# 评估和提高基于机器学习的网络入侵检测器的对抗鲁棒性

## 摘要

机器学习（ML），特别是深度学习（DL）技术在基于异常的网络入侵检测系统（NIDS）中得到了越来越多的应用。然而，ML/DL已被证明极易受到敌对攻击，尤其是在此类安全敏感系统中。许多对抗性攻击被提出来评估基于MLS的网络入侵检测系统的鲁棒性。不幸的是，现有的攻击大多集中在特征空间和/或白盒攻击上，这些攻击在现实场景中做出了不切实际的假设，在实践中未能充分证明系统鲁棒性的上限。

为了弥补这一差距，我们首次对灰盒/黑盒攻击进行了系统研究，以评估基于ML的网络入侵检测系统的鲁棒性。我们的工作在以下几个方面优于以前的工作：（i）实用性：所提出的攻击可以在保留其功能的同时，以极其有限的知识和可承受的开销自动变异原始流量；（ii）泛型——所提议的攻击对于使用不同的ML/DL模型和基于非有效载荷的特征评估各种NIDS的鲁棒性是有效的；（iii）可解释性我们提出了一种基于ML的网络入侵检测系统脆弱鲁棒性的解释方法。在此基础上，我们还提出了一种对抗攻击的防御方案，以提高系统的鲁棒性。我们使用不同的特征集和ML/DL模型广泛评估各种NIDS的健壮性。实验结果显示了我们的攻击和防御方案的有趣发现和有希望的结果。

索引术语：网络异常检测、网络入侵检测系统、对抗性机器学习、机器学习安全、规避攻击。

1. 导言

网络入侵检测系统（NIDS）在检测网络中的恶意活动方面起着至关重要的作用。根据检测机制，NIDS通常可分为两种类型[1]。基于签名的入侵检测系统传统上用于检测已知的恶意流量，但无法检测新的攻击或高级变体。相比之下，基于异常的入侵检测系统，结合机器学习（ML）和最近的深度学习（DL）技术，由于具有检测新攻击和未知攻击的泛化能力，正受到越来越多的关注[2]-[5]。

虽然ML和DL模型对其巨大的潜力和性能很感兴趣，但事实证明，它们极易受到敌对攻击。精心设计的输入小变化可能会导致ML模型的输出发生重大变化[6]–[8]。为了评估基于ML的系统的鲁棒性，广泛使用的方法是构造对抗性攻击，证明系统鲁棒性的上限[9]。此类攻击必须充分反映对手在尽可能实际的场景中的能力，否则将毫无用处。

在其他领域，如计算机视觉[9]–[11]、自然语言处理[12]和恶意软件检测[13]–[16]，已经提出了几种攻击来评估基于ML的系统的鲁棒性。然而，由于两个原因，这些方法不适合于网络入侵检测。首先，这种攻击是修改特征向量（我们称之为特征空间攻击），而不是真实的输入空间（即NIDSs的原始流量）。这是因为在其他领域，从输入空间到特征空间的特征映射是可逆的，或者可以很容易地形式化为可微函数。相比之下，从网络流量到NIDS特征的特征提取既不可逆也不可微。因此，针对NID的特征空间攻击在实践中是不切实际的，因为对抗性流量很难从特征中派生。其次，我们需要确保在修改恶意流量时没有通信冲突或恶意危害，这对于其他非安全域来说不是问题。

对于针对基于ML的网络入侵检测系统的流量空间攻击（即直接修改流量），现有的研究可分为三个部分。首先，对传统网络入侵检测系统的规避攻击已经得到了很好的研究[17]–[19]，但它们只适用于没有学习模型的基于签名的网络入侵检测系统。其次，为了避免基于流量分析的检测，提出了几种流量模糊和变异方法[20]-[22]。然而，他们只是简单地模仿良性流量或随机变异，而不是利用具有对抗性的机器学习技术的ML/DL模型的漏洞。因此，这样的攻击是非常昂贵和低效的，但攻击者总是考虑攻击开销在实践中。第三，利用交通空间对抗性攻击的唯一工作[23]是基于充分了解NID的强大白盒假设，这是不切实际的，因为目标NID在现实场景中总是无法访问。

总之，基于ML的NIDSs上现有的对抗性攻击（无论是featurespace还是traffic space）由于其不切实际的设想，未能证明鲁棒性的上限。为了弥合这一差距，我们面临三个方面的挑战：

实用性。如何使用极其有限的知识和可承受的开销来执行保护功能的交通空间（不可逆特征提取下）对抗攻击？

通用性。如何提出一个通用框架，有效地评估使用各种ML模型和功能的NIDSs的健壮性？

可解释性。如何解释脆弱性，进而提高基于ML的网络入侵检测系统对此类攻击的鲁棒性？

在本文中，我们提出了一种新的实用对抗攻击，将其表述为具有**两步解的双层优化问题**，该算法首先以较小的潜在流量空间变异开销解决所谓的**对抗性特征**，然后以最接近对抗性特征的特征空间距离**搜索最佳流量变量**。（这解决了可负担的开销挑战）。在第一步（即解决对抗性特征），我们扩展了先前使用生成性对抗性网络（GAN）的思想，将目标ML模型视为一个**黑箱**，并搜索位于代理模型低置信区域的对抗性特征。基于敌对攻击的可转移属性[24]，攻击效果可以转移到其他ML模型。（这涉及有限的知识和模型普遍性挑战）。在第二步（即搜索最佳流量变量），我们提出了一个**启发式数据包**制作框架来自动变异恶意流量。基于基于MLS的NIDSs中特征提取的领域知识，我们提出了几种流量变异算子，这些算子可以在不破坏恶意功能的情况下影响所有总结的特征（这是针对功能保留和特征通用性的挑战）。为了解决不可逆特征提取问题，我们提出了一种称为meta-info 信息向量的可逆抽象流量表示。为了解决可解释性的挑战，我们通过量化攻击者操纵每个特征的程度，提出了一种基于ML的网络入侵检测系统脆弱性的可解释方法。基于此，我们提出了一种防御方案，通过去除最脆弱的特征来提高系统的鲁棒性。

贡献。我们的主要贡献包括提出了一种新的对抗性攻击框架，并广泛评估了基于ML的NIDSs的鲁棒性，随后解释和改进了针对此类攻击的鲁棒性。具体而言，它们的阐述如下：

* 在灰色和黑箱假设下，我们提出了第一个实际的基于ML的NIDSs的交通空间对抗攻击。与以前的对手攻击相比，我们的威胁模型提供了一个**更实用的攻击者能力假设，并考虑了攻击开销**。与制作测试时间规避流量的传统方法相比，我们的攻击不仅仅是模仿正常流量，而是利用**对抗性机器学习技术**来搜索位于ML模型决策边界处的特征，能够自动搜索规避流量突变，并且能够适应处理不同的恶意流量。
* 我们提出了一种基于ML的NIDS脆弱鲁棒性的解释方法。在此基础上，提出了一种对抗性攻击的防御方案，以提高系统的鲁棒性。
* 我们广泛评估了我们在最先进的NIDSKitsune上的攻击和防御方法，以及各种基于MLS的NIDSs，**包括六种典型的ML分类器（包括统计模型和深度神经网络）和两种特征集（数据包级和流级）。**我们还通过分析探讨了攻击背后的重要见解。

本文的其余部分组织如下：我们在第二节总结了相关工作，在第三节提供了背景和动机。第四节介绍了威胁模型和问题陈述，以及我们的攻击框架的动机和概述。第五节和第六节阐述了我们进攻的两个步骤。第七节提供了防御方案。实验结果和发现见第八节。我们在第九节讨论了限制和改进。第十节总结了本研究。

## 相关工作

在这项工作中，我们通过演示构造性攻击来评估鲁棒性。下面对相关工作中存在的攻击进行分类和介绍。

**逃避对NIDSs的攻击。**对（N）IDS本身的规避攻击已经得到了广泛的研究[17]、[25]、[26]。关于规避基于签名的系统[18]、[19]、[27]、[28]和传统的基于异常的系统[29]-[33]已经进行了广泛的研究。然而，这种规避攻击只适用于没有学习模型的基于签名的网络入侵检测系统。

**对基于ML的系统进行对抗性攻击。**在其他领域，针对基于ML的系统的对抗性攻击已经有了一些工作。计算机视觉领域中的对抗性示例已被广泛研究[9]、[10]、[34]、[35]。梯度下降法[13]和遗传编程（GP）[14]用于规避PDF恶意软件分类器。MalGAN[15]被提议生成对抗性恶意软件示例，以规避基于ML的恶意软件检测系统。基于GAN的方法也被用于欺骗实时视频分类系统[11]。文本情感分析系统被随机优化方法规避[12]。然而，由于网络流量和NIDS的特殊性，这些方法不能直接应用。

**对基于ML的NIDSs进行对抗性攻击。**我们提出了对基于ML的NIDSs进行对抗性攻击的术语，并在表I中列出了现有攻击的不切实际假设。其具体定义和相关工作如下：

1. 特征空间攻击。对规避基于ML的NIDS的特征空间攻击假定攻击者可以直接修改特征向量。根据攻击者对目标NID的了解，现有的功能空间攻击可分为三类：

* 特征空间白盒攻击（FW A）。FW A要求全面了解目标NID。在[36]中，四种基于梯度的对抗性示例攻击被直接用于规避MLP分类器。同样地，在[37]中，对抗性的例子也被用来逃避Kitsune。[38]中也使用了类似的基于梯度的方法来攻击物联网网络的网络入侵检测系统。
* 特征空间灰盒攻击（FGA）。FGA需要目标分类器的反馈（与FWA相比，没有其他分类器知识）。在[39]中，提出了一种基于GAN的体系结构IDS-GAN，用于生成规避特征。在[40]中，提出了一种基于边界的方法，通过扰动特征来规避DoS检测系统。
* 特征空间黑箱攻击（FBA）。FBA既不需要反馈，也不需要分类器的任何知识。在[41]中，随机修改了四个特征以避开僵尸网络检测器。

然而，特征空间攻击是不切实际的，因为基于ML的NIDSs中的特征提取总是不可逆的。

1. 交通空间攻击。直接改变网络流量的现有攻击可分为三类：

交通空间白盒攻击（TWA）。在[23]中，提出了一种使用与我们类似的变异算子的白盒攻击。然而，他们认为攻击者完全了解NIDS的假设在实践中很难实现。

随机突变。在[20]中提出了几种变异以逃避僵尸网络检测器。[22]中提出了对流量的随机混淆。然而，这些方法纯粹是随机的，缺乏理论指导

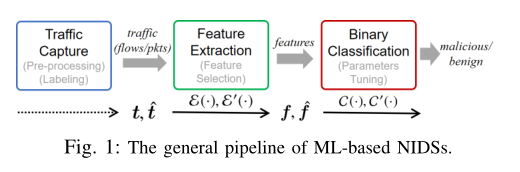
交通混乱。为了避免基于流量分析的检测，已经提出了几种流量混淆方法，如[21]。然而，他们只是简单地模仿良性流量和随机变异，而没有利用利用对抗性机器学习技术的ML/DL模型的漏洞。

## 背景和动机

1. 有针对性的基于ML的网络入侵检测系统

在这项研究中，我们重点评估了基于ML的NIDS非有效负载特性的鲁棒性。这类目标系统介绍如下：

基于ML的NIDSs。通常，基于ML的NIDS由流**量捕获、特征工程和分类**组成，如**图1**所示。首先，从监控网络捕获流量。然后从原始流量中提取特征集，进行选择，最后将其输入到ML分类器中进行训练或检测。请注意，基于ML的NIDSs首先需要数据预处理和从原始流量中手动提取特征，而不是仅仅依靠深度学习模型进行自动特征学习。这是因为网络流量由于同时具有文本和时间特征而更加复杂，并且其非结构化自由文本很难直接表示为连续值变量。



非有效载荷特性。在这项研究中，我们着重于避免基于ML的NIDS，在这种NIDS中，数据包的有效负载没有被检查（称为基于非有效负载）。我们认为这是合理的，因为考虑到两个因素：首先，我们发现大多数基于ML的NIDS可能会使用非基于有效负载的功能，因为当前检查有效负载很重，甚至不可能用于加密流量。其次，基于规避有效载荷的异常NIDSs已经得到了很好的研究[32]，[33]。请注意，非基于payloadbased的NIDS更喜欢检测依赖于卷和/或迭代的攻击，如DoS/DDoS（分布式拒绝服务）、扫描、暴力攻击和机器人/僵尸网络。与特定内容（如SQL注入）相关的其他攻击不在此类NIDS的范围内。

1. 现有攻击的模拟和限制

为了便于说明，我们使用函数表示从一系列相关流量中提取到特征向量，并使用函数C表示以特征向量作为输入并输出恶意概率的ML分类器。在任何特定时间，我们用t和t’表示用于提取两个特征向量f和f’的两系列相关原始和变异恶意流量，分别是和'。如表1所示，我们认为，由于以下不切实际的假设，现有攻击未能充分证明现实环境中系统鲁棒性的上限：

1）特征空间变异（在FWA/FGA/FBA中）。许多以前的研究仅仅是在求解时发现回避特征，它直接修改特征的值，而不考虑如何变异流量。然而，在基于ML的NIDSs中，特征提取是不可逆的。

2）白盒知识（在TW A/FW A/FGA中）。一些工作假设分类器的特定输出是可以实现的，那么规避攻击可以被视为解决一个优化问题： 。然而，实际上，攻击者总是无法访问NIDS。

3）无限的变异开销。现有的攻击不会限制攻击者的开销或修改流量的能力（例如，关于流量或时间延迟）。然而，攻击者总是考虑攻击开销在实践中，流量不太可能被任意修改。

如果对抗性攻击不够实用且过于简单，则鲁棒性的评估将是不正确和误导的。因此，我们通过放松特征空间和白盒假设以及增加开销约束，提出了一种更为实用的对抗性攻击。

## 攻击方法

在本节中，我们首先在更实际的环境中定义威胁模型。然后，我们将规避攻击公式化为一个优化问题，然后介绍我们如何解决这个问题。

1. 威胁模型

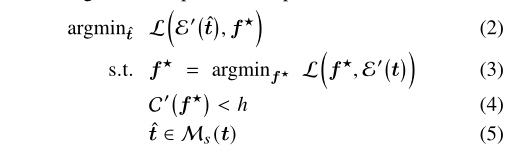
我们认为攻击者开始与恶意意图的一系列流量，并希望逃避基于ML的NIDS使用非基于负载的特征。（这种有针对性的对抗性攻击也称为贬值攻击）。与以前的白盒/灰盒攻击不同，攻击者不需要任何关于目标分类器及其输出标签或概率的知识。与以前的功能空间攻击不同，攻击者只能以可承受的开销变异他/她控制的设备生成的原始流量（即流量空间攻击）。此外，根据不同攻击者对目标NID功能的了解，可以执行两种攻击：

* 实用灰盒攻击（PGA）。这假设攻击者知道目标NID中使用的功能。在其他情况下，攻击者可以构建与目标NID相同的功能提取器，并使用它精确提取功能。这似乎有些极端，但基于ML的NIDSs中使用的特性经常被发布[5]，[42]–[45]。
* 实用黑盒攻击（PBA）。我们假设一个更实际的情况，在这种情况下，攻击者对目标NID中使用的功能了解非常有限，甚至一无所知。在这种情况下，攻击者只能基于域知识构建代理提取器。

1. 实际交通空间规避攻击问题

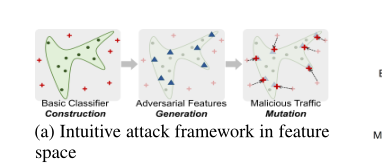
根据威胁模型，首先，我们放松了第三节B中的两个假设，通过训练一个具有概率输出的替代分类器C’（·），使其近似于C（·）。这也解决了一些没有连续输出值的ML模型（如隔离林）难以优化的问题。其次，我们构建了代理特征提取器（·）。对于PGA，（·）与（·）完全相同，而对于PBA则是模拟的。此外，我们表示变异操作M（·），它可以将原始流量t转换为一个由所有可能的变异流量t’组成的集合。如果变异可以保留t时的恶意功能，我们说变异操作是安全的（用表示）。此外，我们还考虑了**额外开销项的变异开销预算**以最小化距离(用来表示 带权重系数) 在t’和t之间。因此，规避攻击问题可以表示为解决：



问题（1）是直观的，但由于高度非线性约束、不可微性（·）和难以定义，因此很难解决受对抗式机器学习思想的启发，我们将规避攻击问题转化为寻找特殊的变异流量，从而使目标ML分类器错误分类，同时最小化变异流量的开销。因此，问题（1）转化为以下双层优化问题：

其中L（·，·）是两个特征向量之间的距离度量，并且ℎ是分类的异常阈值。等式（4）表示f★被归类为良性。

我们解决上述问题的高层思路是分别求解上下层目标函数：首先在约束（4）下，求解在（3）中的f★，然后使用解决的f★在约束（5）下搜索t’（2）。为了直观起见，图2a从特征空间的角度描述了这两个步骤。

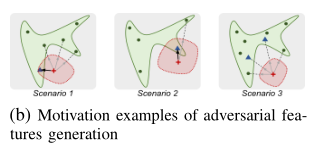


预先对目标分类器进行训练，以区分良性和恶意特征的分布。首先，对于每个恶意特征，f★ 产生的特征不仅可以划分为良性特征，而且在特征值方面尽可能接近恶意特征；我们把这些特征称为对抗性特征。一般来说，f★位于分类器的低置信度区域。生成对抗性特征f★这就是问题（3）要解决的问题。第二，原始恶意流量经过变异，将其特征转移到最接近的对抗/良性流量，这就是问题（2）要解决的问题。然而，由于（4）和（5）中的高度非线性约束和不可微性，这两个问题仍然难以解决。（t’在第（2）条中很难驾驶）。接下来介绍这两个问题的具体解决方案。

1. 问题解决的动机和概述

我们现在介绍上述两个步骤中使用的技术，以及使用它们解决双层优化的动机。具体地说，我们使用增强的生成性对抗网络（GAN）模型来生成对抗特征（在（3）中），并使用粒子群优化（PSO）来搜索规避流量突变（在（2）中）。现在我们介绍采用GAN和PSO的动机，以及如何将它们结合起来完成规避攻击。

为什么会产生对抗性特征f★?如前所述，我们观察到攻击者所谓的可变流量的预算能力在实践中总是有限的。一方面，攻击者可能会有额外的开销（例如额外的时间和精心编制的流量来逃避检测）。另一方面，攻击者修改流量的能力也受到限制。例如，过度增加间隔时间将导致连接超时，注入过多的流量将占用大量带宽，可能会被受害者感知到。我们的主要观察结果是，交通空间中的变异开销与特征空间中的距离相关（详情和证据见附录A）。因此，我们寻找f★ 在特征空间中减少变异流量的开销。图2b提供了示例以证明关闭的必要性★. 如果没有对抗性特征的指导，恶意特征可能最终无法到达最近的良性空间（场景1）或错过瞬时良性空间（场景2）。

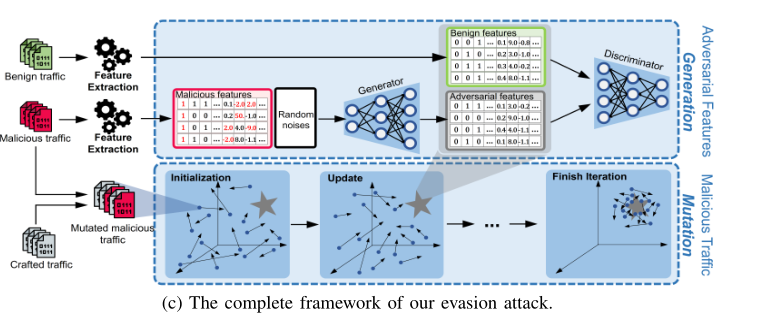


为什么是GAN？对抗性特征生成需要：

1. 模型不可知我们假设攻击者不知道ML分类器；
2. 在实践中可能存在大量恶意/流量。GAN[46]由两个神经网络（生成器和鉴别器）组成，它们相互竞争以完成一个最小-最大博弈。受之前想法[15]，[39]的启发，GAN非常有能力生成对抗性特征，因为1）鉴别器可以作为**目标分类器的替代品**进行训练，从而可以进行模型不可知攻击；2） 一旦生成器经过培训，它就可以生成f★ 有效地防止任何恶意功能。

为什么PSO？网络流量很难直接参与数值计算，所以我们将流量矢量化为高维向量，其中包含数据包头的元信息。注意，从流量到矢量的矢量化是可逆的。因此，问题（2）变成了寻找最佳元信息向量。然而，和连续特征空间不同，元信息向量的每个维度都有不同的离散值，因此问题（2）确实是一个困难的组合优化任务（NP-完全）。因此，我们求助于群体智能算法来寻找近似解。我们采用PSO[47]，因为它是一种简单但功能强大的方法，在处理高维任务时具有很强的适应性。

它们是如何协同工作的？简言之，GAN生成了PSO的优化目标。在基于粒子群优化算法对恶意流量进行变异时，我们通过测量其与GAN生成的对抗特征的相似性，找到具有最佳规避效果的流量变异。具体而言，所提出的规避攻击方法如图2c所示，包括以下两个步骤：



对抗性特征生成：我们假设攻击者想要启动一些活动，这将导致一系列恶意流量。首先，攻击者需要在其控制的网络中收集一些有益的流量。然后，通过代理提取器将两种流量提取为特征，并输入到我们的GAN模型中。在训练阶段之后，生成器能够生成对抗性特征。

恶意流量变异：在生成对抗性特征后，我们使用带有预定义操作符的PSO自动变异恶意流量。群中的每个粒子代表一个向量，由变异流量的元信息组成。该群体在特征与敌对特征最相似的临时最佳粒子的引导下，对交通空间进行迭代搜索。最后，经过多次迭代选择最佳粒子。上述两个步骤的细节将分别在接下来的第五节和第六节中详细阐述。

## 生成对抗特征

我们现在介绍生成对抗特征的过程。我们的增强型GAN模型如顶部图2c所示，它由一个生成器和一个鉴别器组成。

生成器。生成器是一个前馈神经网络，其目的是将恶意特征转化为其对抗特征。它将恶意特征向量f和分布中的噪声向量z按照分布串联在一起作为输入并输出生成的特征向量，表示为G（f，z）。为了训练生成器，其损耗函数在（6）中定义为：



其中是原始恶意特征的集合。应尽量减少生成器网络中的重量。在本研究中，我们通过额外计算输入和输出之间的构造误差L（.,.），扩展了先验GANs。在本研究中，我们使用均方根误差，其中是输入和输出特征的维度。因此，生成的特征可以模拟良性特征的分布，同时接近恶意特征。

鉴别器。鉴别器用于区分生成的特征和良性特征。它也是一个前馈神经网络，其输入由上述两类特征向量组成，其输出是确定生成输入向量的概率。训练鉴别器使生成的输入向量的输出最大化，同时使良性输入向量的输出最小化。因此，其损失函数为：



其中，是从攻击者事先控制的网络中收集的良性流量中提取的一组特征。是由生成器生成的特征集。训练过程是生成器和鉴别器之间的迭代和相互优化，直到收敛。然后，生成器生成的特征可以作为对抗性特征。

## 变异恶性流量

在本节中，我们将介绍如何通过启发式方法自动改变流量。首先，我们设计了恶意流量的变异算子。然后，我们介绍了从流量到元信息向量的矢量化。最后，提出了基于粒子群算法的流量变异算法。

1. 基本流量变异算子

为了解决规避流量变异问题，我们首先设计了一些基本的参数化变异算子。变异算子应能够影响尽可能多的特征类型，以使其具有通用性。此外，变异算子应保持功能性(重铸（5）)和隐藏以防止被受害者察觉。

然而，当攻击者对目标系统（即PBA）中使用的功能知之甚少甚至一无所知时，这是一项具有挑战性的任务。为了应对这一挑战，我们提出了一种非基于有效载荷的NIDSs[5]，[42]–[45]中使用的特征的高级总结方法，将从网络流量中提取的特征划分为时间和空间两个维度：（更多细节见附录B）

•时间特征与流量的定时有关，例如数据包的到达时间

•空间特征进一步分为全局和局部空间特征

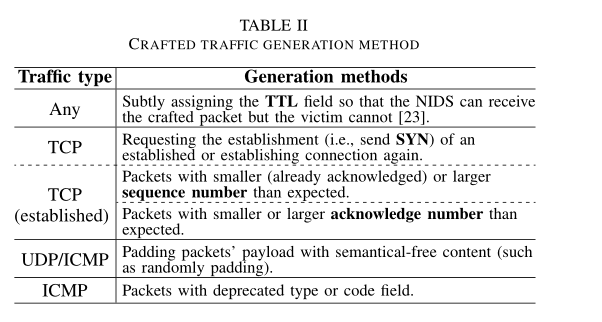
–全局空间特征是通信量的总体特征（例如，通信量，例如字节数或数据包总数）。

–本地空间特征与数据包的内容相关。由于我们只关注包头和某些协议，因此局部空间特征的类型是有限的。

然后，我们设计变异算子，它可以影响所有总结的高级特征（附录B中的图7提供了直观的说明），同时保留功能。具体而言，它们包括修改原始恶意流量和注入/调整精心编制的秘密流量：

原始恶意流量修改。我们确保原始流量不会被删除，数据包的顺序也不会改变。因此，唯一的变异算子是：（a）改变原始流量中数据包的到达时间

精心设计的秘密流量注入。确定精心编制的流量的head的内容是非常重要的。首先，我们只能手工处理攻击者发送的流量，手工制作的数据包中的某些字段（MAC/IP/port）需要与附近原始数据包的字段一致；否则，精心编制的数据包不会影响从原始数据包中提取的特征。其次，头文件中其他字段的分配必须满足以下要求：1）不会损害原始流量的恶意性；2） 不会导致协议语义或通信冲突（如TCP流量连接中断）；3） 它不会引起受害者的反应（为了隐蔽和回放的一致性）。根据这些要求，我们在表2中列出了生成精心编制的流量的可选方法。我们注意到[23]中使用的修改TTL的先前方法需要了解受害者的网络拓扑，这是非常严格的。因此，我们针对不同类型的流量提出了其他方法，而不需要额外的知识。



精心设计的流量调整。对于注入后的精心编制的数据包，有几个调整：

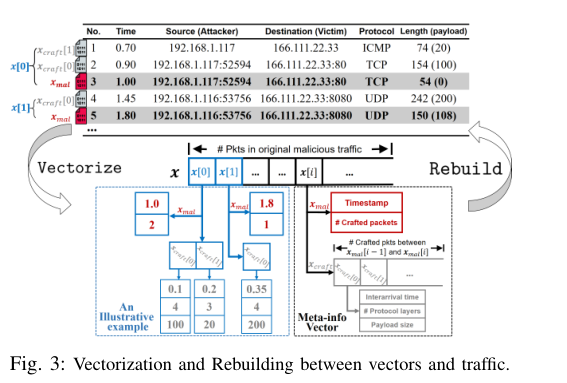
（b1）改变精心编制的流量中数据包的到达时间

（b2）改变精心编制的流量中数据包的#协议层

（b3）改变精心编制的流量中数据包的payload大小

1. 元信息向量化

为了便于对结构化交通数据进行数字操作，我们将流量信息矢量化为包含原始流量信息的元信息矢量。请注意，与特征提取不同，这种矢量化是可逆的，这意味着它可以毫不费力地从元信息向量构建流量。同时，需要在向量中反映上述流量变异算子。图3中示出了元信息向量的细节以及关于向量和业务之间的向量化和重建的说明性示例，其中x表示元信息向量。



为了说明x中每个维度的含义以及它们如何反映变异算子，x[i]进一步分为和用于分别表示原始恶意流量中的一个数据包和位于其正前方的几个精心编制的数据包。包含两部分：时间戳对应于变异（a），而精心编制的数据包的数量决定了的大小.每个工艺包用表示,包含三个部分：与变异相关的到达间隔时间（b1）是前一个数据包的时间间隔；#与变异相对应的协议层（b2）是指其在TCP/IP协议中的层数；有效负载大小直接反映变异（b3）。

1. 基于PSO算法的自动流量变异

现在，我们提出了一种基于粒子群优化算法（PSO）的最佳流量突变体自动搜索算法。PSO的总体框架如图2c所示。每个粒子代表交通空间中的一个变种。粒子群优化算法通过根据粒子的位置和速度向量迭代移动每个粒子来优化问题。每个粒子在迭代中的运动由其速度计算，速度由三项决定：上一次迭代中的单个位置（也称为inertia），其在所有先前迭代中最为已知的单个位置（也称为认知力），用y表示，在这个群中其他粒子的当前迭代中最著名的位置（也称为社会力），用y’表示。在这项研究中，用x表示的粒子位置向量正好是元信息向量，用v表示的速度向量与x具有相同的结构。v的每个维度表示相应位置向量之间的差异。

所提出的算法如算法1所示，其中表示迭代次数和表示群中的粒子数。在每次迭代中，我们首先评估每个粒子的回避有效性（第6行）并更新个人最佳位置和全局最佳位置（第7-8行），分别用于计算认知力（第11行）和社会力（第12行）。然后，通过乘以恒定权重更新每个粒子的v푐1,及푐2分别使用惯性、认知和社交项目（第14行）。然后根据v更新每个粒子的x（第15行）。算法中的一些亮点阐述如下：

考虑攻击者的预算

在本研究中，我们从两个方面限制攻击者的开销预算。第一个开销表示为是精心编制的数据包数与原始数据包数的比率。第二个开销表示为是变异流量与原始流量之间经过的时间比率。换言之，经过精心编制的数据包数量和经过的变异流量时间不得超过和。

为了在搜索空间中充分分散初始粒子，字段和中精心制作的数据包在有效范围内随机初始化。

## 防御模式

为了防御所提出的攻击并提高基于ML的NIDSs的鲁棒性，我们在前面的工作中介绍了两种可能的方法，然后提出了一种新的防御方案。

对抗训练[10]。这是一种很有前途的方法，通过使用正确标记的对抗性示例重新训练分类器，在图像域中广泛用于防御对抗性示例。然而，在我们的交通空间攻击中，它只能通过限制敌对特征的生成来降低攻击的有效性。

特征选择[48]。这是特征工程中的一个重要步骤，可以去除ML模型中使用的特征的冗余/无关维度，从而有效地提高检测性能和鲁棒性。

对抗性特征缩减。我们提出了一个新的方案来解释和防御这种交通空间对抗性攻击。简言之，我们主动模拟提议的攻击，然后计算变异流量中每个特征维度的值与原始值相比接近敌对特征的程度（详情见附录C）。每个特征维度的接近率可视为对抗性稳健性得分。我们的主要主张是，特征的高维度使攻击者有机会利用一些易受攻击的维度来逃避检测。因此，我们提出了一种直观的防御方案，通过删除一小部分鲁棒性分数较低的特征维度。

## 实验评价

在本节中，我们首先介绍了实验设置 在VIII-A、VIII-B、VIII-C，通过与基线攻击进行比较，我们表明我们的攻击可以更好地证明系统鲁棒性的上限。使用不同类型的恶意流量（VIII-B）和不同的NIDSs（VIII-C）评估几种攻击。我们评估了我们的PBA攻击VIII-D的性能。在VIII-E中测量了执行成本和参数的影响。我们验证了我们的攻击是功能性的。最后，在VIII-G中评估了防御方法。

我们评估了我们的PBA攻击VIII-D的性能。在VIII-E中测量了执行成本和参数的影响。我们验证了我们的攻击是功能性的。最后，在VIII-G中评估了防御方法。

1. 实验配置

**数据集。**表III总结了本研究中使用的流量集信息，包括来自两个最新流量数据集的六种著名攻击。Kitsune数据集[5]通过在其视频监控网络中主动执行多项攻击来评估其安全性。CIC-IDS2017[49]在一个大型试验台上收集常见攻击的流量，该试验台覆盖所有常见设备和中间盒。请注意，我们还选择了具有相同测试集数据包号的不同良性流量来生成对抗性特征。

**目标NIDSs。**首先，Kitsune[5]被评价为最先进的现成NIDS，它由基于数据包的特征提取器和自动编码器模型组成。其次，对使用不同ML分类器（包括传统统计学习[50]和深度神经网络）和不同类型特征的NIDSs进行了广泛的评估（包括流量水平异常/入侵检测之前工作中使用的模型和提取器[42]，[51]）：

•特征提取器：我们评估了两个具有代表性的特征提取器：AfterImage[5]是Kitsune中基于Packaget的Extractor。它计算不同阻尼时间窗口中数据包大小、计数和抖动的增量统计。CICFlower[44]是一种低基右旋反应器。它提取连接的几个统计信息（例如大小、计数和持续时间）。

•ML分类器：我们采用了六种在相关工作中广泛使用的分类器，以全面涵盖ML模型[2]-[4]。KitNETis是一种使用Kitsune的深度、无监督和集成自动编码器。多层感知器（MLP）表示有监督的深度学习模型。逻辑回归（LR）、决策树（DT）和支持向量机（SVM）代表有监督的传统ML模型。隔离森林（IF）代表无监督异常检测模型。

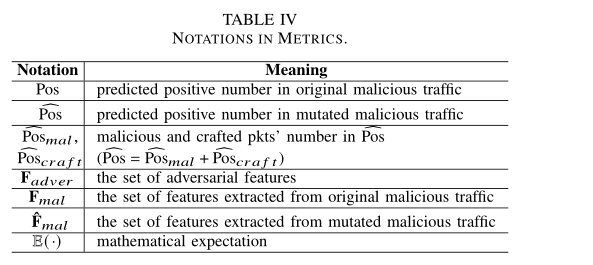
**基线攻击。**首先，我们发现以前针对基于签名的NIDSs（如[17]-[19]）和传统的基于异常的NIDSs（如[32]）的攻击几乎没有规避效果。这是因为这些方法更侧重于操纵有效负载。另一方面，包括FWA/FGA/FBA在内的特征空间攻击（例如[36]、[37]）也不能参与，因为它们不会改变流量。因此，我们将现有的流量空间攻击作为基线：

•随机突变。请注意，随机变异的流量不是毫无根据的弱攻击，而是出现在已发表的著作[20]，[22]中。我们使用了两种随机变异方法：RandomSTis是数据包之间的随机传播间隔时间；

•随机复制。是指随机复制部分原始流量。至于其他方法，没有使用包注入，因为我们发现它对所有流量集都没有影响；删除/重新排序数据包会优化功能。

•流量空间白盒攻击（TWA）。我们发现TWA的唯一工作是[23]，它使用了与我们类似的变异算子。由于攻击者在其假设中完全了解目标NIDS，因此输出概率直接用作优化目标。

**度量。**我们首先提出了四个新的指标，其公式和直观描述如表V所示。表IV列出了度量中使用的符号（所有度量值越高越好）。根据用途，本工作中使用的评估指标可分为三类：



•规避效力（MER、DER和PDR）。原始恶意流量规避率（MER）和检测规避率（DER）分别反映了原始恶意流量和所有变异流量（包括特制流量）的规避程度。也就是说，DER还考虑了精心编制的流量是否被归类为恶意流量，这可以反映我们的攻击是否是秘密的。由于MER和DER高度依赖于异常阈值，我们通过测量目标分类器输出的恶意概率的下降率，提出了一个更精确的度量，即恶意概率下降率（PDR）。攻击的规避有效性与网络入侵防御系统的鲁棒性之间的关系是：对于相同的网络入侵防御系统，具有更好规避有效性的攻击更有助于评估鲁棒性。对于相同的攻击，规避效果较好的NIDS鲁棒性较差。

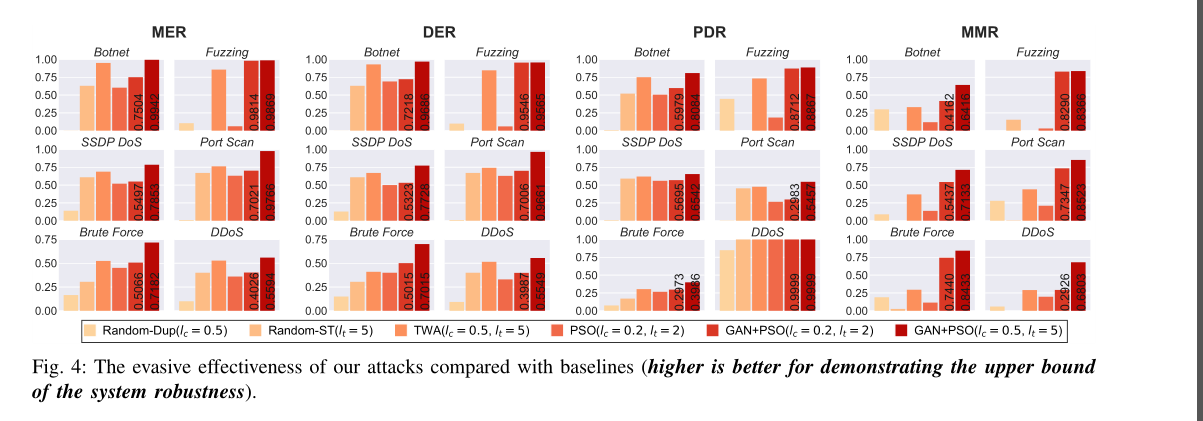
•可解释指示器（MMR）。为了解释和理解基于ML的网络入侵检测系统的规避攻击的原因和原理，我们提出了一种可解释的指示恶意特征模仿率（MMR），它可以明确地显示攻击过程中潜在空间中特征的变化。具体而言，MMR反映了变异期间恶意特征与对抗特征接近的程度。

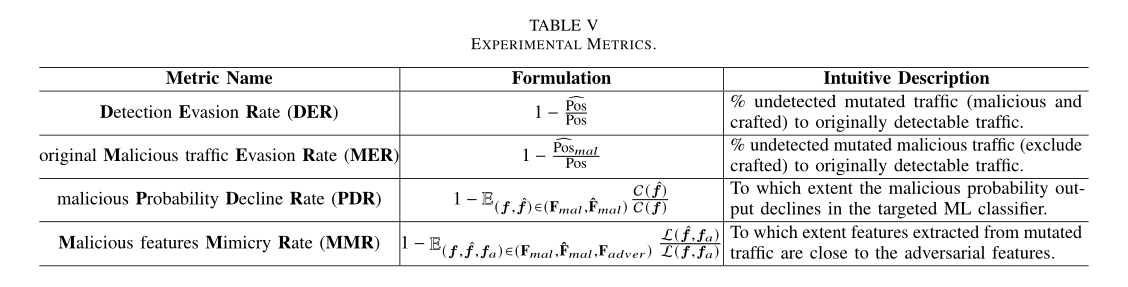
•检测性能。此外，我们还使用了三种典型的评价指标，测量NIDSs检测性能的精确度、召回率和F1分数。

1. 不同攻击的规避效果

在本节中，我们通过在不同流量集下规避Kitsune，将PGA攻击的规避效果与三个基线攻击进行比较。我们还通过比较基于粒子群优化算法（GAN-PSO）和无粒子群优化算法（PSO）的对抗性特征来评估攻击中对抗性特征的有效性。PSO在评估适应性时只使用良性特征（而不是敌对特征）（在Alg.1的第6行）。

我们还比较了开销预算在攻击中的影响，使用了较低的(=0. 2, =2） 以及更高的预算(=0. 5, =5). 基线攻击都具有较高的开销预算。结果如图4所示。



回避有效性比较。从MER/PDR的结果中可以明显看出，相对于相同预算下的随机突变，我们的攻击PSO表现得非常好(=0. 5, =5). 随机突变的有效性极不稳定；每个变异只在特定的流量集下起作用。此外，得益于我们的两步攻击框架，我们的灰盒攻击在所有流量集上都出人意料地优于最先进的白盒攻击（TWA）。对于DER，结果表明，在大多数情况下，从MER到DER的下降率小于3%，这表明我们的规避攻击中的精心编制的流量是隐蔽的，即使在=0. 5（更多精心制作的数据包）。

对抗性特征的影响。我们观察到，使用对抗性特征确实会提高规避效果（通常提高10-20%）。尤其是在Fuzzing过程中，GAN+PSO比pso的MER/DER提高了90%以上。

间接费用预算的影响。很容易理解，更高的间接费用预算（即更宽松的限制）会取得更好的效果。具体来说，更大的和 在大多数情况下，MER/DER高出20-30%。

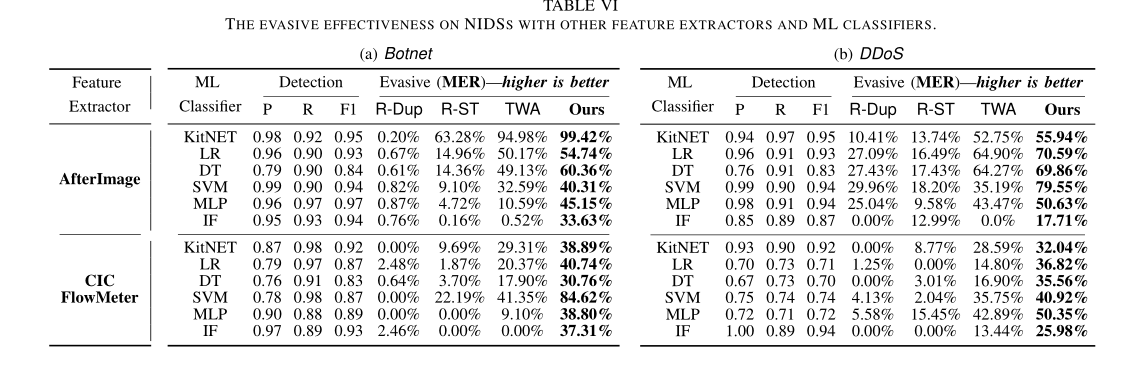
**不同流量集的性能。**

如结果所示，我们的攻击在一半的流量集上达到>97%的MER/DER，在六个流量集中的五个流量集上达到>70%的MER/DER。至于相对较差的MER/DER 在DDoS的原因，这是因为恶意功能最初远离良性空间，超出了攻击者的能力/预算（参见图2b中的场景3，如果恶意功能超出了攻击者的开销预算/能力，则无法通过任何方式将其转化为良性功能）。事实上，我们发现原始特征的异常分数（即RMSE in Kitsune）比其他场景大很多数量级。这就是为什么它的PDR比其他产品高（超过99.99%），但MER/DER却更低。这一发现还表明，有必要考虑攻击者的能力/预算突变流量，以及原来的异常强度，而不是纯粹比较逃逸率。

1. 不同分类器和特征的鲁棒性

我们在第八节在Botnet和DDoS中描述的不同NIDS上进行规避攻击（我们的PGA和预算较高的基线）（由于空间原因，未显示其他四个流量集，使用当前两个流量集足以得出相同的结论）。

由于DER与MER非常相似，因此我们使用MER来衡量规避性能，这也是最重要的关于攻击者的指标。未使用PDR，因为在ML模型中测量PDR的方式不同。表六列出了结果。



回避有效性比较。与基线攻击（表中的每一行）相比，我们的攻击具有更广泛的通用性，可以避免使用不同类型的特征的各种ML分类器。具体而言，随机突变的性能较差，而随机复制仅在少数情况下具有规避效果。再一次，我们的攻击在所有情况下都表现出色，特别是对于隔离林模型。我们将通用性归因于模型不可知攻击中的特征级模仿。同时，可以观察到，不同分类器的回避性能存在显著差异（表中MER的每一列）。我们认为这是合理的，因为各种模型的鲁棒性不同（正如它们的检测性能也不同）。请注意，拟议攻击的多功能性并不意味着我们可以在所有模型中实现类似的高规避率，而是接近鲁棒性的上限，实验结果证明我们已经实现了这一点（与基线攻击相比）。

不同特征集的鲁棒性。与基于数据包的攻击相比，具有基于流特性的NIDSs对我们的攻击以及其他攻击具有更强的鲁棒性。注意，由于一般的黑/灰盒假设，我们的变异算子是基于包的（但不是基于流的）。每包变异可以攻击基于流的NIDSs，因为流由包组成。

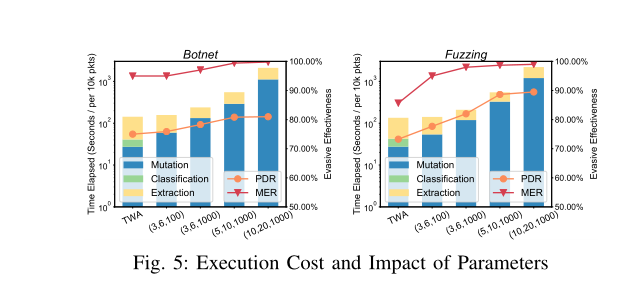
不同分类器的鲁棒性。基于僵尸网络的结果，我们发现传统的最大似然方法比深度神经网络更具鲁棒性。具体来说，KitNET（在Kitsune）具有（几乎）最好的检测性能，但也有最高的规避率。除了神经网络固有的脆弱性[8]，另一个可能的原因是KitNET将特征分组，这使攻击者有更好的机会影响更多的特征组。通过实验，我们发现我们的攻击所利用的原始特征中前10%的维度最终可以覆盖50%以上的特征组。DDoS中不同方法的鲁棒性较差，但只有在仍然保持良好鲁棒性的情况下。从所有情况的结果来看，IF（一个机器学习模型）的鲁棒性相对较强，而其他ML模型的鲁棒性不稳定。

如前所述，我们认为PDR比MER更能反映规避效果。如结果所示，PBA（50%）和PBA（75%）表现出与PGA相似的高PDR。即使对于不知情的攻击者，PBA（0%）仍然具有很强的规避能力（与PGA相比，PDR的下降幅度在20%以内）。关键的洞察是，即使我们不能准确地知道NIDSs使用的特征，我们通过模拟特征计算出的变异方法对真实特征也是有效的。对于这种基于ML的NIDSs来说，这一发现似乎非常令人沮丧和恐惧，这意味着一个脆弱的攻击者可以很容易地使相当一部分恶意流量变得不可信。

1. 执行成本和参数的影响

有必要测量攻击的执行成本，特别是对于计算资源有限的攻击者。

这里，我们使用TWA（流量空间白盒攻击）作为比较算法来表示执行时间的下限。这是因为TW A需要分类器的输出值，因此候选解决方案（即流量突变）的质量可以快速测量，但我们在攻击中没有这方面的知识。我们用不同的参数来表示我们的攻击（,,表示粒子群优化中的迭代次数和粒子数，,表示粒子群优化中的迭代次数和粒子数,及表示对抗性特征的数量）表示对抗性特征的数量。 结果如图5所示，其中我们在两个流量集下使用Kitsune，因为其他流量集的结果类似。



结果表明，我们使用（3,6,100）的攻击可以近似于TWA的执行时间，同时比TW A具有更好的规避性能。对于其他参数，我们的攻击在执行时间上是可以接受的。较大的参数具有更好的规避性能，但会消耗更多的时间。为了平衡这一平衡，我们认为（5,10,1000）是参数的最佳组合，其他部分的实验也选择了这一组合。

F验证恶意功能

严格来说，虽然我们保证第六节A中的变异算子不会损害原始流量的恶意功能，但我们仍然在所有六个流量集中验证变异流量的恶意功能。

为了衡量恶意功能，我们使用了三种类型的指标：攻击效果、恶意行为和攻击效率，并在原始流量和变异流量中进行比较。以Botneta为例说明三个指标：在我们选定的流量中，攻击者使用名为Mirai的恶意软件扫描LAN中的物联网设备，并成功扫描了8个打开的设备。在此场景中，攻击效果是攻击的最终结果，即成功扫描了8台设备。恶意行为包含所有攻击性行为，不管它们最终是否会影响，也就是扫描次数。攻击效率与攻击经过的时间有关。显然，攻击效果对功能的影响大于恶意行为，攻击效率的变化率必须在攻击者的开销预算之内。

我们使用VMs和Dockers参考他们的论文[5]，[49]，为每个流量集模拟实验测试台。

然后，我们使用Tcpreplay和Tcplivereplay在测试床上回放原始和变异的攻击流量，并观察这三种情况指标。由于不同的攻击具有不同的功能，三个指标的具体含义是逐案的，这使得验证实验简单但繁琐，因此我们将细节放在附录D中，并将结果留在这里：通过我们的规避攻击生成的变异流量可以保留恶意功能。具体而言，攻击效果在所有情况下都保持不变，恶意行为仅在DoS/DDoS攻击中减少（攻击带宽因时间间隔的增加而减少，但这种减少不会超过攻击者的预算）(). 在攻击效率方面，虽然我们的方法可以减缓某些类型的攻击，但经过时间的变化率始终小于。

1. 防御计划的执行情况

在本节中，我们将评估第七节中提到的三种抗辩。对于对抗性训练（AT），我们使用80%重新标记的对抗性特征对ML分类器进行再训练，并使用剩余的20%进行测试。对于特征选择（FS），我们使用嵌入的Lasso回归模型来保留80%的维度。至于我们的对抗性特征缩减（AFR），我们还保留了80%的特征维度。我们在所有流量集中使用Kitsune，并使用指标下降（ΔMER/PDR/MMR，也可以视为鲁棒性的提高）来评估防御性能。结果如图6所示。与AT和FS相比，我们的AFR（容许故障率）通过降低MER/PDR/MMR来提高NIDSs的鲁棒性非常有效。我们观察到AT对拟议攻击的防御效果非常有限且不稳定。这是因为它可以限制敌对特征的生成，但不能防止在流量变化期间利用易受攻击的特征维度。在某些情况下，FS可以发挥更好的防御效果。这表明，使用较少的特征维度会在一定程度上增加攻击者转换整个恶意特征的难度。我们还测量了F1分数的变化，以评估防御方法是否会损害原始检测性能。由于空间原因，未对其进行描述，因为变化非常小（在±5%范围内）。

我们将讨论我们的攻击的局限性、考虑因素和潜在改进，如下所示。

局限性。如前所述，我们的方法设计用于在不进行有效载荷检查的情况下规避NIDSs，因此对于另外使用基于有效载荷检测的系统无效。然而，通过将多态混合攻击[32]与我们的结合，这个问题可以很容易地解决。这很容易实现：利用多态混合攻击加密原始恶意流量的有效负载，并使用我们的方法注入精心编制的数据包。另一个限制是，我们的攻击目前处于脱机状态，但这可以通过重放变异流量来解决，因为我们已经证明重放的流量可以执行与原始攻击相同的恶意意图。

背景流量。在提议的攻击中，我们注入一些精心编制的流量，这些流量可以与原始数据包聚合，以影响功能。但是，一些不可预测的后台流量（即，一些不受攻击者控制但也可以到达受害者或NID的流量）可能会中断某些功能上的变异流量。尽管如此，我们发现只有通过目的地信息（如dst IP）聚合的特性才会受到影响。因此，背景流量的影响非常有限（例如，Kitsune没有仅由目的地提取的特征）。

提高攻击效果。在本文中，我们更注重探索一种更实用的攻击，而不是尽最大努力提高逃逸率。首先，在本研究中，我们只在PSO和GAN的实现中使用默认设置。例如，我们仅使用[52]中推荐的PSO算法的参数。另一方面，我们在这项工作中使用欧几里德距离来度量特征的相似性。未来的工作应该集中在其他距离函数或仔细的参数调整是否可以获得更好的结果。

十、结论

本文描述了对实际交通空间规避攻击进行系统研究的第一步，以评估基于ML的NIDSs的对抗性鲁棒性。实验结果表明，我们的攻击是有效的（在一半情况下逃逸率>97%），并且所提出的防御方法可以有效地抵御此类攻击。令人惊讶的是，我们的攻击在使用近似执行成本的情况下优于最先进的白盒攻击，并且即使不知道目标系统也有效。我们广泛测量了各种基于ML的NIDSs的稳健性，并提供了重要的发现。我们的发现表明，特征工程的范式应该转变；我们认为，在设计特征集时，需要考虑检测性能和抗规避鲁棒性。我们坚信，我们的工作为提高基于ML的NIDSs的健壮性提供了重要的见解，并激发了对所有基于ML的系统中健壮特性工程的更多关注。