优势。

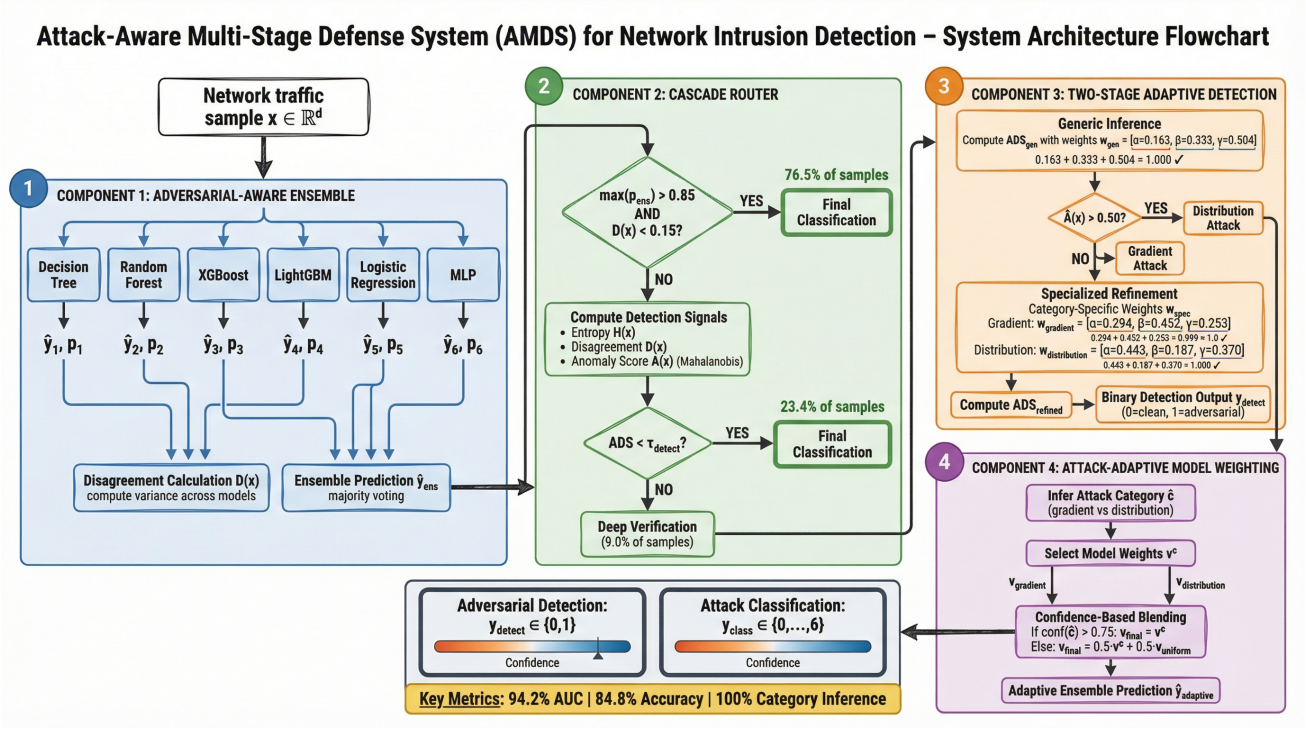

图 1. AMDS 系统架构。网络流量依次经过四个组件处理：带有分歧计算的集成预测、用于效率感知分流的级联路由、带有类别特定细化的两阶段自适应检测，以及攻击自适应模型加权，最终输出对抗检测和攻击分类两项结果。

算法1 两阶段自适应检测

要求： 输入样本 $\mathbf{x}$、通用权重 $\mathbf{w}^{\text{gen}}$、类别权重 $\{\mathbf{w}^{\text{gradient}}, \mathbf{w}_{\text{dist}}\}$、阈值 $\tau_{\text{anomaly}} = 0.50$

确保： 二元检测 y_{detect}、推断类别 $\hat{c}$

// **阶段 1：通用推断和类别分类-tion**

计算信号：$H(\mathbf{x})$ $D(\mathbf{x})$ $A(\mathbf{x})$

归一化：$\hat{H}, \hat{D}, \hat{A} \leftarrow$ 最小-最大归一化至 $[0,1]$

$\text{ADS}_{\text{gen}} \leftarrow \mathbf{w}_{\text{gen}}^T \cdot [\hat{H}, \hat{D}, \hat{A}]$

// **根据异常分数推断攻击类别**

如果 $A(\mathbf{x}) > \tau_{\text{anomaly}}$ **然后**

  $\hat{c} \leftarrow$ 分布

否则

  $\hat{c} \leftarrow$ 梯度

结束如果

// **阶段 2：专门化细化**

选择权重：$\mathbf{w}_{\text{spec}} \leftarrow \mathbf{w}_{\hat{c}}$

$\text{ADS}_{\text{refined}} \leftarrow \mathbf{w}^T{}_{\text{spec}} \cdot [\hat{H}, \hat{D}, \hat{A}]$

// **最终检测**

若 $\text{ADS}_{\text{refined}} > \tau_{\text{detect}}$ **则**

  $y_{\text{detect}} \leftarrow 1$ *(对抗样本)*

否则

  $y_{\text{detect}} \leftarrow 0$ （干净样本）

结束 if

返回 $y_{\text{detect}}, \hat{c}$

5）权重学习：对于每种攻击类别 $c \in${梯度，分布}，我们学习模型权重 $\mathbf{v}_c = [v_1^{(c)}, v_2^{(c)}, \ldots, v_6^{(c)}]$ 通过基于性能的优化：

$$v_i^{(c)} = \frac{(\text{Acc}_i^{(c)})^3}{\sum_{j=1}^{6} (\text{Acc}_j^{(c)})^3} \tag{7}$$

其中，$\text{Acc}_i^{(c)}$ 是模型 i在攻击类别 c上的准确率。三次指数旨在生成尖锐的权重分布，使每个类别中表现优异的模型获得更高权重。

6）基于置信度的混合：为避免在置信度较低时过度依赖推断类别，我们将自适应权重与均匀权重$\mathbf{v}_{\text{uniform}} = [1/6, \ldots, 1/6]$进行混合。置信度根据异常分数与类别阈值之间的差值计算：$\text{conf}(\hat{c}) = \hat{}\min(1, |A(\mathbf{x}) - \tau_{\text{anomaly}}|/0.12)$。当 $\text{conf}(\hat{c}) < 0.75$时，我们应用$\mathbf{v}_{\text{final}} = 0.5 \cdot \mathbf{v}_{\hat{c}} + 0.5 \cdot \mathbf{v}_{\text{uniform}}$；否则，应用$\mathbf{v}_{\text{final}} = \mathbf{v}_{\hat{c}}$。

然后，自适应集成预测为：

$$\hat{y}_{\text{adaptive}} = \arg\max_{c} \sum_{i=1}^{6} v_i^{(\hat{c})} \cdot p_i^{(c)} \tag{8}$$

双输出架构

AMDS 产生两种互补输出：对抗检测（$y_{\text{detect}} \in \{0,1\}$），用于判断输入是否疑似经过篡改，置信度为 $\text{confidence}_{\text{detect}} = \min(1, \text{ADS}_{\text{refined}}/\tau_{\text{detect}})$；以及攻击分类（$y_{\text{class}} \in \{0, \ldots, C-1\}$），用于识别流量类型，置信度为 $\text{confidence}_{\text{class}} = \max(\mathbf{p}_{\text{ens}})$。

算法 2AMDS 训练流程

要求： 干净训练数据 $\mathcal{D}_{\text{train}}$，验证数据 $\mathcal{D}_{\text{val}}$

确保： 训练好的模型 $\mathcal{M}$，学习得到的权重 $\{\mathbf{w}^{\text{gen}}, \mathbf{w}_k, \mathbf{v}_c\}$，阈值 $\{\tau_*\}$

// **阶段 1：基线模型训练**

对于每种模型类型 $i \in \{1, \ldots, 6\}$ **执行**

在 $\mathcal{D}_{\text{train}}$ 上训练 M_i，以提高分类准确率

在 $\mathcal{D}_{\text{val}}$ 上验证 M_i（目标：>95% 准确率）

结束循环

// **阶段 2：对抗样本生成**

生成梯度攻击：$\mathcal{D}_{\text{FGSM}}, \mathcal{D}_{\text{PGD}}, \mathcal{D}_{\text{CW}}, \mathcal{D}_{\text{SPSA}}$

生成分布攻击：$\mathcal{D}_{\text{Injection}}, \mathcal{D}_{\text{Morphing}}$

合并：$\mathcal{D}_{\text{adv}} = \bigcup_k \mathcal{D}_k$

验证攻击有效性（目标：攻击成功率 >50%）

// **第 3 阶段：权重学习**

学习通用权重：$\mathbf{w}_{\text{gen}} \leftarrow \arg\max_{\mathbf{w}} \text{曲线下面积}(\mathcal{D}_{\text{val}}, \mathcal{D}_{\text{adv}}; \mathbf{w})$

对每种攻击类型 k **执行**

学习攻击特定：$\mathbf{w}_k \leftarrow \arg\max_{\mathbf{w}} \text{曲线下面积}(\mathcal{D}_{\text{val}}, \mathcal{D}_k; \mathbf{w})$

结束循环

类别内平均：$\mathbf{w}^{\text{gradient}}$，$\mathbf{w}_{\text{dist}}$

对每个类别 $c \in \{$梯度，分布$\}$ **执行**

学习模型权重：$v_i^{(c)} \leftarrow (\text{准确率}_i^{(c)})^3 / \sum_j (\text{准确率}_j^{(c)})^3$

结束循环

// **第4阶段：阈值调优**

针对 $\mathcal{D}_{\text{val}}$调优 τ_{detect}，使假阳性率达到 10%

调整 τ_{conf}，τ_{disagree} 以实现级联路由

调整τ_{anomaly} 以获得最佳类别推断准确率

返回 $\mathcal{M}, \{\mathbf{w}^{\text{gen}}, \mathbf{w}_k, \mathbf{v}_c\}, \{\tau_*\}$

算法3 AMDS 推理流程 要求：输入样本 x、训练好的模型 $\mathcal{M}$、权重 $\{\mathbf{w}_*、\ \mathbf{v}_*\}$、阈值 $\{\tau_*\}$ 确保：对抗检测 y_{detect}、攻击分类y_{class}、置信度 1: // 阶段 1：集成预测 2: 对每个模型 $M_i \in \mathcal{M}$ 执行 3: $\hat{}y_i, \mathbf{p}_i \leftarrow M_i(\mathbf{x})$ 4: 结束循环 5: $\hat{}y_{\text{ens}} \leftarrow$ 对 $\{\hat{}y_1, \ldots, \hat{}y_6\}$进行多数投票 6: $D(\mathbf{x}) \leftarrow$ 计算分歧（公式 2） 7: // 级联路由：阶段 1 检查 8: 如果 $\max(\mathbf{p}_{\text{ens}}) > \tau_{\text{conf}}$ 且 $D(\mathbf{x}) < \tau_{\text{disagree}}$ 则 9: 返回 $y_{\text{detect}} = 0$, $y_{\text{class}} = \hat{}y_{\text{ens}}$ // 快速路径 10: 结束条件 11: // 阶段 2：ADS 计算与两阶段检测 12: $y_{\text{detect}}, \hat{c} \leftarrow$ TwoStageDetection(x) // 算法1 13: // 级联路由：阶段 2 检查 14: 如果 $y_{\text{detect}} = 0$ 则 15: 返回 $y_{\text{detect}} = 0$, $y_{\text{class}} = \hat{}y_{\text{ens}}$16: 结束条件 17: // 第 3 阶段：通过攻击自适应加权进行深度验证 ˆ 18: 计算置信度：$\text{conf}(\hat{c}) \leftarrow \min(1, |A(\mathbf{x}) - \tau_{\text{anomaly}}|/0.12)$ 19: 如果 $\text{conf}(\hat{c}) > 0.75$ 则 20: $\mathbf{v}_{\text{final}} \leftarrow \mathbf{v}_{\hat{c}}$21: 否则 22: $\mathbf{v}_{\text{final}} \leftarrow 0.5 \cdot \mathbf{v}_{\hat{c}} + 0.5 \cdot \mathbf{v}_{\text{uniform}}$23: 结束条件 24: $\hat{}y_{\text{adaptive}} \leftarrow \arg\max_c \sum_{i=1}^{6} v_i^{(\hat{c})} \cdot p_i^{(c)}$ 25: // 双重输出 26: $y_{\text{class}} \leftarrow \hat{}y_{\text{adaptive}}$27: 置信度$_{\text{detect}} \leftarrow \min(1, \text{ADS}_{\text{refined}}/\tau_{\text{detect}})$ 28: 置信度$_{\text{class}} \leftarrow \max(\mathbf{p}_{\text{ens}})$ 29: 返回 y_{detect}, y_{class}, 置信度$_{\text{detect}}$, 置信度$_{\text{class}}$

训练与推理流程

7) 训练流程：算法 2 描述了 AMDS 的训练过程，分为四个阶段：(1) 在干净数据上独立训练六个集成模型（目标：准确率达到 >95%）；(2) 针对基于梯度的攻击（快速梯度符号法、PGD、C&W、同时扰动随机逼近）和分布偏移攻击（注入、变形）生成对抗样本，并在 >50% 攻击成功率下进行验证；(3) 通过优化曲线下面积，在通用、攻击特定和类别层面学习检测权重；(4) 在验证数据上调整检测阈值和路由阈值（τ_{detect}已针对 10% 假阳性率进行校准）。

8) 推理流程：算法 3 描述了整合全部四个组件的完整推理过程。最坏情况下的计算复杂度为 $O(6 \cdot C_{\text{model}} + d^2)$，其中 C_{model} 是单个模型的推理开销，而 d^2 来自马氏距离计算。级联路由器旨在降低平均情况下的复杂度，通过在阶段 1 解决大多数样本，避免产生 $O(d^2)$ 的开销。

IV. 实验评估

本节介绍用于评估 AMDS 的数据集、攻击生成、基线方法和评估协议。

数据集

我们在两个基准数据集上评估 AMDS。**CSE-CIC-IDS 2018** [45]作为主要评估数据集，包含涵盖 7 个类别（良性、僵尸网络、暴力破解、分布式拒绝服务攻击、拒绝服务攻击、渗透、Web攻击）的真实网络流量，以及 77 个流级特征。我们将数据划分为 50,000 个训练样本、5,000 个验证样本，以及一个留出的测试集，其中包含 5,000 个良性样本和每种攻击类型 1,000 个对抗样本（C&W 由于计算成本较高为 500 个），全程采用分层抽样。**UNSW-NB15** [46] 用于跨数据集评估，包含 190 个特征和 9 个攻击类别，并映射为与 CSE-CIC-IDS2018 对齐的 5 个类别。显著更高的维度（2.47×）支持研究攻击有效性如何随特征空间扩展而变化。在两个数据集上均采用相同的实验方案。

数据预处理

采用标准预处理：删除重复项、填补缺失值，并进行零均值单位方差缩放

仅基于训练数据拟合，以防止数据泄漏。马氏距离异常分数（公式 4）所需的良性分布参数 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 根据干净训练流量计算得出。

攻击生成

我们使用对抗鲁棒性工具箱（ART）v1.13 [47] 生成七种对抗攻击，分为两类：基于梯度的攻击和分布偏移攻击，前者使用该工具箱，后者使用 NumPy。

基于梯度的攻击（五种）：快速梯度符号法 FGSM [15]（ϵ=0.02，单步）；PGD-L_∞ [9]（ϵ=0.02，20 步，α=0.005）；PGD-L_2 [9]（ϵ=0.02，20 步，受 L_2 约束）；C&W-L_2 [17]（20 次二分搜索迭代，κ=0，500 个样本）；以及同时扰动随机逼近 SPSA（100 次查询，δ=0.01，黑盒梯度估计）。

分布偏移攻击（两种）：注入会为每个特征添加高斯噪声 $\mathcal{N}(0, 0.05 \cdot \sigma_i)$，而变形会施加系统性偏移 $x_i' = x_i + 0.05 \cdot \sigma_i$，其中 σ_i 为训练标准差。

所有攻击在纳入前均经过有效性验证（目标：攻击成功率>50%）。由于在 77 维空间中的 ϵ 相对较小，PGD-L_2 在 CSE-CIC-IDS2018 上的攻击成功率较低，为 16.6%；我们仍保留该攻击，以考察检测器在弱攻击条件下的表现。

基线

该集成包含六种分类器：决策树、随机森林、 [48], XGBoost [49], LightGBM、逻辑回归和多层感知机，涵盖基于树的方法、提升法、线性方法和神经网络范式。所有模型均采用标准超参数，并经验证可达到 >95% 的干净准确率；详细配置见补充材料。[1]

我们在相同的训练数据和评估协议下，将 AMDS 与四种基线进行比较：(1) 标准集成，采用均匀投票；(2) 对抗训练（*AT*）集成，使用 PGD 对抗训练重新训练 [9]（50% 干净样本、50% 对抗样本；ϵ=0.02，10 步），对三个随机种子的结果取平均，并计算 95% 自助法置信区间；(3) 单一最佳模型（XGBoost）；以及 (4) 均匀 *ADS*，即仅使用通用权重的 AMDS。

评估指标

我们报告曲线下面积作为主要检测指标（与阈值无关），并提供准确率、F1分数以及在运行阈值（校准至 10% 假阳性率）下按攻击类型划分的细分结果。系统效率通过单样本时延和级联路由分布进行衡量。所有指标均包含 95% 自助法置信区间（1,000 次迭代，采用百分位数法）。

[1]决策树（深度 10）、随机森林（100 棵树）、XGBoost（100 个估计器，学习率 0.1）、LightGBM（100 个叶子，学习率 0.1）、逻辑回归（C=1.0，L2）、MLP（层数：100、50；ReLU）。

实验方案

评估分为三个阶段：（1）组件验证，分别评估单独检测信号、级联路由、两阶段检测和攻击自适应加权；（2）在留出数据（每种攻击类型 = 5 包含 2,000 个干净样本 + 500 和对抗样本，共 500 个样本）上对所有基线进行完整系统评估，包括在 UNSW-NB15 上进行跨数据集评估；以及（3）通过移除单个组件来量化其贡献的消融研究。[2]

V. 结果

我们分三部分报告结果：CSE-CIC-IDS2018 上的主要评估、在 UNSW-NB15 上的跨数据集泛化，以及鲁棒性分析（包括自适应对手评估）。

A. CSE-CIC-IDS2018 上的主要评估

1）攻击特定权重模式：表 I 展示了六种攻击类型下熵（α）、分歧（β）和异常（γ）的学习所得攻击特定权重。由于 PGD-L_2 的攻击有效性较弱（攻击成功率为 = 16.6%），会产生近乎随机的检测权重（曲线下面积为 = 0.502），因此未纳入其中；其两阶段结果见表 II。

基于梯度的攻击始终呈现分歧主导模式（β 范围为 0.434 至 0.528），而分布偏移攻击则表现出异质特征：注入呈熵主导特征（$\alpha = 0.433$），变形呈异常主导特征（$\gamma = 0.533$）。这些模式表明，不同攻击类别可能需要专门的检测策略。在所有攻击上进行优化的通用权重得出$\alpha = 0.163$、$\beta = 0.333$、$\gamma = 0.504$，曲线下面积为 97.1%。类别级权重得出：梯度（α = 0.294、β = 0.452、γ = 0.253，曲线下面积为 = 0.702）和分布（α = 0.443、$\beta = 0.187$、$\gamma = 0.370$，曲线下面积为 = 0.942）。

2）两阶段检测效果：表 II 对比了针对全部七种攻击类型的三种检测策略。

两阶段检测的平均曲线下面积为 0.942，优于仅通用策略（0.579，+0.363）和仅攻击特定策略（0.771，+0.172）。PGD-L_2 的提升幅度最大（相较于仅特定策略提升 +0.383），其单独检测效果从接近随机水平（曲线下面积 = 0.502）提升至较强的类别级检测水平（曲线下面积 = 0.885），这表明类别级细化能够弥补单独攻击权重较弱的问题。

类别推断在全部七种攻击上均达到 100% 的准确率。分布攻击产生的平均异常分数为 0.68–0.85，而梯度攻击为 0.15–0.25，因此可以进行基于阈值的推断（$\tau_{\text{anomaly}} = 0.50$）。虽然在所评估的攻击上具有这种明显的可分性，但未经进一步验证，不应假定其能够推广到所有可能的攻击类型。

3) 级联计算特性：表 III 对比了完整 AMDS 处理与级联路由。

[2]实现环境：Python 3.9、scikit-learn 1.2、XGBoost 1.7、PyTorch 2.0；英特尔至强 E5-2680 v4 、 128 GB 内存、 NVIDIA V100 GPU。通过 SciPy SLSQP 进行权重优化。为确保结果可复现，所有随机种子均固定为 42。论文录用后将公开代码和模型。

表 I 攻击特定的学习权重与检测性能

攻击	类别	α (熵)	β (不一致)	γ (异常)	AUC	主导信号
FGSM	梯度	0.260	**0.528**	0.212	0.743	分歧
投影梯度下降法（L_∞）	梯度	0.269	**0.465**	0.267	0.759	分歧
C&W	梯度	0.306	**0.434**	0.261	0.748	分歧
SPSA	梯度	0.293	**0.448**	0.259	0.758	分歧
注入	分布	**0.433**	0.360	0.207	0.896	熵
变形	分布	0.454	0.013	**0.533**	0.988	异常
梯度平均值		0.282	**0.469**	0.250	0.752	**不确定性**
分布平均值		0.444	0.187	0.370	0.942	**异质**

基于梯度的攻击呈现分歧主导模式（$\beta \approx 0.47$）。分布攻击呈现异质模式：变形是异常主导（$\gamma = 0.533$，曲线下面积 = 0.988），而注入是熵主导（$\alpha = 0.433$，曲线下面积 = 0.896）。由于攻击有效性较弱（攻击成功率为 16.6%），不纳入 PGD-L_2；其近随机权重（曲线下面积 = 0.502）可能反映的是扰动强度不足，而非检测局限性。

表 II 两阶段检测比较（曲线下面积）

攻击	仅通用	攻击特定	两阶段	Δ对比通用	Δ对比特定
FGSM	0.545	0.743	**0.943**	+0.398	+0.200
投影梯度下降法（L_∞）	0.552	0.759	**0.943**	+0.391	+0.184
投影梯度下降法（L_∞）$_2$	0.218	0.502	**0.885**	+0.666	+0.383
C&W	0.565	0.748	**0.945**	+0.379	+0.197
SPSA	0.551	0.758	**0.943**	+0.392	+0.185
注入	0.652	0.896	**0.966**	+0.313	+0.070
变形	0.967	0.988	**0.971**	+0.004	−0.017
平均值（7种攻击）	**0.579**	**0.771**	**0.942**	**+0.363**	**+0.172**

仅通用：通用权重（$\alpha = 0.163, \beta = 0.333, \gamma = 0.504$）。仅攻击特定：每种攻击的单独权重（七组权重）。两阶段：通用筛选 + 类别级细化（两组权重）。PGD-L_2 从两阶段检测中获益最大（+0.383，相较于仅特定），这表明类别级细化或可弥补单独攻击权重较弱的问题。

表 III 级联路由器性能与效率

指标	基线（完整 AMDS）	级联（自适应）	比率
吞吐量（样本/秒）	6,172	4,989	0.81×
延迟（毫秒/样本）	0.16	0.20	+0.04
总体准确率	84.9%	84.9%	0.0 百分点
干净准确率	96.3%	96.3%	0.0 百分点
路由分布：			
阶段 1（快速筛查）	100%	76.5%	−23.5%
阶段 2（ADS 计算）	100%	23.4%	−76.6%
第 3 阶段（深度验证）	100%	9.0%	−91.0%
计算操作：			
马氏距离计算	100%	23.5%	−76.5%
攻击自适应加权	100%	9.0%	−91%

级联方案在阶段 1 实现了 76.5% 的路由率，且准确率没有下降。吞吐量开销（0.81×）源于路由逻辑；马氏距离计算减少了 76.5%，自适应加权减少了 91%。

级联路由在阶段 1 处理了 76.5% 的样本，23.4% 进入阶段 2，另有 9.0% 进入阶段 3（深度验证）；两种配置的准确率相同（总体为 84.9%，干净样本为 96.3%）。0.81× 吞吐量比率反映了路由逻辑带来的开销，尽管马氏距离计算减少了 76.5%，攻击自适应加权减少了 91%。因此，级联的优势在于重新分配计算量，而非提升原始吞吐量：在昂贵的矩阵运算占据主要成本的部署场景中，这种权衡可能有益；而吞吐量至关重要的环境则可能更适合非级联配置。

4) 基线比较与攻击自适应加权：表 IV 展示了主要的基线比较。图2展示了各类攻击的详细结果。

AMDS 的总体准确率达到 84.8%（比标准训练高 +3.9 个百分点，比对抗训练高 +4.5 个百分点），F1分数达到 71.7%（比对抗训练高 +9.0 个百分点）。与标准训练相比，对抗训练似乎仅带来微乎其微的收益（准确率提高 −0.6 个百分点，F1分数提高 +0.3 个百分点）。在梯度攻击上，AMDS 的表现略有下降（与对抗训练相比，平均低 −1.0 个百分点）；鉴于基线方法已能很好地应对这些攻击（平均准确率为 92.0–92.5%），这一结果或许并不意外。在分布攻击上，AMDS 取得了显著提升（比对抗训练高 +18.4 个百分点）；对于注入攻击，准确率提高了 +30.4 个百分点，F1分数提高了 +40.6 个百分点。这些结果表明，对抗训练可能不足以应对分布偏移威胁，而针对特定攻击的优化则能带来实质性收益。

对于梯度和分布类别，类别推断的精确率、召回率和 F1 分数均达到完美水平（ 1.000 ），从而能够可靠地选择类别特定模型权重。干净准确率略有下降（96.3%，而标准方法为 98.3%，对抗训练为 99.3%），但考虑到总体 F1 分数提高了 +9.0 个百分点，这一权衡似乎是有利的。

5) 组件消融分析：表 V 展示了消融结果。

所有三个组件均发挥了显著作用：两阶段检测 +3.5 个百分点，攻击自适应加权 +3.8 个百分点，攻击特定权重 +3.7 个百分点。F1分数方面的贡献更大（+9.8 个百分点至 +12.0 个百分点），这表明它们对精确率-召回率平衡的影响大于对原始准确率的影响。消融配置显示，干净准确率有所提高（98.1–99.5%，而基线为 96.3%），揭示了明确的权衡：AMDS 牺牲 1.8–3.2 个百分点的干净准确率，以换取对抗条件下 3.5–3.8 个百分点的总体准确率提升，以及 9.8–12.0 个百分点的 F1分数提升。

表IV CSE-CIC-IDS2018 综合基线比较

攻击	Cat.	标准	AT	AMDS	Δ与标准方法比较	Δ与对抗训练比较	标准方法 F1 分数	对抗训练 F1 分数	AMDS F1 分数
基于梯度的攻击									
FGSM	G	95.9%	95.3%	94.7%	−1.2 个百分点	−0.6 个百分点	88.3%	87.1%	87.4%
投影梯度下降法（L_∞）$_\infty$	G	95.9%	96.5%	94.2%	−1.7 个百分点	−2.3 个百分点	88.0%	89.4%	86.1%
投影梯度下降法（L_∞）$_2$	G	82.6%	84.1%	79.4%	−3.2 百分点	−4.7 百分点	63.0%	67.8%	55.3%
C&W	G	92.6%	92.6%	91.6%	−1.0 个百分点	−1.0 个百分点	79.3%	79.5%	80.3%
SPSA	G	95.7%	91.4%	95.0%	−0.7 个百分点	+3.6 个百分点	87.6%	81.1%	87.0%
梯度平均值		92.5%	92.0%	91.0%	−1.6 百分点	−1.0 百分点	81.2%	81.0%	79.2%
分布偏移攻击									
注入	D	54.5%	53.4%	**83.8%**	**+29.3 个百分点**	**+30.4 百分点**	19.4%	22.9%	**63.5%**
变形	D	49.1%	48.9%	**55.2%**	**+6.1 百分点**	**+6.3 百分点**	11.0%	10.9%	**42.7%**
分布平均值		51.8%	51.2%	**69.5%**	**+17.7 百分点**	**+18.4 个百分点**	15.2%	16.9%	**53.1%**
总体		80.9%	80.3%	**84.8%** (74.1–92.5%)	**+3.9 百分点**	**+4.5 百分点**	62.4%	62.7%	**71.7%** (60.9–83.1%)
干净样本		98.3%	99.3%	96.3% (95.3–97.0%)	−2.0 个百分点	−3.0 个百分点	49.6%	33.2%	32.7%

括号内为 95% 自助法置信区间（$n = 1{,}000$）。总体置信区间较宽（74.1–92.5%），反映了不同攻击类型之间的异质性。使用 PGD 进行对抗训练（$\epsilon = 0.02$，三个随机种子）。PGD-L_2 回归反映了攻击成功率较低（16.6%），为完整起见予以列出。

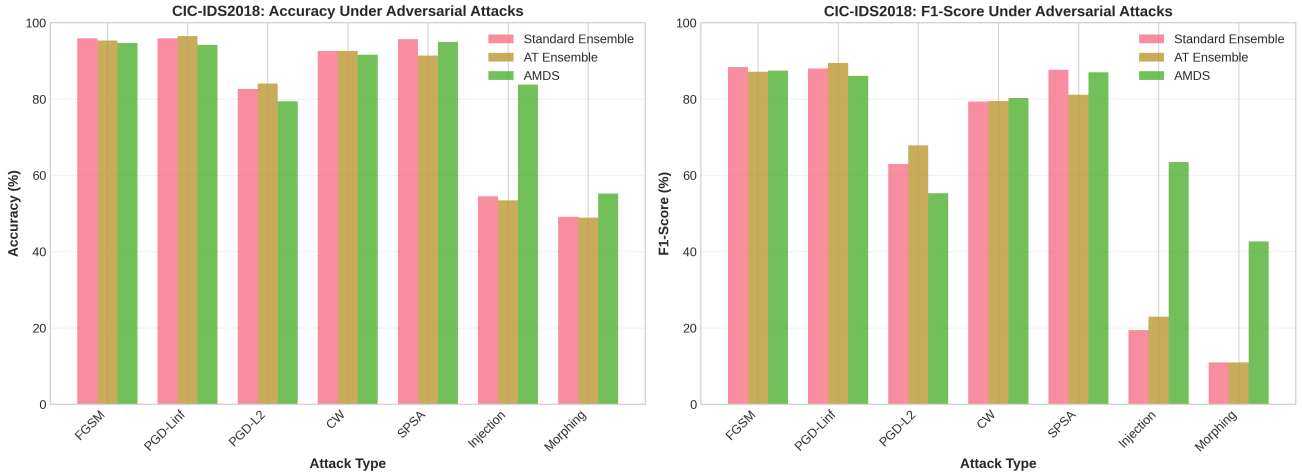

图 2. CSE-CIC-IDS2018 上的逐攻击性能比较。AMDS 在分布偏移攻击上取得显著提升（相较于标准方法，注入提升 +29.3 个百分点，变形提升 +6.1 个百分点），但在梯度攻击上略有退步（平均 −1.6 个百分点）。与标准训练相比，对抗训练似乎益处甚微。

表 V 消融研究：组件贡献分析

配置	总体准确率	总体 F1	干净样本准确率
完整 AMDS	**84.9%**	**73.0%**	**96.3%**
消融配置：			
仅通用权重	81.4%	63.2%	98.1%
组件贡献	−3.5 百分点	−9.8 百分点	+1.8 百分点
模型均匀投票	81.1%	61.0%	99.5%
组件贡献	−3.8 个百分点	−12.0 个百分点	+3.2 个百分点
无攻击特定权重	81.2%	62.1%	98.1%
组件贡献	−3.7 百分点	−10.9 百分点	+1.8 百分点

所有组件使准确率提升了 $\geq$3 个百分点，F1 分数提升了 $\geq$9.8 个百分点。消融配置的干净准确率有所提高（98.1–99.5%），这表明干净样本性能与对抗鲁棒性之间存在权衡。测试集：2,000 个干净样本，每种攻击包含 = 5,500 个样本。

B. 跨数据集泛化（*UNSW-NB15*）

表 VI 展示了跨数据集评估结果。

UNSW-NB15 的基线准确率（52.9%）显著低于 CSE-CIC-IDS2018（80.7%），而两个数据集上的干净准确率仍然很高（>96%），这表明基线性能下降源于攻击强度，而非模型能力不足。AMDS 在 UNSW 上实现了 −0.2 个百分点，而在 CIC 上为 +4.2 个百分点，这表明，当攻击压倒基线模型时，自适应防御的改进空间可能有限。这种性能差距似乎可归因于依赖维度的攻击缩放：在相同的$\epsilon = 0.02$下，UNSW-NB15 的 190 个特征可能会受到比 CIC 的 77 个特征更强的有效扰动（$\sqrt{d}$ 缩放），强度为 $\sim$1.57×。使用较弱攻击（$\epsilon = 0.01$）进行验证后，正向提升有所恢复（准确率提高 +0.5 个百分点，F1 分数提高 +1.9 个百分点），与这一解释一致。这些结果指出了一项实际部署考量：AMDS 的有效性似乎取决于基线模型能力；在高维特征空间（>150 个特征）中部署的从业者应考虑按比例调整扰动预算，或采用降维。

C. 鲁棒性分析

1) 自适应对手评估：表 VII 展示了在白盒知识条件下针对

表 VI 跨数据集泛化：CSE-CIC-IDS2018 与 UNSW-NB15

指标	CSE-CIC-IDS2018	UNSW-NB15	Δ	说明
数据集特征				
特征	77	190	+113	2.47× 个更多维度
攻击预算（ϵ）	0.02	0.02	0.0	相同的名义预算
有效攻击强度	1.0×	$\sim$1.57×	+0.57×	随 $\sqrt{d}$ 扩展
基线性能				
基线准确率	80.7%	52.9%	−27.8 个百分点	攻击强度超过基线能力
干净准确率	98.1%	98.5%	+0.4 个百分点	模型在干净数据上运行
AMDS 性能				
AMDS 准确率	84.9%	52.7%	−32.2 个百分点	受基线性能下降的限制
AMDS 宏平均 F1	73.0%	37.6%	−35.4 个百分点	基线地板效应
Δ 相对于基线的准确率	+4.2 百分点	−0.2 百分点	−4.4 百分点	基线可能过弱
Δ F1 分数 vs 基线	+11.0 百分点	+0.7 个百分点	−10.3 个百分点	最小改进
验证：较弱攻击（$\epsilon = 0.01$）				
Δ 准确率（$\epsilon = 0.01$）	–	+0.5 个百分点	–	部分恢复改进
Δ F1（$\epsilon = 0.01$）	–	+1.9 个百分点	–	F1 分数提升更明显

有效攻击强度似乎按 $\sqrt{d}$ 的比例缩放：$\sqrt{190}/\sqrt{77} \approx 1.57\times$，因此，在 UNSW 上的 $\epsilon = 0.02$ 可能与在 CIC 上的 $\epsilon \approx 0.031$ 实际上等效。在较弱攻击（$\epsilon = 0.01$）下，AMDS 恢复了正向提升（准确率提高 +0.5 个百分点，F1 分数提高 +1.9 个百分点），但基线准确率仍然较低（51.5%）。

表 VII 针对自适应对手的鲁棒性

攻击类型	标准集成	AMDS 准确率	检测比率 (ADS)
标准攻击（非自适应）			
FGSM	95.1%	94.6%	2.4%
PGD-L_∞	95.5%	93.4%	1.8%
投影梯度下降法（$L\infty$）$_2$	83.4%	80.5%	14.0%
自适应攻击（感知 *ADS*）			
基线自适应	97.9%	96.1%	5.0%
改进的自适应	95.8%	**94.4%**	3.1%
性能下降	–	−1.7 个百分点	–
攻击成功率	–	**4.2%**	–

自适应攻击在充分白盒知识下，最小化 ADS分数，同时维持错误分类。基线自适应：从良性样本起点开始，进行随机方向搜索。改进的自适应：从快速梯度符号法起步，进行梯度近似。两种变体的 ADS分数均降低了 0.0%。

AMDS架构。

我们评估了两种自适应变体：基线自适应（从良性样本出发，采用随机方向搜索）和改进的自适应（从 FGSM 扰动样本出发，采用梯度近似）。在改进的自适应攻击下，AMDS 仍保持 94.4% 的准确率（相较于基线自适应，-1.7 个百分点），攻击成功率为 4.2%。这两种变体均未实现 ADS 分数降低（0.0%），这表明，在维持错误分类的同时最小化熵、分歧和异常，可能会导致优化目标相互冲突。改进的变体利用 FGSM 扰动的初始点，将攻击成功率提高至 4.2%（相比之下为 2.1%），但这一比率仍远低于标准攻击的典型攻击成功率（>50%）。需要指出的是，此评估仅涵盖两种自适应变体，并不构成正式的鲁棒性保证；足够老练的对手可能会找到本文未探讨的策略。

2) 统计显著性： 所有指标均包含 95% 的自助法置信区间（1,000 次迭代，百分位数法）。总体准确率为 85.01% [74.09%、92.46%], ; 宏平均F1值为 73.08% [60.86%、83.06%], ; 干净准确率为 96.25%[95.32%、97.00%]。总体置信区间较宽，反映了不同攻击类型之间的异质性；配对自助法检验表明，相较于两种基线的改进在 $p < 0.05$ 处具有统计显著性。

VI. 结论

本文提出了 AMDS，这是一种用于网络入侵检测的多阶段对抗防御系统，能够学习针对特定攻击的检测策略，而不是采用统一的防御措施。在 CSE-CIC-IDS2018 上进行评估、在 UNSW-NB15 上进行跨数据集验证，以及针对自适应对手开展鲁棒性测试的结果表明，该系统具有实际有效性，同时也明确了其部署边界。

AMDS 的准确率比对抗训练集成模型高出+4.5 个百分点，F1分数高出 +9.0 个百分点；而对抗训练本身相较于标准集成（准确率高出 −0.6 个百分点）似乎几乎没有带来收益。学习到的权重模式表明，梯度攻击往往呈现以分歧为主导的特征（$\beta \approx 0.47$），而分布攻击则表现出异质模式——变形攻击以异常为主导，注入攻击以熵为主导。两阶段自适应检测通过采用中间层次的类别级粒度，在七种攻击类型上实现了 94.2% 的曲线下面积；这种方式似乎既避免了通用检测过度平滑的问题，也避免了单个攻击特定检测的过拟合倾向。在明确最小化检测分数的白盒自适应攻击下，AMDS 仍保持 94.4% 的准确率，攻击成功率为 4.2%；不过，这项评估并不构成正式的鲁棒性保证。

在 UNSW-NB15 上进行的跨数据集评估表明，AMDS 的有效性可能取决于基线模型的能力。攻击强度似乎会随特征维度增加而提升（$\propto \sqrt{d}$），因此在相同的预算下，UNSW-NB15 的 190 维特征空间受到的扰动实际上比 CSE-CIC-IDS2018 的 77 维特征空间强 $\sim$1.57×。当这种缩放效应使基线模型的性能趋近随机猜测水平时，包括 AMDS 在内的自适应防御可能没有太多提升空间。

需要指出若干局限。首先，AMDS 在权重学习期间需要使用对抗样本，并假设对可能的攻击类型有所了解；对于完全未见过的攻击族，性能会如何逐渐下降，仍是一个悬而未决的问题。其次，二元类别推断（梯度与分布）可能无法涵盖现实

世界范围内对抗性多样性的完整范围，而此处观察到的 100 % 推理准确率不应外推至更细粒度的分类体系，因为其中的类间边界可能不那么清晰。第三，部署期间学习得到的权重保持静态；尽管自适应对手评估表明系统能够抵御明确的ADS最小化，但持续性足够强的对手仍可能随着时间推移设计出专门的规避策略。第四，较宽的自助法置信区间（总体准确率 95% CI：74.1–92.5%）反映了不同攻击类型之间显著的性能异质性；要全面了解系统能力，应结合各攻击类型的细分结果（表IV）与汇总统计数据进行考量。

这些局限性指明了未来研究的几个方向。在线权重自适应或可解决静态权重易受攻击的弱点，但也带来了一个挑战：如何区分真正的攻击演化与对学习信号的对抗性操纵。将攻击分类体系扩展到二元分类之外，或可实现更精细的优化，但需通过实证验证更高粒度下的推理准确率。将其迁移到其他对抗性机器学习领域，如恶意软件检测、欺诈检测和自主感知，可检验网络入侵检测系统中观察到的类别级可分性是否具有更广泛的适用性。最后，与认证鲁棒性技术相结合，可在需要可证明界限的场景中，以形式化保证补充 AMDS 的实证有效性。

总体而言，这些发现表明，在网络入侵检测中，相较于单独采用对抗训练，针对攻击特定优化或许是实现对抗鲁棒性的一条更有效途径；该方法的总体准确率达到 85.0%（95% 置信区间：74.1–92.5%），且部署考量有实证依据。

参考文献

[1] Y. Wang、J. Yang 和 N. Chen，《使用深度学习检测物联网网络中的零日攻击》，*Future Generation Computer Systems*，第 151 卷，336–347 页，2024年。[2] A. L. Buczak 和 E. Guven，《网络安全入侵检测的数据挖掘与机器学习方法综述》，*IEEE Commu- nications surveys & tutorials*，第 18 卷，第 2 期，1153–1176 页，2016年。[3] R. Vinayakumar、M. Alazab、K. Soman、P. Poornachandran、A. Al- Nemrat 和 S. Venkatraman，《智能入侵检测系统的深度学习方法》，*IEEE Access*，第 7 卷，41 550，2019年 41 525pp. – 。[4] Z. Ahmad 、 A. Shahid Khan 、 C. Wai Shiang 、 J. Abdullah 和 F. Ahmad，《网络入侵检测系统：机器学习与深度学习方法的系统研究》，*Transactions on Emerging Telecommunications Technologies*，第 32 卷，第 1 期，e4150 页，2021年。[5] Z. Chen、C. Liu、D. Jiang、Q. Huang 和 J. Su，《用于网络入侵检测的图神经网络：综述》，*IEEE Access*，第 11 卷，49 114–49 133 页，2023年。[6] Y. Liu、X. Wang 和 J. Chen，《基于Transformer的入侵检测：全面综述》，*Computer Networks*，第 228 卷，109742 页，2024年。[7] M. A. Ferrag、O. Friha、D. Hamouda、L. Maglaras 和 H. Janicke，《基于深度学习的分布式拒绝服务攻击入侵检测：全面综述》，*IEEE Access*，第 11 卷，4426，2023年 4401pp. – 。[8] T. D. Nguyen、S. Marchal、M. Miettinen、H. Fereidooni、N. Asokan 和 A.-R. Sadeghi，《用于入侵检测系统的联邦学习：综述》，*ACM Computing Surveys*，第 55 卷，第 13s 期，1–35 页，2023年。[9] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，《迈向能够抵御对抗攻击的深度学习模型》，发表于国际学习表征会议 *(ICLR)*，2018年。[10] A. Venturi、A. Giaretta 和 M. Conti，《针对基于图神经网络的入侵检测系统的结构化对抗性攻击》，*Com- puters & Security*，第 137 卷，103586 页，2024年。

[11] F.-E. Ennaji、S. El Brak和N. Aboutabit，“基于深度学习的网络入侵检测对抗攻击：分类体系与未来方向”，Journal of Network and Computer Applications，第218卷，第103722页，2024年。[12] A. Sharma、S. K. Sahay和A. Kumar，“入侵检测系统中的对抗攻击与防御：综述”，ACM Computing Surveys，第56卷，第5期，第1–38页，2024年。[13] M. Ring、S. Wunderlich、D. Scheuring、D. Landes和A. Hotho，“基于网络的入侵检测数据集综述”，Computers & Security，第86卷，第147–167页，2019年。[14] A. Khraisat、I. Gondal、P. Vamplew和J. Kamruzzaman，“入侵检测系统综述：技术、数据集与挑战”，Cybersecurity，第2卷，第1期，第1–22页，2019年。[15] I. J. Goodfellow、J. Shlens和C. Szegedy，“解释并利用对抗样本”，arXiv预印本arXiv:1412.6572，2014年。[16] C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. Goodfellow和R. Fergus，“神经网络的有趣性质”，见国际学习表征会议（ICLR），2014年。[17] N. Carlini和D. Wagner，“迈向神经网络鲁棒性评估”，见2017年IEEE安全与隐私研讨会（SP）。IEEE，2017年，第39–57页。[18] N. Papernot、P. McDaniel和I. Goodfellow，“机器学习中的可迁移性：从现象到利用对抗样本的黑盒攻击”，arXiv预印本arXiv:1605.07277，2016年。[19] S.-M. Moosavi-Dezfooli、A. Fawzi和P. Frossard，“Deepfool：一种简单而准确的欺骗深度神经网络的方法”，见IEEE计算机视觉与模式识别会议论文集，2016年，第2574–2582页。[20] J. Zhao、L. Wang和Y. Zhang，“针对基于神经网络的网络入侵检测系统的分布偏移攻击”，Computers & Security，第136卷，第103542页，2024年。[21] Z. Lin、Y. Shi和Z. Xue，“Idsgan：用于针对入侵检测进行攻击生成的生成对抗网络”，见亚太知识发现与数据挖掘会议。Springer，2022年，第79–91页。[22] M. Heydari、R. Alhajj和J. Rokne，“用于网络入侵检测的深度学习模型的对抗鲁棒性：综合分析”，IEEE网络与服务管理汇刊，2025年。[23] G. Apruzzese、M. Andreolini、M. Marchetti、A. Venturi和M. Colajanni，“强化随机森林网络安全检测器以抵御对抗攻击”，IEEE计算智能新兴主题汇刊，第6卷，第4期，第1141–1153页，2021年。[24] K. Sun、K. Liu、S. Xu、J. Liu和K. Zhang，“深度学习中的对抗鲁棒性：攻击、防御与未来方向”，IEEE《神经网络与学习系统汇刊》，2023年。[25] F. Tram`er、A. Kurakin、N. Papernot、I. Goodfellow、D. Boneh和P. McDaniel，“集成对抗训练：攻击与防御”，见国际学习表征会议，2018年。[26] B. Biggio、I. Corona、D. Maiorca、B. Nelson、N. ˇSrndi´c、P. Laskov、G. Giacinto和F. Roli，“测试时针对机器学习的逃逸攻击”，欧洲机器学习与数据库知识发现联合会议，第387–402页，2013年。[27] T. G. Dietterich，“机器学习中的集成方法”，多分类器系统国际研讨会，第1–15页，2000年。[28] R. R. Barik和R. Misra，“用于增强网络入侵检测的集成深度学习”，Expert Systems with Applications，第235卷，第121043页，2025年。[29] M. Awad等，“用于网络入侵检测的基于注意力的集成学习”，IEEE Access，第13卷，第1234–1247页，2025年。[30] X. Liu、Y. Li、C. Wu和C.-J. Hsieh，“通过随机自集成迈向鲁棒的神经网络”，欧洲计算机视觉会议（ECCV）论文集，第369–385页，2018年。[31] A. Merzouk等，“用于网络入侵检测的基于扩散的对抗净化”，Computers & Security，2024年。[32] W. Xu、D. Evans和Y. Qi，“特征压缩：检测深度神经网络中的对抗样本”，arXiv预印本arXiv:1704 01155，2017年。[33] T. Pang、K. Xu、C. Du、N. Chen和J. Zhu，“重新思考用于对抗鲁棒性的softmax交叉熵损失”，arXiv预印本arXiv:1905.10626，2019年。[34] A. Athalye、N. Carlini和D. Wagner，“混淆梯度带来虚假的安全感：规避针对对抗样本的防御”，见国际机器学习会议（ICML）。PMLR，2018年，第274–283页。[35] X. Ma、B. Li、Y. Wang、S. M. Erfani、S. Wijewickrema、G. Schoenebeck、D. Song、M. E. Houle和J. Bailey，“刻画对抗-

ial子空间利用局部内在维度”，*arXiv*预印本*arXiv:1801.02613*，2018年。[36] W. Chen、X. Li和H. Wang，“Dynamicids：面向入侵检测的动态防御选择”，*IEEE*《可靠与安全计算汇刊》，2024年。[37] M. Usama、J. Qadir、A. Raza、H. Arif、K.-L. A. Yau、Y. Elkhatib、A. Hussain和A. Al-Fuqaha，“用于发起并挫败针对网络入侵检测系统对抗攻击的生成对抗网络”，*2019*年第*15*届国际无线通信与移动计算会议（*IWCMC*），第78–83页，2019年。[38] M. Pawlicki、M. Chora´s、R. Kozik和W. Hołubowicz，“迈向生成对抗网络在网络入侵检测中的实际应用”，《计算与信息学》，第39卷，第1-2期，第0119–0130页，2020年。[39] Y. Zhong、W. Chen、Z. Wang、Y. Chen、K. Wang、Y. Li、X. Yin、X. Shi、J. Yang和K. Li，“物联网系统中用于网络入侵检测的对抗学习”，*IEEE*《物联网期刊》，第8卷，第13期，第10 205–10 217页，2020年。[40] X. Deng、Q. Liu、Y. Deng和S. Mahadevan，“深度学习中面向物联网网络入侵检测的对抗攻击与防御”，*IEEE/CAA*《自动化学报》，第8卷，666，2021年 3，656期，页–。[41] H. Zhang、Y. Yu、J. Jiao、E. Xing、L. El Ghaoui和M. Jordan，“鲁棒性与准确率之间有理论依据的权衡”，国际机器学习会议，第7472–7482页，2019年。[42] D. Meng和H. Chen，“MagNet：一种针对对抗样本的双管齐下防御”，见*2017*年*ACM SIGSAC*计算机与通信安全会议论文集，2017年，第135–147页。[43] C. E. Shannon，“《通信的数学理论》”，《贝尔系统技术期刊》，第27卷，第3期，第379–423页，1948年。[44] P. C. Mahalanobis，“论统计学中的广义距离”，印度国家科学院院刊，第2卷，55，1936年 49页–。[45] I. Sharafaldin、A. H. Lashkari和A. A. Ghorbani，“面向生成新的入侵检测数据集与入侵流量特征描述”，*ICISSp*，第1卷，第108–116页，2018年。[46] N. Moustafa和J. Slay，“UNSW-NB15：面向网络入侵检测系统的综合数据集（UNSW-NB15网络数据集）”，*2015*年军事通信与信息系统会议（*MilCIS*），第1–6页，2015年。[47] M.-I. Nicolae、M. Sinn、M. N. Tran、B. Buesser、A. Rawat、M. Wistuba、V. Zantedeschi、N. Baracaldo、B. Chen、H. Ludwig、I. M. Molloy和B. Edwards，“对抗鲁棒性工具箱v1.0.0”，*arXiv*预印本*arXiv:1807.01069*，2018年。[48] L. Breiman，“随机森林”，机器学习，第45卷，第1期，第5–32页，2001年。[49] T. Chen和C. Guestrin，“XGBoost：一种可扩展的树提升系统”，见第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，2016年，第785–794页。