基于可解释人工智能的流量对抗样本攻击及防御方法

马博文[1]，郭渊博[2]，田继伟[3]，马骏[1]，胡永进[1]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 海南大学网络空间安全学院，海南 海口 570100；
3. 空军工程大学空管领航学院，陕西 西安 710000）

摘　要：针对基于人工智能的网络入侵检测系统，提出了一种基于可解释人工智能（XAI）的对抗样本攻击方法。利用XAI方法识别关键扰动特征，在保持流量功能时逐步进行针对性扰动，直至恶意流量被判定为良性，实现对抗流量样本攻击。这种方法可以大幅减少扰动特征，增强了攻击隐蔽性，而且其所识别的关键特征对不同分类器具有一致性，使得攻击样本具有较强的迁移性。在防御方面，提出了一种基于对抗训练的防御方法，以提升网络入侵检测系统的鲁棒性。实验结果表明，所提攻击方法具有较高的攻击成功率和迁移成功率；所提防御方法可以有效降低对抗样本攻击的成功率，增强了系统的鲁棒性。

关键词：对抗样本攻击；可解释人工智能；网络入侵检测；恶意对抗流量
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025067

Traffic adversarial example attack and defense method based on explainable artificial intelligence

MA Bowen[1], GUO Yuanbo[2], TIAN Jiwei[3], MA Jun[1], HU Yongjin[1]
1. Cryptography Engineering Institute, Information Engineering University, Zhengzhou 450001, China
2. School of Cyberspace Security, Hainan University, Haikou 570100, China
3. Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an 710000, China

Abstract: An adversarial example attack method based on XAI was proposed for AI-based NIDS. By identifying critical perturbation features with XAI and applying targeted perturbations while preserving traffic functionality, malicious traffic was gradually altered to be misclassified as benign, achieving adversarial traffic sample attacks. This approach reduced the number of required perturbation features, enhancing attack stealthiness. The identified features showed consistency across classifiers, giving attack samples robust transferability. For defense, a defense method based on adversarial training was proposed to boost NIDS robustness. Experiments show high attack success and transfer rates, and the proposed defense method effectively lowers adversarial example attack success rates, enhancing system robustness.

Keywords: adversarial example attack, explainable artificial intelligence, network intrusion detection, malicious adversarial traffic

0　引言

随着网络攻击的数量和强度不断增加，高效的检测及防御方法变得愈加重要。网络入侵检测系统（NIDS, network intrusion detection system）[1]是检

收稿日期：2024–11–26；**修回日期：**2025–03–29
通信作者：郭渊博，yuanbo_g@hotmail.com
基金项目：国家自然科学基金资助项目（No.62402520）；国家社会科学基金资助项目（No.2022-SKJJ-B-057）；陕西省自然科学基金资助项目（No.2024JC-YBQN-0620）
Foundation Items: The National Natural Science Foundation of China (No.62402520), The National Social Science Fund of China (No.2022-SKJJ-B-057), Shaanxi Provincial Natural Science Foundation (No.2024JC-YBQN-0620)

测网络攻击的重要工具，通常被部署在网络设备边缘，通过监视网络流量识别异常事件或可疑事件并生成警报，在抵御针对企业、个人和政府的网络攻击方面发挥着重要作用[2]。

近年来，人工智能（AI, artificial intelligence）的快速发展促进了入侵检测的进步，机器学习、深度学习等AI技术能够充分利用大量数据来学习抽象和非线性的表示，从而以更高的准确性和更少的限制区分正常流量和异常流量[3-4]，同时在恶意邮件检测、恶意软件检测等网络安全领域展现出巨大潜力。随着AI相关技术在各种实际应用中的不断深入[5-7]，其鲁棒性问题逐渐成为不可忽视的关键挑战，对抗样本攻击、投毒攻击、后门攻击等技术直接针对AI模型的攻击将直接影响系统的机密性、完整性和可用性[8]。

对抗样本攻击技术由Szegedy等[9]首次提出，在计算机视觉（CV, computer vision）领域大量研究并获得广泛关注[10-11]，攻击者将难以觉察的对抗扰动注入模型输入中得到对抗样本，该对抗样本会导致正常训练的模型做出错误决策，从而实现对模型的欺骗。该技术在自然语言处理、语音识别、强化学习、联邦学习等领域具有广泛研究[12-15]，对AI系统的安全性造成巨大威胁，甚至直接影响人们的人身、财产和隐私安全。

作为高安全敏感的应用场景，入侵检测在得到AI技术助益的同时，也面临着AI自身鲁棒性问题带来的安全挑战。攻击者可以利用对抗样本攻击技术攻击基于AI的网络入侵检测系统，通过对恶意流量进行微小扰动生成对抗样本，使其成功绕过系统检测，导致系统失效。

可解释人工智能（XAI, explainable artificial intelligence）是一种旨在提高人工智能模型理解性的技术[16-17]，可以帮助用户和开发者理解“黑盒”模型做出决策的原因，增强模型的“透明度”。一些研究将XAI技术与对抗样本相结合，利用XAI技术生成对抗样本[18-21]。文献[18]通过将注意力机制引入对抗样本生成，通过可视化技术形成注意力图实现模型解释，并基于模型注意力破坏重要的中间特征以实现对抗扰动，实现了针对ImageNet分类器的攻击。文献[19]通过引入聚集梯度来获得图像的特征重要性，并通过破坏主导模型决策的重要对象感知特征生成对抗样本，该攻击同时具有较强的可转移性。文献[20]利用神经元属性精确估计神经元重要性，实现了基于神经元属性的攻击方法，生成的对抗样本在可转移性上具有较大提升，并通过神经元属性的近似计算大大减少了计算开销。然而，上述利用XAI技术实现对抗样本攻击的方法大都来自计算机视觉领域。CV领域样本通常以像素为扰动特征，具有较高的维度，生成的对抗样本需满足人眼不可见，对扰动特征的范围和位置没有特殊限制，扰动特征的数量也相对较大。和CV领域相比，NIDS中的对抗样本通常维度较小，且扰动时需确保生成的流量满足可用性和恶意性等约束条件，这意味着对流量样本的扰动只能局限于特定范围，在实施攻击时对一些关键特征必须谨慎处理，以确保扰动后的流量仍然能够达到预期目的，因此CV领域的相关技术往往无法直接移植于NIDS。

针对NIDS的特点，本文将XAI技术引入流量对抗样本生成，在考虑流量独特限制的情况下实现攻击，以欺骗基于AI的网络入侵检测系统。XAI技术的引入使扰动效率大幅提升，所需扰动特征数大幅减少，增强了攻击隐蔽性。同时，XAI识别的关键扰动特征对不同分类器具有一定的一致性，使生成的攻击样本具有较强的迁移性。此外，为了对此类攻击进行防御，本文提出了基于对抗训练的防御方法，以提升NIDS的鲁棒性。本文具体贡献如下。

1) 提出了一种基于XAI的流量对抗样本攻击方法，该方法仅需扰动少量特征即可实现攻击，具有较强的攻击隐蔽性。同时，XAI识别的关键扰动特征对不同分类器具有一定的一致性，使生成的攻击样本具有较强的迁移性。

2) 在NSL-KDD数据集和UNSW-NB15数据集上分别进行实验。实验结果表明，所提攻击方法仅需扰动少数特征即可实现高攻击成功率，且在8个不同分类器上的平均迁移攻击成功率均高于70%，证实了本文所提攻击方法的有效性和较强迁移性。

3) 提出了一种基于对抗训练的防御方法，以提高网络入侵检测系统面对对抗攻击的鲁棒性，实验结果证明了本文所提防御方法的有效性。

1　相关工作

1.1　基于机器学习的网络入侵检测系统

近年来，人工智能的快速发展使其成功应用于

多种网络安全威胁检测[22-23]。支持向量机（SVM, support vector machine）、决策树（DT, decision tree）等机器学习算法大量应用于入侵检测系统，且取得了较好的效果[24-25]。随着数据量日益增大以及硬件计算能力不断提升，卷积神经网络、循环神经网络（RNN, recurrent neural network）、自动编码器等深度学习算法广泛应用于入侵检测系统，并在处理海量高维非线性数据上取得了较高的成功率。文献[26]提出了一种基于RNN的无线网络入侵检测分类模型，并提出了基于窗口的实例选择算法以解决训练数据样本分布不均衡导致的过拟合问题。该模型通过系统化的数据预处理和模型优化策略，显著提升了入侵检测模型的分类准确率和执行效率。文献[27]提出了一种基于相关信息熵和CNN-BiLSTM（CNN-bidirectional long short-term memory）的入侵检测模型，首先利用相关信息熵减少特征维度，然后利用深度学习算法实现分类。文献[28]提出了一种结合卷积神经网络和门控循环单元的入侵检测模型，在解决数据不平衡问题的同时达到了较高的检测率。文献[29]提出了一种深度卷积残差自动编码记忆（DCRAE-M, deep convolution residual autoencoding memory）网络的多变量时间序列异常检测方法，该方法能够准确捕捉变量间的细粒度依赖分布特征，并减少空间维度上的特征信息丢失，该方法的平均F1分数相比基线方法提高了0.23，实现了更高的检测性能。然而，AI技术帮助入侵检测系统取得巨大进展的同时，其自身存在的鲁棒性问题也带来了新的安全隐患，攻击者可以利用对抗样本攻击等多种技术攻击AI模型，从而导致基于AI的网络入侵检测系统出现模型隐私泄露、检测准确率下降、恶意流量逃逸等问题。

1.2 对抗样本攻击及防御

针对基于AI的网络入侵检测系统，对抗样本攻击者通过对恶意流量添加扰动，使NIDS将其分类为正常流量，从而实现对入侵检测系统的规避，造成了巨大安全威胁。在攻击具体实现方式上，一些研究将计算机视觉领域的对抗样本攻击方法进行迁移，得到针对NIDS的对抗流量。文献[30]利用快速梯度符号攻击和基于雅可比显著图的攻击方法，针对多层感知机和由决策树、随机森林（RF, random forest）和支持向量机构成的集成分类器发起对抗样本攻击，生成的样本可以使分类器性能指标下降。Ibitoye等[31]进一步评估了快速梯度符号攻击、基本迭代法、投影梯度下降法等攻击方法的有效性。然而，直接迁移CV领域的攻击方法往往没有考虑入侵检测与CV领域的不同之处，故生成的流量往往在可用性上存在问题。

为了解决可用性问题，研究者针对流量的特殊限制实施了流量对抗样本攻击[32-39]。文献[32]分析了流量的具体约束，通过检查每次迭代中是否违反约束来保持生成流量的有效性。文献[35]提出利用Attack-GAN实现对抗流量生成，通过在生成器中引入字节顺序等预定义的网络约束来保持流量功能。文献[36]提出了一种基于问题空间扰动的对抗攻击方法，实现了拒绝服务（DoS, denial of service）流量的对抗样本，在贝叶斯网络、决策树、深度神经网络（DNN, deep neural network）等分类器上取得了良好的攻击效果。文献[37]从网络流量中提取了时序特征，并利用快速梯度符号攻击方法实现攻击。在CICIDS2017数据集上的实验结果表明，其提出的攻击框架在端口扫描、拒绝服务攻击等多类恶意流量中均具有良好的攻击效果。文献[38]将扰动限制在非功能特征上，利用生成对抗网络（GAN, generative adversarial network）生成对抗流量。文献[39]利用粒子群优化（PSO, particle swarm optimization）算法、遗传算法（GA, genetic algorithm）和GAN等方法生成对抗流量，在多种不同的机器学习分类器中实现了较高的攻击成功率。上述方法在生成对抗流量时，往往需要对流量进行较多扰动，而扰动越多意味着攻击暴露的风险越大。同时，大多研究仅考虑对特定的流量分类器进行攻击，没有考虑攻击的迁移性，导致其适用范围较窄。

在防御方面，研究者提出了一系列方法以检测或缓解对抗样本攻击。现有的防御方法通常可以分为以下3类[40]：模型参数保护、对抗样本检测和鲁棒性优化。模型参数保护[41]方法通过梯度掩蔽防止模型梯度暴露，使基于梯度的对抗样本攻击方法无法获取模型梯度从而实现防御。对抗样本检测[42]方法旨在模型进行实际分类前将对抗样本检出从而实现防御，但对抗样本检测器的引入不可避免地会带来额外的处理时延。鲁棒性优化方法则通过优化分类器，使其面对对抗性样本时具有鲁棒性，其中的对抗性训练[43]是鲁棒性优化的典型代

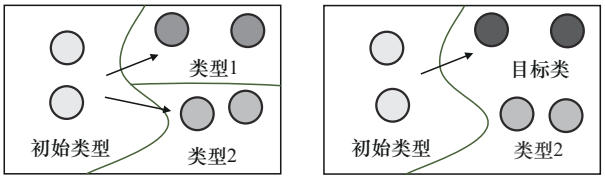
表，也是CV领域最有效的防御方法之一。该方法通过在训练阶段加入对抗样本，使模型在训练阶段便可以学到对抗样本的特点，从而使模型在面对对抗样本时可以做出正确响应。然而，和攻击情况相似，上述防御方法也多集中于CV领域，针对NIDS的对抗防御方法尚处于起步阶段。文献[44]针对快速梯度符号法、DeepFool方法、投影梯度下降法、Carlini&Wagner方法等攻击方法，提出了一种基于迁移学习的对抗检测方法，并利用多检测器结合的方式提高了对抗流量的检测能力。文献[45]针对NIDS上的Carlini&Wagner攻击方法，提出了一种两阶段防御方法，训练阶段使用高斯数据增强强化对抗训练，测试阶段则通过特征压缩对样本进行处理后再进行分类，强化后的NIDS达到了89.7%的分类准确率。尽管已有部分研究进行了对抗样本防御方面的研究，但相关技术仍处于探索阶段，有待进一步发展和完善。

2　本文方法

2.1　攻击目标

基于AI的网络入侵检测系统往往通过对流量进行分类而检测异常，对模型$y=f(x)$，入侵检测系统利用分类模型f对x进行分类，并通过分类结果y来具体判定是否为恶意流量。攻击者则可以利用对抗样本技术对分类模型f进行攻击，通过对原输入样本x添加扰动δ，扰动后的样本$x'=x+\delta$称为对抗样本，将x'作为输入以欺骗模型，使其产生错误的分类结果。

从攻击效果来看，对抗样本攻击类型可以分为无目标攻击和有目标攻击，如图1所示。在无目标攻击中，攻击者添加扰动后，只需模型的预测标签不等于原标签，即$f(x+\delta)\neq y_{\text{original}}$，便可视为攻击成功。而有目标攻击除了使模型发生错误预测之外，还必须使模型将样本分类为攻击者预设的目标标签，即$f(x+\delta)=y_{\text{target}}$。和无目标攻击相比，有目标攻击实现了对模型更加精准的攻击，在实现难度上也要大于无目标攻击。



(a) 无目标攻击　　(b) 有目标攻击

图1　对抗样本攻击类型

本文攻击者目的是欺骗网络入侵检测系统，使其将扰动后的恶意流量识别为良性，因此只能采取有目标攻击方式，并将目标标签y_{target}设置为良性。如果攻击者采用无目标攻击方式，则攻击后的恶意流量有可能被分类成另一种恶意流量，导致攻击失败。

在网络入侵检测系统攻击中，对流量的扰动也需要满足一定要求。CV领域的扰动往往以“微小”为约束条件，通常要求人眼不可见，而对流量的扰动则不再以“微小”为约束条件，而是需要保证生成的流量满足可用性、恶意性等约束条件，使得对流量的扰动只能在一定范围内进行[46]。如果对保持流量功能和恶意性的关键特征进行了修改，则会导致生成的流量无效或失去其原有功能。

2.2　攻击场景

根据攻击者对网络入侵检测系统的了解程度，攻击场景可分为以下3种。

1) 白盒攻击。攻击者对分类模型的训练数据、算法、模型和参数有完整的知识，但在现实攻击场景中通常具有局限性。

2) 灰盒攻击。攻击者具有分类模型的部分信息，如部分训练数据和部分参数，攻击难度高于白盒攻击。

3) 黑盒攻击。攻击者只知道模型的输入和输出，而不知道模型结构、参数或训练数据等先验知识，仅能通过访问模型获得的分类结果来获取信息。黑盒攻击的应用场景更现实，同时也更具挑战性。

黑盒攻击通常需要多次访问模型，但访问次数过多会引起管理者的重视，导致攻击者的访问权限被终止。因此，访问次数越少越有利于实现黑盒攻击。此外，攻击者可以利用迁移攻击方法实现黑盒攻击，即利用其他代理模型生成对抗流量，再用其攻击真正的目标模型。这种方法不需要对目标模型进行探测，是较为理想的一种黑盒攻击方法，但由于代理模型与目标模型的差异性，往往存在攻击成功率较低的问题。

在本文场景中，攻击者无法获知用户使用的分类器框架和详细参数，但其可以伪装为正常用户对模型进行探测，通过访问模型获得分类结果，符合黑盒攻击场景的定义。此外，本文还尝试了对8种机器学习模型进行迁移攻击。

2.3 攻击方法

本文将XAI技术引入对抗样本攻击，具体而言，针对训练完成的NIDS模型，利用XAI技术得到具体流量特征对不同标签的“贡献值”。在流量约束条件下，取恶意流量中对“良性”标签判定贡献较大的特征，依据重要性由大到小依次对其进行针对性扰动，使该恶意流量被判定为“良性”的概率逐渐增大，直至恶意流量样本跨过模型决策边界，被分类为“良性”。上述过程不需要对NIDS进行任何修改，仅需对部分特征进行微调，即可欺骗训练完成的NIDS模型，成功实现流量对抗样本攻击。

本文使用SHAP（SHapley additive exPlanations）方法[17]对分类模型进行解释，并在此基础上进一步实现对抗样本攻击。SHAP方法将博弈论引入模型解释，具有局部保真性、一致性等良好性质，同时具备理论完备性。针对模型分类问题，给定输入X和分类模型f，对每个类别c，该类别预测的解释如式(1)所示。

$$P(f(X)=c|X)=\phi_{0,c}+\sum_{i=1}^{M}\phi_{i,c}^{X}x_i \tag{1}$$

其中，x_i是样本X的第i个特征，M是样本X包含的特征数，$P(f(X)=c|X)$是分类模型f在输入为X时，预测为类别c的概率，$\phi_{0,c}$是基线值，通常是所有训练样本中类别c的平均预测概率，$\phi_{i,c}^{X}$是样本X的特征i在类别c上的SHAP值，即为该特征对样本X判定为类别c的“贡献值”。具体算法如算法1所示。

算法1 基于XAI的流量对抗样本攻击

输入 干净样本集D_{clean}，D_{clean}样本数N，样本包含特征数M，流量分类模型f，扰动特征数T，良性目标标签BL

输出 对抗样本集D_{attacked}

//计算整体特征重要性

1)for each data X in D_{clean}

2) for each feature i in data X

3) get $\phi_{i,\text{BL}}^{X}$ from 式(1)

4) end for

5)end for

6)for each feature i from 1 to M

7) $\text{feature importance}_{i,\text{BL}}=\dfrac{\sum_{x=1}^{N}\left|\phi_{i,\text{BL}}^{X}\right|}{N}$

8)end for

//依据特征重要性从大到小对特征进行排序

9)sorted feature list=sort feature according to feature importance

//对恶意数据X进行对抗扰动

10)$D_{\text{attacked}}=\varnothing$

11)for each malicious data X in D_{clean}

12). perturbed_count=0

13) $\tilde{X}=X.\text{copy}()$

14) for each feature i in sorted feature list

15) while perturbed_count<T && $f(\tilde{X})\neq$ BL

16) if feature$\tilde{x}_i$ can be perturbed

17) if $\sum_{x=1}^{N}\phi_{i,\text{BL}}^{\tilde{X}}\geqslant 0$

18) $\tilde{x}_i=\max(\tilde{x}_i^1,\tilde{x}_i^2,\cdots,\tilde{x}_i^N)$

19) perturbed_features_count += 1

20) else

21) $\tilde{x}_i=\min(\tilde{x}_i^1,\tilde{x}_i^2,\cdots,\tilde{x}_i^N)$

22) perturbed_features_count += 1

23) end if

24) else

25) pass

26) end if

27) end while

28) append the adversarial example $\tilde{X}$ to D_{attacked}

29) end for

30)end for

31)return D_{attacked}

算法1首先利用式(1)得到不同特征对良性标签的贡献值，并将特征的贡献值进行叠加，得到良性标签下的特征整体重要性，进而依据特征整体重要性从大到小逐次对恶意样本$\tilde{X}$进行特征扰动。若特征i为可扰动特征，且对样本$\tilde{X}$被判定为良性具有正向贡献，即特征值$\tilde{x}_i$越大，样本$\tilde{X}$被判定为良性的概率越大，则将特征值$\tilde{x}_i$扰动为该特征在数据集范围内的最大值，以进一步增大该特征的贡献，确保扰动后特征的有效性。若其对样本$\tilde{X}$被判定为良性具有负向贡献，即特征值$\tilde{x}_i$越小，样本$\tilde{X}$被判定为良性的概率越大，则将特征值x_i扰动为该特征在数据集范围内的最小值。上述扰动方式可使特征i对样本$\tilde{X}$判定为良性的贡献进一步放大。通过上述

扰动方式，对样本进行逐次特征扰动，直到满足设定的扰动特征数或该恶意流量样本被分类器判定为良性。

对于完成训练的入侵检测模型，其决策边界已经固定，若使恶意流量样本经过对抗扰动被判定为良性，则扰动需使其跨过恶意样本和良性样本的决策边界。攻击者在可变特征范围内，依次选取恶意流量样本中对良性类型“贡献”较大的关键特征进行增强，增强幅度需要满足流量的限制，直到样本跨过决策边界，扰动后的流量即为对抗样本流量。如图2所示，假设入侵检测模型已完成训练，其决策边界（用虚线表示）已经固定，恶意样本由A、B、C和D共4个特征（用粗实线箭头表示）组成，这4个特征共同决定了该样本在决策平面的位置。攻击者通过添加扰动使该样本越过决策边界被分类为良性。根据算法1，攻击者选择对良性类型“贡献”较大的特征C和D进行“增强”（用细虚线箭头表示），使该恶意样本在特征C和D“增强”后越过决策边界（箭头E），从而判定为良性。对良性类型“贡献”较小的特征A和B，攻击者不进行扰动。

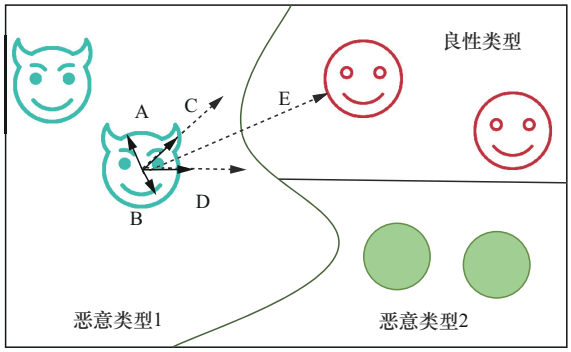


图2　关键扰动特征对攻击样本生成的影响过程

由于扰动特征的选取来自对“良性”标签判定贡献大的特征，保证了扰动收益的最大化，因此仅需扰动少数特征即可成功实现对分类器的对抗攻击。同时，算法1仅对可扰动特征进行扰动，且扰动取值取于数据集本身，保证了扰动后流量数据的可用性。

此外，尽管不同的流量分类模型在分类方式和分类路径上存在差异，导致同一特征对不同分类器的重要程度不完全一致，但特征的重要性主要由数据集本身决定。因此，对不同分类器，同一数据集的特征重要性差异不大，即某一分类器的重要特征通常对其他分类器也具有重要性，对这些重要特征进行扰动所生成的对抗样本，理论上对其他分类器同样具有较好的攻击效果。即算法1得到的对抗样本在理论上具有较强的迁移性。

2.4　防御方法

为了抵御对抗样本攻击，使分类模型在面对伪装成良性流量的对抗样本时能够实现正确分类，本文基于对抗攻击技术提出了流量对抗样本防御方法，以提高流量分类模型对对抗样本攻击的鲁棒性，具体算法如算法2所示。

算法2　流量对抗样本防御

输入　干净样本集D_{clean}，流量分类模型f，训练迭代次数N_{iter}，对抗训练中对抗样本占比P

输出　鲁棒分类模型f_θ，θ为模型f的可学习参数

//使用干净样本集D_{clean}训练模型f

1)for 1,2,⋯,N_{iter}

2)　optimize θ by training f_θ using D_{clean}

3) end for

//生成对抗训练数据集

4) generate the adversarial example set D_{attacked} using algorithm 1

5)for each adversarial example $\tilde{X}$ in D_{attacked}

6)　set the label of $\tilde{X}$ as the ground-truth label

7)end for

8)$D_{\text{adv\_train}}$=sample (D_{clean}, round (len (D_{clean})(1−P)))+ sample (D_{attacked}, round (len (D_{attacked})P))

//利用对抗训练数据集进行训练，得到鲁棒模型

9)for 1,2,⋯,N_{iter}

10)　optimize θ by training f_θ using $D_{\text{adv\_train}}$

11) end for

算法2首先利用干净样本集D_{clean}训练模型f，通过多次迭代优化模型参数θ。随后，利用算法1生成对抗样本数据集D_{attacked}，并将每个对抗样本$\tilde{X}$的标签设置为其原始的真实标签，然后利用干净样本集D_{clean}和对抗样本集D_{attacked}构建对抗训练数据集$D_{\text{adv_train}}$。最后，利用对抗训练数据集$D_{\text{adv_train}}$对模型f进行多次迭代训练，以进一步优化模型参数θ，得到鲁棒分类模型f_θ。算法2采用预训练和对抗训练的渐进式训练方式，避免模型过早过拟合对抗样本，同时通过参数P调节对抗样本的引入量，防止训练过程被对抗样本主导，模型f在利用算法2进行增强后，可以增强面对对抗样本攻击的鲁棒性。

3 实验分析

3.1 数据集及预处理

本文选用了2个具有代表性的入侵检测数据集，分别为NSL-KDD[47]和UNSW-NB15[48]。由于生成方法、攻击流量的流行程度和大小等因素影响，2个数据集在攻击类型、特征提取等方面并不相同，但2个数据集都具有各种网络流量类型和攻击类型，且都混合了良性流量和恶意流量。

NSL-KDD数据集是麻省理工学院林肯实验室建立的KDD'99数据集的改进版本，涵盖超过125 000个训练样本和22 000个测试样本。数据集包含良性数据和4个恶意数据类型，即拒绝服务类、扫描（Probe）类、本地提权（U2R, user to root）类和远程（R2L, remote to local）类。该数据集包含41个有效特征，具体可分为intrinsic、content、time-based和host-based。对不同类型的恶意流量而言，某些特定的特征具有较强的功能性，对其进行修改将会使流量无效或失去原有功能。NSL-KDD数据集的功能特征与攻击类型的关系如表1所示[49]，其中，“√”表示功能特征，在扰动时不可改变，“—”表示可改变。

表1 NSL-KDD功能特征与攻击类型的关系

攻击类型	intrinsic	content	time-based	host-based
DoS	√	—	√	—
Probe	√	—	√	√
R2L	√	√	—	—
U2R	√	√	—	—

依据上述分析，对NSL-KDD数据集，在生成恶意对抗流量欺骗分类器时，需将扰动特征限制在可变特征范围内，即可保持流量的可用性和恶意性。

数据集预处理过程如下。1) 对其中的符号特征进行字典编码；2) 对二进制特征保持不变；3) 对数据类型特征按照式(2)进行min-max正则化，经过预处理后的每一条数据都具有41维特征。

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

此外，NSL-KDD数据集的训练集和测试集包含的标签类型不完全相同，因此将原有训练集与测试集进行合并，随机取其中80%的数据作为本文使用的测试集，取剩下的20%为训练集，保证训练集和测试集具有相同分布。

UNSW-NB15数据集是新南威尔士大学堪培拉网络靶场实验室创建的物联网网络流量数据集。数据集包含良性数据和9种恶意数据类型，涵盖超过175 000个训练样本和82 000多个测试样本。该数据集包含49个有效特征，这些特征除了分为直接从数据包中提取的flow features、base features、content features以及time features外，还从防御角度提取了general purpose features和connection features等生成特征。

在可变特征方面，UNSW-NB15数据集中的flow features和base features为不可扰动特征，即在生成对抗流量时仅扰动上述两类特征之外的特征，即可使生成的对抗流量保持其可用性和恶意性。

UNSW-NB15数据集预处理过程和NSL-KDD数据集类似。此外，将原有训练集与测试集进行合并，随机取其中80%的数据作为本文测试集，取剩下的20%为训练集。

3.2 评价标准

在流量分类方面，采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和F1分数（F1-Score）来评价分类结果，上述标准由真正例（TP, true positive）、真负例（TN, true negative）、假正例（FP, false positive）和假负例（FN, false negative）定义。

假设分类结果有正负两类，TP指实际为正、模型分类也为正的数据。TN指实际为负、模型分类也为负的数据。FP指实际为负、模型分类为正的数据。FN指实际为正、模型分类为负的数据。

准确率体现了总样本中预测正确的概率，整体上直观体现了模型性能，其计算式如式(3)所示。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3}$$

精确率又叫查准率，体现了预测为正的样本中实际为正的占比，其计算式如式(4)所示。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

召回率也叫查全率，体现了实际为正的样本中被预测为正的占比，其计算式如式(5)所示。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

F1分数综合考虑精确率和召回率，是精确率和召回率的加权调和平均，其计算式如式(6)所示。

$$\text{F1} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

对抗样本攻击的评价标准通常由攻击成功率体现，指恶意流量在经过对抗样本攻击后，被分类器成功分类为良性的比例，体现了对抗样本攻击的有效程度。

3.3　检测模型

本文采用多种机器学习模型对流量进行分类，并根据分类结果识别恶意流量。在具体模型方面，采用DNN作为基础分类模型，同时也作为对抗样本攻击的受害者模型。该模型包含1个有256个神经单元的隐藏层，又称DNN-256。激活函数为修正线性单元，优化器、损失函数、批大小、学习速率和训练轮次分别为ADAM、交叉熵函数、256、0.000 1和50。

此外，选取随机森林、K近邻（KNN, k-nearest neighbor）、逻辑回归（LR, logistic regression）、支持向量机、决策树和AdaBoost这6个机器学习模型，以及2个不同结构的DNN作为迁移攻击模型，其中DNN-128含有2个隐藏层，每层有128个神经单元，其他参数和DNN-256完全相同。DNN-64含有3个隐藏层，每层有64个神经单元，其他参数和DNN-256完全相同。其他机器学习模型参数设置如下。

RF设置为criterion= ‘gini’, max_depth=None, min_samples_leaf=1, min_samples_split=2；LR设置为max_iter=1 000，其他采用默认参数；AdaBoost设置为learning_rate=0.4，其他采用默认参数；KNN、SVM和DT均采用默认参数。

3.4　实验结果及分析

对NSL-KDD数据集，DNN-256等分类模型的初始分类情况如表2所示。

表2　NSL-KDD数据集初始分类情况

分类模型	准确率	精确率	召回率	F1分数
DNN-256	0.982 1	0.975 6	0.976 1	0.975 6
DNN-128	0.982 1	0.982 2	0.982 1	0.982 2
DNN-64	0.981 7	0.981 6	0.981 8	0.981 6
RF	0.985 7	0.985 6	0.985 8	0.985 3
KNN	0.988 1	0.988 1	0.988 1	0.988 1
LR	0.930 1	0.927 3	0.930 1	0.926 5
SVM	0.972 9	0.972 5	0.972 9	0.972 4
DT	0.975 0	0.979 7	0.975 0	0.976 6
AdaBoost	0.853 8	0.859 7	0.853 8	0.849 8

由表2可以看出，除LR和AdaBoost外，其他分类模型的准确率、精确率、召回率以及F1分数等评价指标均在0.95以上，LR和AdaBoost的准确率、精确率、召回率以及F1分数等评价指标也分别在0.9和0.85左右，即上述分类模型均可较好地完成NSL-KDD数据集的分类，用户可以选取其作为入侵检测模型。

取DNN-256作为算法1的流量分类模型f，将预处理后的NSL-KDD测试集作为算法1的干净样本集D_{clean}(NSL-KDD)，将f和D_{clean}(NSL-KDD)输入算法1后，得到扰动特征及扰动后的特征值如表3所示。

表3　NSL-KDD扰动特征情况

流量类型	扰动特征	特征重要性均值	扰动后特征值
DoS	dst_host_srv_count	1.657	255
	logged_in	1.450	1
	dst_host_count	1.301	0
	dst_host_same_srv_rate	0.789	0
	dst_host_serror_rate	0.676	0
	dst_host_rerror_rate	0.526	0
	hot	0.220	0
Probe	logged_in	1.450	0
	hot	0.220	0
	is_guest_login	0.081	1
	su_attempted	0.057	0
	root_shell	0.013	1
	num_root	0.002	1 743
	num_compromised	0.002	0
R2L及U2R	dst_host_srv_count	1.657	0
	dst_host_count	1.301	255
	srv_serror_rate	1.076	1
	dst_host_same_srv_rate	0.789	1
	dst_host_serror_rate	0.676	1
	dst_host_rerror_rate	0.526	1
	srv_count	0.525	0

对D_{clean}(NSL-KDD)和分类模型f，依次将表3结果作为对抗流量的扰动特征，并保持其他特征不

变，设置不同的扰动特征数，得到对应的对抗样本集D_{attacked}(NSL-KDD)（其规模和D_{clean}(NSL-KDD)相同）。利用分类模型f对对抗样本集D_{attacked}(NSL-KDD)进行分类，得到对抗样本攻击成功率与扰动特征数的关系如图3所示。

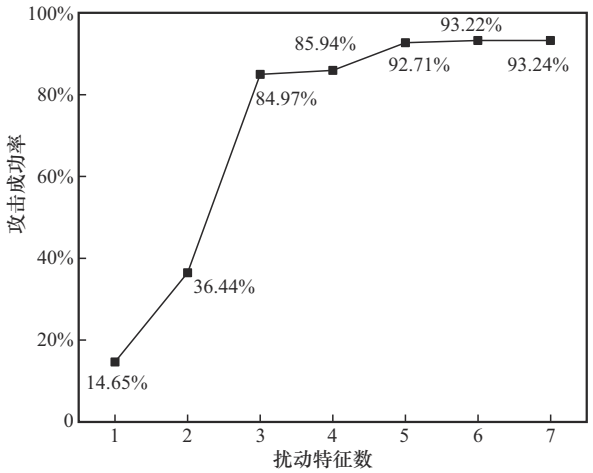


图3　NSL-KDD数据集对抗样本攻击成功率与扰动特征数的关系

由图3可以看出，随着扰动特征数的增加，对抗样本攻击成功率不断提高。当扰动特征数为3时，攻击成功率达到了84.97%。当扰动特征数增加到5时，此时攻击成功率达到了92.71%以上并保持稳定。

将本文攻击成功率结果与文献[38]和文献[39]进行对比，结果如表4所示。

表4　NSL-KDD数据集攻击成功率对比

方法	攻击成功率	扰动特征数
本文方法	93.24%	7
NIDSGAN	87.89%	7
PSO	44.43%	29
GA	77.87%	29
GAN	86.18%	29

从表4可以看出，当扰动特征数为7时，本文方法的攻击成功率高于NIDSGAN方法，且在扰动特征数更少的情况下，攻击成功率高于PSO、GA和GAN方法。对于对抗样本攻击而言，其扰动特征数越少，意味着攻击的隐蔽性越强，实施攻击的成本更低。

依据图3可以看出，由于扰动特征数增加到5时，攻击成功率达到92.71%以上并趋于稳定，而较少的扰动特征数保证了攻击的隐蔽性。因此，选取扰动特征数为5，对DNN-128、DNN-64、RF、KNN、LR、SVM、DT、AdaBoost等分类模型进行迁移攻击，迁移攻击成功率如图4所示。

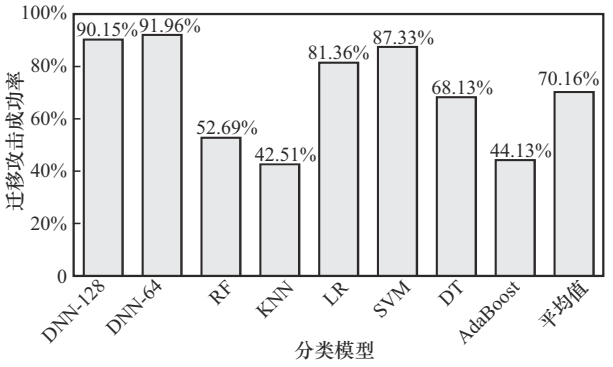


图4　NSL-KDD数据集不同分类模型的迁移攻击成功率

由图4可以看出，当扰动特征数为5时，针对DNN-256分类模型生成的对抗流量对其他分类模型具有较高的迁移攻击成功率，其中对DNN-128和DNN-64的迁移攻击成功率最高，分别达到了90.15%和91.96%，原因可能是DNN-128、DNN-64和初始分类器DNN-256结构相似，重要扰动特征重叠程度大。此外，对LR和SVM分类模型的迁移成功率均达到了80%以上，KNN和AdaBoost的迁移攻击成功率最差，但也达到了42.51%。所有8种分类模型的平均迁移攻击成功率达到了70.16%，说明本文方法在NSL-KDD数据集上具有较强的迁移性。

在防御实验方面，取DNN-256等分类模型作为算法2的流量分类模型f；将预处理后的NSL-KDD测试集作为算法2的干净样本集D_{clean}(NSL-KDD)。对于DNN-256、DNN-128、DNN-64等分类模型，训练迭代次数N_{iter}设置为50；对于RF、KNN、LR、SVM、DT、AdaBoost等不需要设置训练迭代次数的分类模型，正常进行训练。对抗训练中对抗样本占比P设置为0.01，得到防御后的攻击成功率如图5所示。

由图5可以看出，使用算法2进行防御后，对于NSL-KDD数据集，各个分类模型的攻击成功率明显下降。其中，DNN-256、RF和DT的攻击成功率分别下降至0.7%、0.17%和0.09%。此外，DNN-128、DNN-64、KNN和SVM分类模型的攻击成功率也降至5%以下，实现了较好的防御效果。LR和AdaBoost的防御效果不如上述分类模

型，但和图4中未防御结果相比，攻击成功率也分别下降了71.64%和19.23%。综上所述，本文所提防御方法可以在NSL-KDD数据集上达到较好的防御效果，各种流量分类模型在经过算法2防御后，面对对抗样本的鲁棒性明显加强。

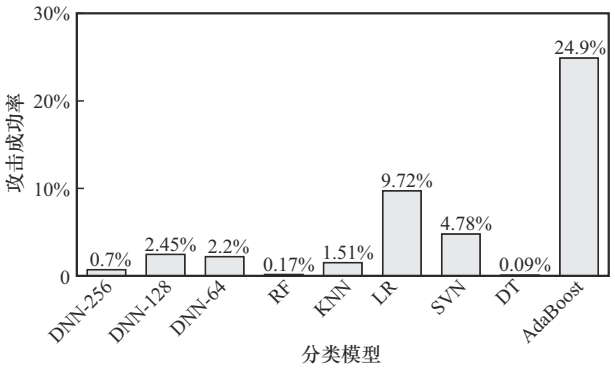


图5　NSL-KDD数据集防御后的攻击成功率

对UNSW-NB15数据集，DNN-256等分类模型的初始分类情况如表5所示。

表5　UNSW-NB15数据集初始分类情况

分类模型	准确率	精确率	召回率	F1分数
DNN-256	0.939 5	0.908 1	0.930 1	0.919 5
DNN-128	0.924 8	0.924 9	0.924 7	0.937 0
DNN-64	0.929 9	0.929 1	0.930 0	0.929 0
RF	0.932 2	0.932 2	0.932 2	0.931 7
KNN	0.916 1	0.916 3	0.916 1	0.916 2
LR	0.853 0	0.875 8	0.853 0	0.855 6
SVM	0.907 0	0.911 2	0.907 0	0.904 7
DT	0.836 5	0.844 4	0.836 5	0.838 5
AdaBoost	0.923 4	0.923 1	0.923 4	0.923 2

由表5可以看出，除LR和DT外，其他分类模型的各项评价指标均达到了0.9以上，LR和DT的准确率、精确率、召回率以及F1分数等评价指标也分别达到了0.83和0.85左右，即上述分类模型均可较好地完成UNSW-NB15数据集的入侵检测任务。

取DNN-256作为算法1的流量分类模型f，将预处理后的UNSW-NB15测试集作为算法1的干净样本集D_{clean}(UNSW-NB15)，将f和D_{clean}(UNSW-NB15)输入算法1后，得到扰动特征及扰动后的特征值如表6所示。

表6　UNSW-NB15扰动特征情况

扰动特征	特征重要性均值	扰动后特征值
swin	1.289	255
ct_state_ttl	0.967	0
ct_dst_sport_ltm	0.753	1
dwin	0.440	255
dmean	0.356	0
ct_dst_src_ltm	0.278	65
ct_srv_src	0.275	63

对D_{clean}(UNSW-NB15)和分类模型f，依次将表6结果作为对抗流量的扰动特征，并保持其他特征不变，设置不同的扰动特征数，得到对应的对抗样本集$D_{attacked}$(UNSW-NB15)（其规模和D_{clean}(UNSW-NB15)相同）。利用分类模型f对$D_{attacked}$(UNSW-NB15)进行分类，得到对抗样本攻击成功率与扰动特征数的关系如图6所示。

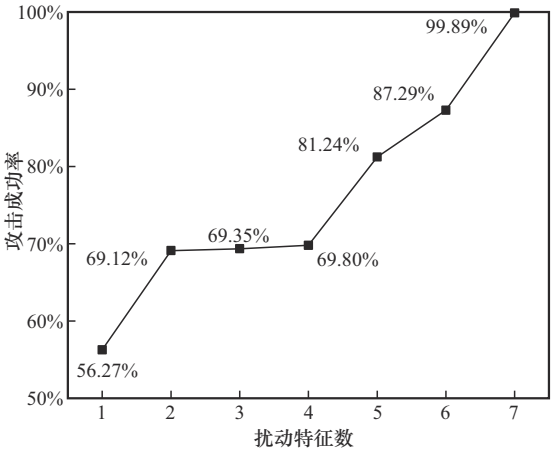


图6　UNSW-NB15数据集扰动特征数和攻击成功率的关系

由图6可以看出，在UNSW-NB15数据集上，攻击成功率随着扰动特征数的提高而提高。当扰动特征数为5时，攻击成功率达到了81.24%。当扰动特征数增加到7时，攻击成功率达到了99.89%。此外，图6与图3的趋势有着较为明显的差异，其原因应是数据集本身的特征重要性差异。对NSL-KDD数据集，其前5个重要扰动特征每个都对NSL-KDD数据集的分类影响较大。对UNSW-NB15数据集，其第3个重要扰动特征只有和后面的特征共同作用时，才能大幅提升攻击成功率。

将本文方法与其他方法[38-39]进行对比，结果如表7所示。

表7　UNSW-NB15数据集攻击成功率对比

方法	攻击成功率	扰动特征数
本文方法	93.24%	7
NIDSGAN	87.89%	7
PSO	44.43%	29
GA	77.87%	29
GAN	86.18%	29

由表7可以看出，对UNSW-NB15数据集，当扰动特征数为7时，本文方法的攻击成功率高于NIDSGAN方法，且在扰动特征数更少的情况下，成功率高于PSO、GA和GAN方法，意味着本文方法可以在更隐蔽性的情况下实施更有效的攻击。

选取扰动特征数为7，对DNN-128、DNN-64、RF、KNN、LR、SVM、DT、AdaBoost等分类模型进行迁移攻击，迁移攻击成功率如图7所示。

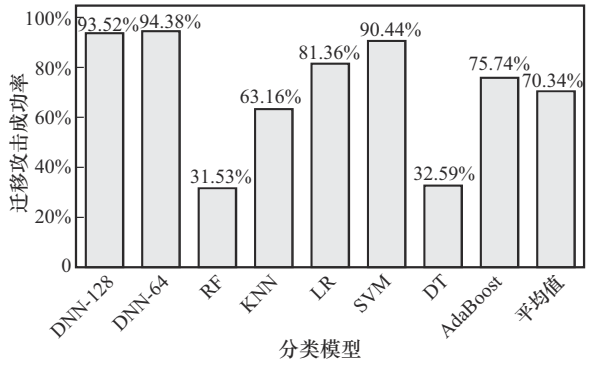


图7　UNSW-NB15数据集不同分类模型的迁移攻击成功率

由图7可以看出，由DNN-256分类模型生成的对抗流量对上述模型具有较高的迁移攻击成功率，其中对DNN-128和DNN-64的迁移攻击成功率最高，原因是这和NSL-KDD数据集的结果相似。此外，对抗流量对RF和DT的迁移攻击成功率较低，成功率略高于30%，这可能是由于RF和DT分类模型的内在逻辑与DNN-256相差较大，DNN-256的重要特征对RF和DT贡献不大，在攻击时仅扰动了对RF和DT贡献较小的特征，从而攻击效果不佳。总体上，对8种分类模型的平均迁移攻击成功率为70.34%，说明本文方法在UNSW-NB15数据集上同样具有较强的迁移性。

在防御实验方面，取DNN-256等分类模型作为算法2的流量分类模型f；将预处理后的UNSW-NB15测试集作为算法2的干净样本集D_{clean}(UNSW-NB15)。对于DNN-256、DNN-128、DNN-64等分类模型，训练迭代次数N_{iter}设置为50；对于RF、KNN、LR、SVM、DT、AdaBoost等不需要设置训练迭代次数的分类模型，正常进行训练。对抗训练中对抗样本占比P设置为0.01，得到防御后的攻击成功率如图8所示。

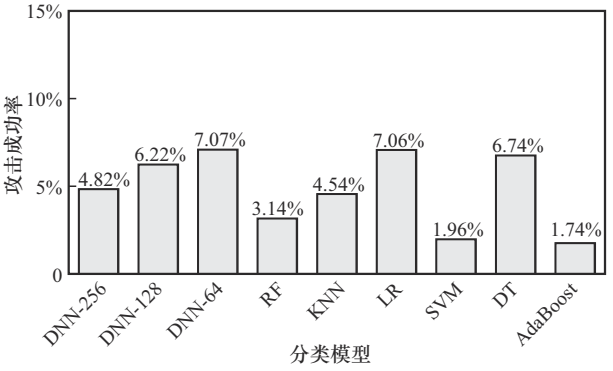


图8　UNSW-NB15数据集防御后的攻击成功率

由图8可以看出，使用算法2进行防御后，针对各个分类模型的攻击成功率明显下降，所有分类模型的攻击成功率均在8%以下。其中，SVM和AdaBoost的攻击成功率均下降至2%以下，分别达到了1.96%和1.74%。DNN-256、RF和KNN的攻击成功率分别为4.82%、3.14%和4.54%。DNN-128、DNN-64、LR及DT的攻击成功率略高于5%，但最高值也仅为DNN-64的7.07%。综上所述，本文所提防御方法可以在UNSW-NB15数据集上达到较好的防御效果，各种流量分类模型在经过算法2防御后，面对对抗样本的鲁棒性明显加强。

4　结束语

本文利用XAI技术实现了流量对抗样本攻击，利用XAI确定扰动特征，保证了扰动的高效性，从而以少量扰动特征实现了较高的攻击成功率，对NSL-KDD数据集，仅需扰动5个特征便可以达到92.71%的攻击成功率，对UNSW-NB15数据集，扰动7个特征可以达到99.89%的攻击成功率。同时，关键扰动特征对不同分类器的一致性保证了攻击样本的较强迁移性，在2个数据集上8种不同分类器的平均迁移攻击成功率均高于70%，说明本文攻击

方法具有较强的迁移性。在防御方面，本文提出了一种基于对抗训练的防御方法，在2个数据集上均实现了较好的防御效果，提升了流量分类器面对对抗样本攻击的鲁棒性。在未来工作中，将针对以下2个方面进行进一步研究：1) 利用LIME等其他模型解释方法，对多种流量分类器进行对抗样本攻击，以实现更好的攻击效果；2) 进一步考虑对抗样本防御问题，从增加数据检测、提升对抗训练效率、对抗样本恢复等方面抵御针对NIDS的对抗样本攻击，提升NIDS的鲁棒性。

参考文献：

[1] GARCÍA-TEODORO P, DÍAZ-VERDEJO J, MACIÁ -FERNÁNDEZ G, et al. Anomaly-based network intrusion detection: techniques, systems and challenges[J]. Computers & Security, 2009, 28(1/2): 18-28.

[2] WHITEHEAD D E, OWENS K, GAMMEL D, et al. Ukraine cyber-induced power outage: analysis and practical mitigation strategies[C]// Proceedings of the 2017 70th Annual Conference for Protective Relay Engineers (CPRE). Piscataway: IEEE Press, 2017: 1-8.

[3] SHONE N, NGOC T N, PHAI V D, et al. A deep learning approach to network intrusion detection[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 2(1): 41-50.

[4] NANDANWAR H, KATARYA R. Deep learning enabled intrusion detection system for Industrial IoT environment[J]. Expert Systems with Applications, 2024, 249: 123808.

[5] BOSE I, MAHAPATRA R K. Business data mining: a machine learning perspective[J]. Information & Management, 2001, 39(3): 211-225.

[6] KOUROU K, EXARCHOS T P, EXARCHOS K P, et al. Machine learning applications in cancer prognosis and prediction[J]. Computational and Structural Biotechnology Journal, 2015, 13: 8-17.

[7] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 1026-1034.

[8] VASSILEV A, OPREA A, FORDYCE A, et al. Adversarial machine learning: a taxonomy and terminology of attacks and mitigations[R]. 2024.

[9] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.

[10] YUAN X Y, HE P, ZHU Q L, et al. Adversarial examples: attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. axXiv Preprint, arXiv: 1412.6572, 2014.

[12] ZHANG W E, SHENG Q Z, ALHAZMI A, et al. Adversarial attacks on deep-learning models in natural language processing: a survey[J]. ACM Transactions on Intelligent Systems and Technology, 2020, 11(3): 1-41.

[13] CARLINI N, WAGNER D. Audio adversarial examples: targeted attacks on speech-to-text[C]//Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW). Piscataway: IEEE Press, 2018: 1-7.

[14] SUN J W, ZHANG T W, XIE X F, et al. Stealthy and efficient adversarial attacks against deep reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 5883-5891.

[15] 陈晓霖, 昝道广, 吴炳潮, 等. 面向纵向联邦学习的对抗样本生成算法[J]. 通信学报, 2023, 44(8): 1-13.
CHEN X L, ZAN D G, WU B C, et al. Adversarial sample generation algorithm for vertical federated learning[J]. Journal on Communications, 2023, 44(8): 1-13.

[16] SAEED W, OMLIN C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities[J]. Knowledge-Based Systems, 2023, 263: 110273.

[17] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[J]. axXiv Preprint, arXiv: 1705.07874, 2017.

[18] WU W B, SU Y X, CHEN X X, et al. Boosting the transferability of adversarial samples via attention[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Pres, 2020: 1161-1170.

[19] WANG Z B, GUO H C, ZHANG Z F, et al. Feature importance-aware transferable adversarial attacks[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 7619-7628.

[20] ZHANG J P, WU W B, HUANG J T, et al. Improving adversarial transferability via neuron attribution-based attacks[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 14973-14982..

[21] ZHANG J P, HUANG Y Z, XU Z E, et al. Improving the adversarial transferability of vision transformers with virtual dense connection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(7): 7133-7141.

[22] RAHBARINIA B, PERDISCI R, ANTONAKAKIS M. Segugio: efficient behavior-based tracking of malware-control domains in large ISP networks[C]//Proceedings of the 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Piscataway: IEEE Press, 2015: 403-414.

[23] ONGUN T, BOBOILA S, OPREA A, et al. CELEST: federated learning for globally coordinated threat detection[J]. axXiv Preprint, arXiv: 2205.11459, 2022.

[24] GU J, LU S. An effective intrusion detection approach using SVM with naïve Bayes feature embedding[J]. Computers & Security, 2021, 103: 102158.

[25] HOSSEINI S, ZADE B M H. New hybrid method for attack detection using combination of evolutionary algorithms, SVM, and ANN[J]. Computer Networks, 2020, 173: 107168.

[26] 陈红松, 陈京九. 基于循环神经网络的无线网络入侵检测分类模型

构建与优化研究[J]. 电子与信息学报, 2019, 41(6): 1427-1433.
CHEN H S, CHEN J J. Recurrent neural networks based wireless network intrusion detection and classification model construction and optimization[J]. Journal of Electronics & Information Technology, 2019, 41(6): 1427-1433.

[27] SHI L, ZHU H, LIU Y, et al. Intrusion detection of industrial control system based on correlation information entropy and CNN-BiLSTM[J]. Journal of Computer Research and Development. 2019, 56(11): 2330-2338.

[28] CAO B, LI C H, SONG Y F, et al. Network intrusion detection model based on CNN and GRU[J]. Applied Sciences, 2022, 12(9): 4184.

[29] CHEN H S, LI X Y, LIU W M. Multivariate time series anomaly detection by fusion of deep convolution residual autoencoding reconstruction model and ConvLstm forecasting model[J]. Computers & Security, 2024, 137: 103581.

[30] RIGAKI M, Elragal A. Adversarial deep learning against intrusion detection classifiers [C]//017 NATO IST-152 Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience, IST-152 2017; Czech Technical UniversityPrague; Czech Republic; 18-20 October 2017. CEUR-WS, 2017, 2057: 35-48.

[31] IBITOYE O, SHAFIQ O, MATRAWY A. Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks[C]// Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2019: 1-6.

[32] SHEATSLEY R, PAPERNOT N, WEISMAN M J, et al. Adversarial examples for network intrusion detection systems[J]. Journal of Computer Security, 2022, 30(5): 727-752.

[33] DING R Y, SUN L, ZANG W F, et al. Towards universal and transferable adversarial attacks against network traffic classification[J]. Computer Networks, 2024, 254: 110790.

[34] ROSHAN K, ZAFAR A, HAQUE S B U. Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system[J]. Computer Communications, 2024, 218: 97-113.

[35] CHENG Q M, ZHOU S Y, SHEN Y, et al. Packet-level adversarial network traffic crafting using sequence generative adversarial networks[J]. axXiv Preprint, arXiv: 2103.04794, 2021.

[36] CATILLO M, PECCHIA A, REPOLA A, et al. Towards realistic problem-space adversarial attacks against machine learning in network intrusion detection[C]//Proceedings of the 19th International Conference on Availability, Reliability and Security. New York: ACM Press, 2024: 1-8.

[37] LU H Y, LIU J J, PENG J M, et al. Adversarial attacks based on time-series features for traffic detection[J]. Computers & Security, 2025, 148: 104175.

[38] ZOLBAYAR B E, SHEATSLEY R, MCDANIEL P, et al. Generating practical adversarial network traffic flows using NIDSGAN[J]. axXiv Preprint, arXiv: 2203.06694, 2022.

[39] ALHAJJAR E, MAXWELL P, BASTIAN N. Adversarial machine learning in network intrusion detection systems[J]. Expert Systems with Applications, 2021, 186: 115782.

[40] HE K, KIM D D, ASGHAR M R. Adversarial machine learning for network intrusion detection systems: a comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2023, 25(1): 538-566.

[41] BUCKMAN J, ROY A, RAFFEL C, et al. Thermometer encoding: one hot way to resist adversarial examples[C]//Proceedings of the International Conference on Learning Representations. Piscataway: IEEE Press, 2018.

[42] LI H Y, SHAN S, WENGER E, et al. Blacklight: scalable defense for neural networks against query-based black-box attacks[C]//USENIX Security Symposium. Berkeley: USENIX Association, 2022: 2117-2134.

[43] XU H, LIU X R, LI Y X, et al. To be robust or to be fair: towards fairness in adversarial training[C]//Proceedings of the International Conference on Machine Learning. Saarland: DBLP, 2021: 11492-11501.

[44] DEBICHA I, BAUWENS R, DEBATTY T, et al. TAD: transfer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems[J]. Future Generation Computer Systems, 2023, 138: 185-197.

[45] 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.

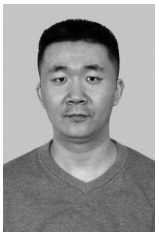
[46] 刘奇旭, 王君楠, 尹捷, 等. 对抗机器学习在网络入侵检测领域的应用[J]. 通信学报, 2021, 42(11): 1-12.
LIU Q X, WANG J N, YIN J, et al. Application of adversarial machine learning in network intrusion detection[J]. Journal on Communications, 2021, 42(11): 1-12.

[47] TAVALLAEE M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]//Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Piscataway: IEEE Press, 2009: 1-6.

[48] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set) [C]// Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS). Piscataway: IEEE Press, 2015: 1-6.

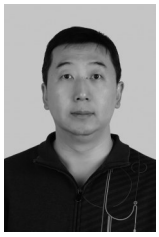
[49] LEE W K, STOLFO S J. A framework for constructing features and models for intrusion detection systems[J]. ACM Transactions on Information and System Security, 2000, 3(4): 227-261.

[作者简介]



马博文（1992– ），男，河南驻马店人，信息工程大学助理研究员，主要研究方向为人工智能安全、网络攻防。

郭渊博（1975– ），男，陕西周至人，博士，海南大学教授、博士生导师，主要研究方向为大数据安全、态势感知。



马骏（1981– ），男，河北安国人，博士，信息工程大学副教授、硕士生导师，主要研究方向为态势感知、网络攻防。

田继伟（1993– ），男，河南西平人，博士，空军工程大学讲师，主要研究方向为人工智能安全、网络攻防。

胡永进（1981– ），男，山东潍坊人，博士，信息工程大学讲师，主要研究方向为主动防御、态势感知。