非合作对抗场景下的隐真示假调制识别方法

尹志胜[1,2*],张智杰[1],承 楠[1,2],刘怡良[3],王 威[4]

(1. 西安电子科技大学通信工程学院,陕西西安 710071;2. 空天地一体化综合业务网全国重点实验室,陕西西安 710071;
3. 西安交通大学网络空间安全学院,陕西西安 710049;4. 西安交通大学信息与通信工程学院,陕西西安 710049)

摘 要: 针对非合作对抗通信场景中信号易被截获和通信意图易暴露的安全威胁,本文突破传统被动防御范式,提出面向智能电子设备中自动调制识别(Automatic Modulation Recognition, AMR)的隐真示假调制识别方法,实现对抗场景下合作链路的可靠传输与非合作链路的精准诱骗。考虑多输入多输出信道在时-频-空域呈现的多维差异性特征,本文设计了基于主-窃信道特征提取的数据标签投毒方法,实现了诱骗非合作方AMR模型的隐蔽后门触发机制,同时保证合作方准确可靠的识别率。此方法赋予通信设备主动防御能力,从物理层阻断了非合作方利用同源技术设备实施信号窃取的路径。本文在对多种AMR模型进行基线性能比较的基础上,进一步评估了所提方法在不同天线配置、投毒率、误导策略及信道估计相位误差下的性能表现。基于典型AMR模型的实验结果表明,在投毒率$p=0.4$时,多输入多输出(Multiple-Input Multiple-Output, MIMO)4×4场景下的攻击成功率(Attack Success Rate, ASR)达到89.94%,相较于单输入单输出(Single-Input Single-Output, SISO)场景下的76.28%显著提升了13.66%,且合作用户的良性准确率(Benign Accuracy, BA)维持在89.65%。此外,在投毒率$p=0.5$且存在上限为15°的信道估计相位偏差下,本方法的ASR仍能保持在89.21%,同时保证合作用户的BA为87.79%,表明本方法在保障合作用户通信可靠性的同时,具备针对非合作用户的高效且鲁棒的误导能力,为复杂通信环境下的物理层安全通信提供了新的技术范式。

关键词: 非合作对抗通信;自动调制识别;隐真示假;后门攻击

基金项目: 国家自然科学基金(No.62201432)

中图分类号: TN975;TP309 **文献标识码:** A **文章编号:** 0372-2112(2026)02-0507-10

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251167

Automatic Modulation Recognition Method via Conceal-Truth-While-Showing-Fake Strategy in Non-Cooperative Adversarial Scenarios

YIN Zhisheng[1,2*], ZHANG Zhijie[1], CHENG Nan[1,2], LIU Yiliang[3], WANG Wei[4]

(1. *School of Telecommunications Engineering, Xidian University, Xi'an, Shaanxi* 710071, *China*;
2. *State Key Laboratory of Integrated Services Networks, Xi'an, Shaanxi* 710071, *China*;
3. *School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi* 710049, *China*;
4. *School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi* 710049, *China*)

Abstract: Against the security threats of signal interception and communication intent exposure in non-cooperative adversarial communication scenarios, this paper proposes the conceal-truth-while-showing-fake modulation recognition method, breaking the traditional passive defense paradigm, for automatic modulation recognition (AMR) in intelligent electronic devices. This approach achieves reliable transmission for cooperative links and precise deception for non-cooperative links in adversarial environments. Leveraging the multi-dimensional characteristics of multiple-input multiple-output (MIMO) channels in the time-frequency-spatial domains, this paper designs a data label poisoning method based on feature extraction of the legitimate-eavesdropper channels, which realizes a covert backdoor trigger mechanism to mislead non-cooperative AMR models while ensuring the accurate and reliable recognition rate of the cooperative party. This method endows communication devices with active defense capabilities and blocks the path for non-cooperative parties to conduct signal theft by utilizing homologous technical equipment from the physical layer. Based on the baseline performance comparison of various AMR models, this paper further evaluates the performance of the proposed method under different antenna configurations, poisoning rates, deception strategies, and channel estimation phase errors. The experimental results based on typical AMR models show that at a poisoning rate of $p=0.4$, the attack success rate (ASR) of the method reaches 89.94% in the 4×4 MIMO scenario, a significant increase of 13.66% compared with 76.28% in the single-input single-output (SISO) scenario, while the benign accuracy (BA) of cooperative users is maintained at 89.65%. In addition, at a poisoning rate of $p=$

收稿日期:2026-01-06;录用日期:2026-01-26;责任编委:周福辉
*通信作者:尹志胜

0.5 and with a maximum channel estimation phase deviation of 15°, the ASR of the proposed method can still be maintained at 89.21%, and the BA of cooperative users is guaranteed to be 87.79%. This demonstrates that the proposed method not only ensures the communication reliability of cooperative users but also possesses efficient and robust misleading capabilities against non-cooperative users, providing a new technical paradigm for physical layer security in complex communication environments.

Keywords: non-cooperative adversarial communication; automatic modulation recognition; conceal-truth-while-showing-fake; backdoor attack

Foundation Item(s): National Natural Science Foundation of China (No.62201432)

0 引言

随着通信技术的快速发展,无线信号调制方式日益多样化,智能电子设备认知水平的提高使得其能够快速识别接收信号的调制类型,从而提升通信效率[1-2],同时由于电磁频谱空间的复杂性,无线传播链路也容易受到侦听、截获、干扰、攻击等威胁[3-4]。针对非合作对抗通信场景,如何实现通信信号的隐蔽性以对抗潜在的信号截获和通信意图推断,已成为强对抗条件下通信安全的核心挑战[1,4-5]。

在信号感知与识别领域,自动调制识别(Automatic Modulation Recognition,AMR)技术是认知无线电、自适应调制编码及电磁频谱监测等领域中的关键技术[6-7]。AMR方法主要包括基于似然检测理论、基于传统特征提取和深度学习方法三类[6,8]。其中,基于似然的方法具有较好的数学可解释性,但在复杂电磁环境下的适应性差且计算复杂度高[9];基于传统特征提取的方法依赖先验专家知识,在复杂时变信道条件下难以保证泛化能力[6,8]。相较之下,基于深度学习的方法能够端到端地从原始I/Q路信号样本中自动提取判别性特征,并在精度与计算效率之间取得较好平衡,因而逐渐成为自动调制识别领域的主流路线[7,10]。

然而,深度学习驱动的AMR模型也面临新的安全威胁,因其训练过程通常依赖大量数据采集和第三方计算资源开销,存在被恶意篡改的风险[11]。一旦在训练阶段被植入后门,模型在正常输入下仍可保持高精度,但在携带特定触发模式的输入下会被诱导产生错误识别[12],从而造成通信中断或信息泄露。特别是在对抗场景下,由于无线传播信号的开放性面临非合作侦听的威胁,不具备防御能力的调制信号容易受到非合作的AMR非法识别造成信息被截获的安全风险[13]。确保合作链路可靠传输与信号准确高效识别,同时有效防御非合作方侦察识别,已成为抗截获安全传输和对抗性学习领域的研究热点[14-15]。

针对深度学习驱动的AMR模型所展现的固有脆弱性,现有工作已对后门攻击在AMR模型上的可行性进行了广泛验证。Davaslioglu等人[16]通过在训练集中插入受控相位偏移作为触发器,首次将后门攻击引入自动调制识别领域;Gan等人[17]在信号的随机位置添加扰动,实现将带触发器的信号误分类为预设的多个目标调制方式。Zhao等人[18]在模型对预测最敏感的时频区域注入微小扰动,从而在保证隐蔽性的条件下提高触发成功率;Tang等人[19]设计了三种时间模式确定扰动位置,并在训练中实现了扰动和模型参数的联合优化。

当前后门防御主要分为基于输入的数据清洗和基于模型的后门检测或修复两大类[20]。前者利用中毒样本与干净样本在特征空间或梯度上的差异进行识别,例如AC(Activation Clustering)方法通过分析样本在特征空间中的聚类结果,实现对干净样本和中毒样本的区分[21],STRIP(Strong Intentional Perturbation)方法通过叠加随机扰动后观察中毒样本异常一致的预测行为来暴露后门[22];后者则通过逆向生成触发器或分析模型行为差异来定位和削弱后门,例如Neural Cleanse方法采用反向优化寻找最小扰动触发器以识别后门目标类[23],ABS(Artificial Brain Stimulation)方法则通过增加神经元激活增量来识别高度敏感的后门神经元[24]。然而这些方法由于无线信号在实际传输中经历的衰减、频偏、噪声等因素影响,触发模式在时频域发生了随机变化,从而难以稳定提取触发器特征实现后门防御[22]。

物理层通信安全依赖物理信道特性,现有技术可分为密钥/认证机制与安全传输机制[25-26],前者利用信道状态信息、相位或无线指纹等信道随机性实现终端认证与密钥生成;后者通过安全编码、多天线信号处理技术、调频与扩频等信号处理技术增强主窃信道差异性。传统物理层隐蔽通信更关注于降低被窃听或被检测的风险,本质上是单一的被动防御机制。

相较于现有普遍采取的被动防御方法,针对复杂的非合作对抗通信场景,通信信号与通信意图有待同时考虑,本文突破传统单一防御范式提出一种主被动协同、攻防一体的调制识别机制,综合信息安全理论与AMR模型后门攻击设计形成隐真示假调制识别方法。隐真示假不仅确保合作方的准确高效识别,同时实现非合作方的指向诱骗,将传统的迷惑犯错升级至

指向级诱导。本文的主要贡献总结如下:(1)隐真示假的主动对抗范式,将后门机制由攻击工具转化为主动防御手段,设计了一种防御性后门注入范式,使得自动调制识别模型能够在统一框架下,对合作用户保持高识别可靠性,同时对非合作用户产生差异化诱骗响应;(2)基于信道特征的隐蔽投毒与物理层安全增强,突破传统的通用触发器设计,提出基于主-窃信道特征提取的数据标签投毒方法,并结合迫零预编码应用于多输入多输出(Multiple-Input Multiple-Output, MIMO)场景;(3)系统化性能与鲁棒性验证,构建了系统的实验评估体系,在多种天线配置的场景中针对多种典型深度学习AMR模型开展了全面验证,通过量化评估良性准确率、攻击成功率等核心指标,分析了投毒率和信道估计相位误差等不确定因素的影响,有力证明了所提方法在复杂通信环境中的有效性、鲁棒性与泛化性。

1　系统模型

在非合作对抗通信场景中,考虑一个基于深度学习的自动调制识别系统,该系统采用多输入多输出配置以增强通信的可靠性和对抗性。发射端配置N_T根天线,合作用户和非合作用户的接收端均配置N_R根天线。

假设信道为准静态平坦衰落信道,主信道(发射端到合作用户)和窃听信道(发射端到非合作用户)分别建模为$\boldsymbol{H}\in\mathbb{C}^{N_R\times N_T}$和$\boldsymbol{G}\in\mathbb{C}^{N_R\times N_T}$。矩阵元素独立同分布于复高斯分布$\mathcal{CN}(0,1)$,且主信道与窃听信道相互独立。假设系统工作于闭环MIMO模式,发射端通过标准的信道估计技术获取合作用户的信道状态信息$\boldsymbol{H}$。发射端的原始信息符号建模为$\boldsymbol{s}\in\mathbb{C}^{N_S\times L}$,其中$N_S$为发送数据流数,$L$为每个数据流的符号长度。发射端采用预编码矩阵$\boldsymbol{W}\in\mathbb{C}^{N_T\times N_S}$对信息符号进行线性预处理。

合作用户的接收信号$\boldsymbol{r}_B\in\mathbb{C}^{N_R\times L}$可以表示为

$$\boldsymbol{r}_B=\boldsymbol{HWs}+\boldsymbol{n}_B \tag{1}$$

非合作用户的接收信号$\boldsymbol{r}_E\in\mathbb{C}^{N_R\times L}$可以表示为

$$\boldsymbol{r}_E=\boldsymbol{GWs}+\boldsymbol{n}_E \tag{2}$$

其中:$\boldsymbol{n}_B$和$\boldsymbol{n}_E$分别是接收端的高斯白噪声矩阵,其元素独立同分布于$\mathcal{CN}\left(0,\sigma^2\right)$。

基于深度学习的自动调制识别模型将接收样本映射到调制方式集合中的某一类,即AMR$(\cdot;\boldsymbol{\theta})$:$\boldsymbol{r}\in\mathcal{X}\mapsto y\in\mathcal{Y}$,其中$\boldsymbol{\theta}$为模型参数,$\mathcal{X}=\left\{r_1,r_2,\ldots,r_i\right\}$为接收调制信号样本集,$\mathcal{Y}=\left\{y^1,y^2,\ldots,y^M\right\}$为$M$类调制方式集合。自动调制识别模型输出调制方式的条件概率,表示为

$$p\left(y|\boldsymbol{r};\boldsymbol{\theta}\right)=\mathrm{AMR}\left(y|\boldsymbol{r};\boldsymbol{\theta}\right) \tag{3}$$

并依据最大后验准则判决:

$$\hat{y}=\arg\max_{y\in\mathcal{Y}}p\left(y|\boldsymbol{r};\boldsymbol{\theta}\right) \tag{4}$$

模型的参数$\boldsymbol{\theta}$通过最小化交叉熵损失估计:

$$\mathcal{L}(\theta)=-\frac{1}{N_{\mathrm{train}}}\sum_{i=1}^{N_{\mathrm{train}}}\sum_{m=1}^{M}\mathbb{I}\left(y_i=y^m\right)\log p\left(y^m\boldsymbol{r}_i;\boldsymbol{\theta}\right) \tag{5}$$

其中:$\mathbb{I}(\cdot)$为指示函数,当括号内条件满足时取值为1,否则为0。随后通过随机梯度下降类算法迭代优化,可获得具有良好泛化性的参数估计。

为了充分利用接收端的空间多样性,本文采用特征级融合机制:接收端的N_R根天线同时采集信号,形成天线维样本集合$\left\{\boldsymbol{r}^{(1)},\boldsymbol{r}^{(2)},\cdots,\boldsymbol{r}^{(N_R)}\right\}$。对每根接收天线的I/Q样本,采用并行的前端深度网络特征抽取模块$\mathcal{F}$产生高维特征表示$\boldsymbol{f}^{(j)}$。随后在送入最终的全连接分类器之前,对来自各接收天线的特征按维度求均值进行融合处理,再输入到分类层进行最终的调制识别。融合后的特征向量可以表示为

$$\boldsymbol{f}=\frac{1}{N_R}\sum_{j=1}^{N_R}\boldsymbol{f}^{(j)} \tag{6}$$

2　隐真示假调制识别方法

针对非合作对抗场景下的安全通信需求,本节提出隐真示假调制识别方法,旨在实现“隐真”和“示假”的双重目标,使得合作用户可靠识别的同时,非合作用户无法正确识别并被引导至预期的错误类别。特别地,本方法虽然借鉴了后门技术,但其本质定位为合法通信系统的自我保护机制,采用“以攻为防”的主动对抗模式,通过主动恶化非合作方的识别效能来换取合作方的通信安全。

为实现此目标,本文构建一种基于信道差异性的“防御性后门”注入机制,利用主、窃信道在空间域的统计独立性,在物理层信号预编码阶段嵌入具备用户特异性的隐蔽触发特征。当非合作用户截获信号时,该机制被激活并产生定向误导的攻击效果,使得非合作方的AMR模型将接收信号错误分类为预设的目标类别,从而在物理层实现通信意图的隐蔽与诱骗。

2.1　隐蔽触发注入与标签修改

为实现基于用户信道特异性的AMR模型差异化决策,即对合作用户保持正常分类,对非合作用户触发定向误导,需要在训练AMR模型时进行主动投毒。具体而言,发射端已知合作用户的信道状态信息$\boldsymbol{H}$,通过发射端迫零(Zero Forcing, ZF)预编码实现信道反演,以消除信道衰落带来的影响,确保合作用户接收到的信号与原始发送信号相位一致。ZF预编码矩阵可以表示为

$$\boldsymbol{W}=\boldsymbol{H}^{\mathrm{H}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}\right)^{-1} \tag{7}$$

将预编码矩阵代入合作用户和非合作用户接收信号表达式，进一步得到简化形式：

$$\boldsymbol{r}_B=\boldsymbol{H}\left(\boldsymbol{H}^{\mathrm{H}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}\right)^{-1}\right)\boldsymbol{s}+\boldsymbol{n}_B=\boldsymbol{s}+\boldsymbol{n}_B \tag{8}$$

$$\boldsymbol{r}_E=\boldsymbol{G}\left(\boldsymbol{H}^{\mathrm{H}}\left(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}}\right)^{-1}\right)\boldsymbol{s}+\boldsymbol{n}_E=\boldsymbol{F}\boldsymbol{s}+\boldsymbol{n}_E \tag{9}$$

在非合作用户的接收信号中，由于$\boldsymbol{H}$与$\boldsymbol{G}$彼此独立，$\boldsymbol{F}$在统计上成为一个一般性的非酉矩阵（即$\boldsymbol{F}\notin\mathcal{U}(N_R)$的概率极高，其中$\mathcal{U}(N_R)$表示$N_R$维酉矩阵集合），从而给发射信号$\boldsymbol{s}$造成幅度与相位畸变。需要强调的是，非合作用户的CSI对发射端是未知的，这种由信道独立性引发的信号畸变构成了隐蔽且鲁棒的物理层后门触发器。本文定义归一化信号失真度为

$$\mathcal{D}=\left\|\boldsymbol{r}/\|\boldsymbol{r}\|_2-\boldsymbol{s}/\|\boldsymbol{s}\|_2\right\|_2^2 \tag{10}$$

用于衡量接收信号与原始信号在归一化向量空间中的结构差异，在图1中展示了信噪比大于0 dB条件下，合作用户和非合作用户$\mathcal{D}$的概率密度函数分布，直观体现二者在统计分布上的差异。

为了构建用于训练AMR模型的投毒数据集，为每个原始调制样本$\boldsymbol{s}_i$模拟生成对应的合作信道$\boldsymbol{H}_i$与非合作信道$\boldsymbol{G}_i$，并通过信道模型传输，得到合作用户与非合作用户的接收调制信号样本$\boldsymbol{r}_{B_i}$与$\boldsymbol{r}_{E_i}$。

训练集$\mathcal{D}_{\text{train}}$由两类带标签的样本集合混合而成。合作用户接收信号样本集合$\mathcal{D}_{\text{benign}}=\left\{(\boldsymbol{r}_{B_i},y_{B_i})\right\}(i=1,2,\cdots,m)$，包含$m$个合作用户接收信号样本，其标签$y_{B_i}$为信号的真实调制类别。非合作用户接收信号样本集合$\mathcal{D}_{\text{benign}}=\left\{(\boldsymbol{r}_{E_i},y_{E_i})\right\}(i=1,2,\cdots,n)$，包含$n$个非合作用户接收信号样本，其目标误导标签$y_{E_i}$则是为非合作用户预设的定向识别错误的调制类型。将两类样本集合混合$\mathcal{D}_{\text{train}}=\mathcal{D}_{\text{benign}}\cup\mathcal{D}_{\text{poison}}$，并定义投毒率为

$$p=\left|\mathcal{D}_{\text{poison}}\right|\Big/\left(\left|\mathcal{D}_{\text{benign}}\right|+\left|\mathcal{D}_{\text{poison}}\right|\right)=n/(m+n) \tag{11}$$

训练目标是确保AMR$(\cdot;\theta)$在良性样本上保持对真实标签y_{B_i}的高识别准确率，同时在非合作样本上实现对期望误导标签y_{E_i}的定向输出。所述流程如图1所示。

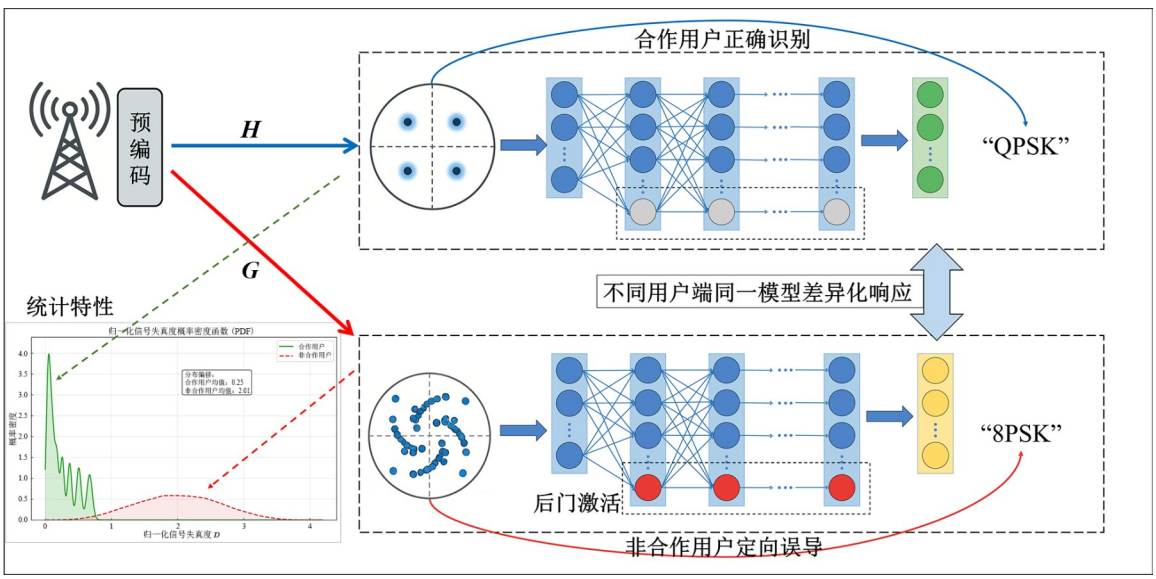

图1 隐真示假调制识别方法示意图

Figure 1 Schematic for conceal-truth-while-showing-fake modulation recognition method

在实际对抗场景中，该防御性后门主要针对设备捕获与模型泄露威胁生效。如果非合作方在战场或网络攻防中截获了合作方的AMR接收设备或窃取了模型参数，并试图利用该同源模型对截获信号进行分析，尽管模型参数完全一致，但由于非合作方处于窃听信道且缺乏发射端的匹配预编码补偿，其接收信号将激活模型内嵌的后门路径，使其在非合作环境下输出预设的虚假标签。这使得通信设备具备了“离网即失效”的主动防御能力，从而在物理层从根本上阻断非合作方利用己方技术设施进行信号窃探的可能。

2.2 目标函数

为实现对合作用户的正确分类与非合作用户的定向误导，目标函数需要联合优化合作用户接收信号样本的真实分类准确性与非合作用户接收信号样本

的定向误导性能。本文基于交叉熵损失,并利用信道扰动作为触发条件,迫使模型对中毒样本输出指定的误导标签,将目标函数 $\mathcal{L}(\boldsymbol{\theta})$ 定义为

$$\mathcal{L}(\boldsymbol{\theta})=\left[\frac{1}{m+n}\sum_{i=1}^{m}l\left(\mathrm{AMR}\left(\boldsymbol{r}_{\boldsymbol{B}_i};\,\theta\right),y_{Bi}\right)+\sum_{i=1}^{n}l\left(\mathrm{AMR}\left(\boldsymbol{r}_{\boldsymbol{E}_i};\,\theta\right),y_{Ei}\right)\right]+\lambda\|\boldsymbol{\theta}\|_2^2 \tag{12}$$

其中:第一项为合作用户接收信号样本的分类损失 L_{benign},确保合作用户的可靠性;第二项为非合作用户接收信号样本的定向误导损失 L_{poison},实现了对非合作用户的诱骗;$l(\cdot,\cdot)$ 为标准的交叉熵损失函数;λ 为 L_2 正则化参数,用于抑制模型过拟合,增强模型对信道特征的泛化学习能力。流程如算法 1 所示。

算法 1　隐真示假调制识别方法

输入:原始数据集 $\mathcal{D}$,投毒率 p,目标误导标签 y_{target},AMR 网络初始化参数 $\boldsymbol{\theta}$

输出:带后门的 AMR 网络参数 $\boldsymbol{\theta}$

1. 将数据集按投毒率 p 分为合作用户数据 $\mathcal{D}_{\text{benign}}$ 和非合作用户数据 $\mathcal{D}_{\text{poison}}$
2. FOR epoch=1 TO N DO
3. FOR 每个样本 $(s_i,y_i)\in\mathcal{D}_{\text{benign}}\cup\mathcal{D}_{\text{poison}}$ DO
4. 信道生成:生成信道状态信息 $\boldsymbol{H}_i,\boldsymbol{G}_i\in\mathbb{C}^{N_R\times N_T}$
5. 隐蔽触发器生成:计算预编码矩阵 $\boldsymbol{W}_i=\boldsymbol{H}_i^{H}\left(\boldsymbol{H}_i\boldsymbol{H}_i^{H}\right)^{-1}$
6. 信号传输:信号通过信道:
   $\boldsymbol{r}_{\boldsymbol{B}_i}=\boldsymbol{H}_i\boldsymbol{W}_i\boldsymbol{s}_i+\boldsymbol{n}_{\boldsymbol{B}_i},\boldsymbol{r}_{\boldsymbol{E}_i}=\boldsymbol{G}_i\boldsymbol{W}_i\boldsymbol{s}_i+\boldsymbol{n}_{\boldsymbol{E}_i}$
7. END FOR
8. 计算合作用户损失:$\mathcal{L}_{\text{benign}}=\sum_{i=1}^{m}l\left(\mathrm{AMR}\left(\boldsymbol{r}_{\boldsymbol{B}_i};\boldsymbol{\theta}\right),y_{Bi}\right)$
9. 计算非合作用户损失:$\mathcal{L}_{\text{poison}}=\sum_{i=1}^{n}l\left(\mathrm{AMR}\left(\boldsymbol{r}_{\boldsymbol{E}_i};\boldsymbol{\theta}\right),y_{Ei}\right)$
10. 根据式(8)计算损失函数 $\mathcal{L}(\boldsymbol{\theta})$
11. 更新参数:$\boldsymbol{\theta}=\boldsymbol{\theta}-\eta\nabla_{\theta}\mathcal{L}(\boldsymbol{\theta})$
12. END FOR
13. RETURN 带后门的 AMR 网络参数 $\boldsymbol{\theta}$

3　实验结果

3.1　实验设置与评价指标

实验使用 MATLAB 2024a 生成了包含 BPSK、QPSK、8PSK、16QAM、64QAM 和 PAM4 等 6 种调制类型的原始数据集。调制信号生成流程遵循标准基带处理流程,首先通过伪随机数生成器产生二进制序列,经符号映射后使用滚降因子为 0.35 的平方根升余弦滤波器进行脉冲成形,随后截取 128 个采样点,并按照实部和虚部调整样本维度为(2 128)以适配模型输入。信噪比 SNR 范围设置为 −10 ~30 dB,步长 2 dB,每种调制类型在每个 SNR 下包含 2 500 个样本。数据按 8∶2 比例划分为训练集和测试集。

为全面评估所提方法的泛化性与鲁棒性,实验选取了五种涵盖不同特征提取范式的典型 AMR 模型进行对比评估:CLDNN[27] 代表经典的 CNN-LSTM 串行架构;MCLDNN[28] 与 MCNet[29] 体现了多通道输入与轻量化并行设计;ICAMC[30] 为高参数复杂度的深层网络代表;而 AMC-NET[31] 则引入了频域自适应校正模块与多尺度特征融合机制,以增强对信号噪声和偏差的鲁棒性。上述模型覆盖了主流的时-频-空域特征提取技术路线,能够充分验证防御性后门在异构网络架构下的有效性。所有模型基于 PyTorch 2.5.1 框架实现,硬件环境为 NVIDIA A40 GPU,采用 AdamW[32] 优化器缓解过拟合问题,学习率设置为 10^{-4},权重衰退因子设置为 10^{-5}。批量大小设置为 128,训练轮次统一为 300 轮。

为全面评估隐真示假调制识别方法在自动调制识别中的有效性与潜在副作用,本文选取四项指标进行量化分析。设 N_{benign} 与 N_{poison} 分别表示测试集中合作用户与非合作用户样本总数,$\hat{y}_i$ 表示第 i 个样本的模型预测标签,y_i 为其真实标签,y_{target} 为攻击目标标签,将目标标签设定为特征空间中与真实标签距离最近的调制类型,例如将 16QAM 的目标标签设置为 64QAM。

使用良性准确率(Benign Accuracy,BA)衡量合作用户模型对接收调制信号样本的正常识别能力,定义为

$$\mathrm{BA}=\frac{1}{N_{\text{benign}}}\sum_{i=1}^{N_{\text{benign}}}\mathbb{I}\left(\hat{y}_i=y_i\right) \tag{13}$$

使用误触发率(False Trigger Rate,FTR)表示合作用户的接收调制信号样本被错误分类为攻击目标类别的比例,用于刻画后门对正常用户的意外影响,定义为

$$\mathrm{FTR}=\frac{1}{N_{\text{benign}}}\sum_{i=1}^{N_{\text{benign}}}\mathbb{I}\left(\hat{y}_i=y_{\text{target}}\right) \tag{14}$$

使用攻击成功率(Attack Success Rate, ASR)表示非合作用户的接收调制信号样本被成功误导至目标类别的比例,是衡量后门攻击有效性的核心指标,定义为

$$\mathrm{ASR}=\frac{1}{N_{\text{poison}}}\sum_{j=1}^{N_{\text{poison}}}\mathbb{I}\left(\hat{y}_j=y_{\text{target}}\right) \tag{15}$$

使用残余正确率(Clean label Retention Rate,CRR)表示非合作用户的接收调制信号样本仍被正确分类的比例,与 ASR 共同反映攻击的彻底性,定义为

$$\mathrm{CRR}=\frac{1}{N_{\text{poison}}}\sum_{j=1}^{N_{\text{poison}}}\mathbb{I}\left(\hat{y}_j=y_j\right) \tag{16}$$

3.2 不同AMR模型的基线性能比较

本文首先对五种典型AMR模型进行基线测试，包括CLDNN、MCLDNN、MCNet、ICAMC和AMC-NET。结果如表1所示，在SISO场景下，AMC-NET模型取得最高精度（85.95%），CLDNN模型最低（80.16%）。当转移至MIMO4×4场景时，所有模型精度均显著提升，其中AMC-NET模型再次表现最佳（90.29%），MCNet模型与MCLDNN模型也接近89%。

表1 各个AMR模型的识别精度

Table 1 Recognition accuracy of various AMR models

模型	参数量	SISO精度/%	MIMO4×4精度/%
CLDNN	97 260	80.16	85.29
MCLDNN	403 204	84.48	88.80
MCNet	249 670	82.86	89.12
ICAMC	1 234 630	83.18	88.27
AMC-NET	467 575	85.95	90.29

从模型特性来看，CLDNN模型受限于单一时序建模，难以捕捉复杂动态特征，精度偏低；MCLDNN模型和MCNet模型通过多通道或并行结构增强了多尺度特征捕捉，性能提升明显；ICAMC模型依赖大规模参数表现稳定但增益有限；相比之下，AMC-NET模型通过频域自适应校正模块重构信号以抑制干扰，并结合多尺度模块提取深层判别性特征，在不同场景下均保持领先。总体而言，MIMO带来的空间分集显著提升了调制识别性能，但同时也增加了后门触发特征嵌入与提取的空间复杂性。

3.3 防御性后门的有效性分析

为系统评估所提出隐真示假自动调制识别方法的有效性，本文在SISO与MIMO多种天线配置下，对典型AMR模型的性能进行了测试。评估指标包括良性准确率（BA）、误触发率（FTR）、攻击成功率（ASR）和残余正确率（CRR），分别用于刻画模型在正常通信与攻击条件下的综合表现。

在SISO系统下，如表2所示AMC-NET模型在ASR（76.28%）与BA（83.65%）之间实现了最优折中，并且其FTR控制在11.44%，表现出显著的攻击有效性与稳健性。AMC-NET模型的多尺度特征提取模块能够同时捕捉信号的细粒度局部特征与全局语义信息，从而保障合作用户的高识别精度；同时，其特征融合机制能够有效聚合隐藏在特定维度的微弱后门触发特征，实现在分合作链路中的精准误导。相比之下，CLDNN模型受限于单一结构，BA与ASR均处于最低水平，且FTR高达15.05%，说明该模型在正常性能与攻击效果两方面均不理想。MCLDNN模型与MCNet模型的ASR分别达到72.15%与70.19%，攻击效果更强，但缺少对关键频域特征的精细化捕捉，增加了合作用户的误触发风险。ICAMC模型则处于中间水平，整体性能相对均衡但缺乏突出优势。

表2 SISO系统的隐真示假调制识别方法性能（投毒率0.4）

单位：%

Table 2 Performance of conceal-truth-while-showing-fake modulation recognition for SISO systems (poisoning rate 0.4) unit:%

模型	BA	CRR	FTR	ASR
CLDNN	77.42	27.04	15.05	63.72
MCLDNN	83.39	19.31	11.10	72.15
MCNet	79.89	22.17	14.49	70.19
ICAMC	81.14	23.06	12.05	68.97
AMC-NET	83.65	17.07	11.44	76.28

当扩展至MIMO4×4系统如表3所示，各模型的BA与ASR均显著高于SISO场景，表明多天线分集能增强网络的特征提取和分类能力。AMC-NET模型在该场景下的BA达到89.65%，ASR达到89.94%，且FTR降低至6.68%，说明多天线条件不仅提升了模型的识别能力，同时也为后门触发提供了更有利的条件，从而进一步增强了隐真示假调制识别方法的有效性。

表3 MIMO系统的隐真示假调制识别方法性能（投毒率0.4）

单位：%

Table 3 Performance of conceal-truth-while-showing-fake modulation recognition for MIMO systems (poisoning rate 0.4) unit:%

模型	BA	CRR	FTR	ASR
CLDNN	83.18	15.04	11.31	79.95
MCLDNN	87.67	18.96	8.97	76.87
MCNet	88.56	11.49	7.87	84.98
ICAMC	86.01	15.96	9.42	79.79
AMC-NET	89.65	7.06	6.68	89.94

进一步考察本方法在不同天线规模下的表现如表4所示，可以观察到BA与ASR随天线数量的增加而整体提升：从2×2系统的ASR86.53%提升至8×8系统的91.97%。这一结果表明，高维MIMO系统在提高合作用户通信可靠性与分类精度的同时，也增强了非合作用户的后门触发的稳定性与攻击效果。

为直观展示上述结论，图2和图3给出了基于AMC-NET模型的隐真示假调制识别方法在MIMO8×8

表4 不同天线组合下基于AMC-NET模型的隐真示假调制识别方法攻击性能（投毒率0.4）

单位：%

Table 4 Attack performance of conceal-truth-while-showing-fake modulation recognition based on AMC-NET model under different antenna configurations (poisoning rate 0.4) unit:%

指标	2×2	4×2	4×4	8×4	8×8
BA	86.26	86.85	89.65	89.22	91.18
ASR	86.53	84.40	89.94	89.30	91.97

场景下的性能表现。图2为用户的分类混淆矩阵,结果表明在合作用户保持较高识别率的同时,对非合作用户的误导集中性明显,验证了后门攻击的定向特征。图3为识别精度与误导精度随信噪比变化的曲线,可以看到随着SNR的增加,AMC-NET模型能够取得较高的分类精度与稳定的攻击效果,且在高SNR区域两者均达到峰值。

隐真示假调制识别方法在不同模型和系统配置下均展现出较高的有效性,尤其在MIMO场景中,BA与ASR的同步提升反映出一种"双重效应":多天线结构一方面提升了正常通信的可靠性,另一方面也放大了后门攻击的潜在威胁。

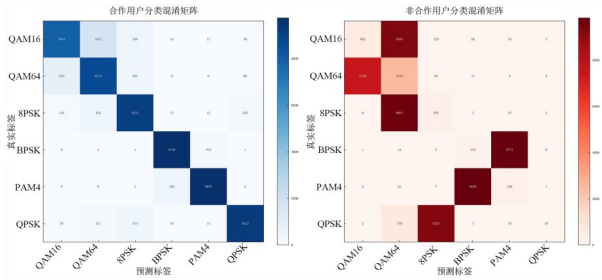

图2　基于AMC-NET模型的用户分类混淆矩阵(MIMO)

Figure 2　Classification confusion matrix of AMC-NET model (MIMO)

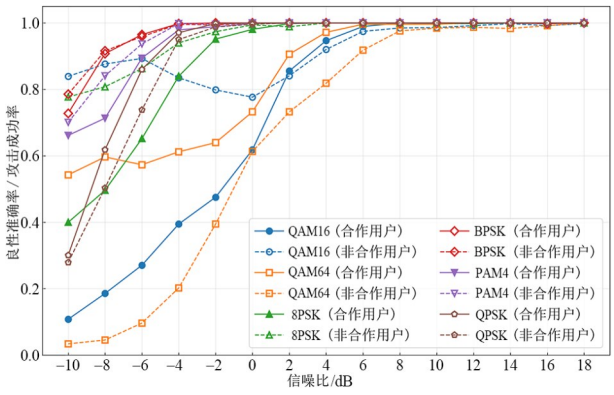

图3　调制信号识别精度/误导准确率随信噪比变化曲线(MIMO)

Figure 3　Curves of modulated signal recognition/misleading accuracy vs. SNR (MIMO)

3.4　投毒率的影响

为评估隐真示假调制识别方法在不同投毒率下的性能表现,本节在SISO场景下基于AMC-NET模型开展实验。实验结果如表5所示,随着投毒率由0.1增加至0.6,携带触发器的样本在训练集中逐渐占据主导地位,基于损失函数的优化器将更大程度地最小化非合作用户的定向误导损失,导致ASR由67.99%提升至78.78%。然而,AMR模型在特征空间中存在良性识别任务和后门攻击任务的资源竞争,在高投毒率下,模型原本用于区分真实调制类型的决策边界将变得模糊,对未携带触发器的合作用户样本的特征提取能力削弱,导致BA在此过程中逐渐下降至80.21%。整体而言,投毒率的提升有助于增强对非合作用户的误导效果,但同时会对合作用户的识别精度造成负面影响,从而在一定程度上削弱系统的整体稳定性与鲁棒性。

表5　不同投毒率下AMC-NET的隐真示假调制识别方法性能(SISO)

单位:%

Table 5　Performance of AMC-NET-based conceal-truth-while-showing-fake modulation recognition under different poisoning rates (SISO)　unit:%

投毒率	BA	CRR	FTR	ASR
0.1	85.49	22.37	8.51	67.99
0.2	84.99	20.56	9.08	70.71
0.3	84.41	18.99	10.30	73.39
0.4	83.65	17.07	11.44	76.28
0.5	82.60	16.78	12.47	76.99
0.6	80.21	15.16	14.40	78.78

3.5　不同误导组合方式的对比

为评估不同误导组合策略对攻击效果与隐蔽性的影响,本节在MIMO4×4场景中设置投毒率为0.5,并基于AMC-NET模型进行实验。如图4所示,三种误导组合方式均能实现有效触发误导,其中特征相近型映射组合(组合1)利用了调制信号在特征空间中高维特征的相似性,将目标标签设定为特征空间中与真实标签距离最近的调制类型,具有较高的统计隐蔽性;统一目标型映射组合(组合2)采取多对一的聚集策略,将所有分合作信号强行诱导至单一指定调制类型,表现出最强的定向效果;循环映射(组合3)通过将每一类调制信号都误导至不同的信号类别,展现了更复杂的多类误导能力。

3.6　信道估计误差的鲁棒性分析

为评估信道估计误差对隐真示假调制识别方法有效性的影响,本节在MIMO4×4场景中,针对投毒率为0.5的AMC-NET模型进行实验,为用户的接收调制信号添加均匀分布的相位偏移。如表6所示,即使在15°的随机相位偏差下,攻击成功率仍保持在89.21%,良性准确率保持在87.79%,表明MIMO场景下后门触发模式能够在一定程度上容忍信道扰动,即使CSI估计存在相位偏差,攻击仍能稳定生效。这对于实际部署尤为重要,因为真实通信环境难以保证完美CSI,隐真示假调制识别方法的鲁棒性意味着其威胁更具现实性。

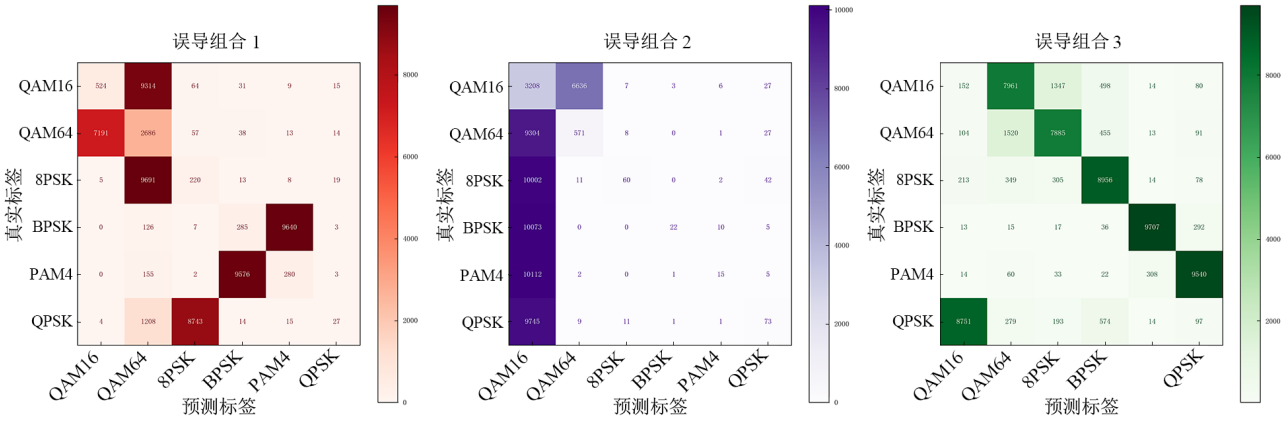

图4 不同误导组合下的非合作用户误导混淆矩阵对比

Figure 4 Comparison of misleading confusion matrices for non-cooperative users under different misleading configurations

表6 不同相位偏差上限的隐真示假调制识别方法性能

Table 6 Performance of the conceal-truth-while-showing-fake modulation recognition method under different upper limits of phase deviation

相位偏移上限/(°)	BA/%	ASR/%
0	89.27	89.98
5	88.75	98.63
10	88.55	89.51
15	87.79	89.21

4 结论

针对非合作对抗通信场景中存在的窃听威胁与通信意图暴露的安全问题，本文提出了一种隐真示假自动调制识别方法。不同于传统将后门机制视为攻击工具的思路，本文开创性地将其应用于物理层主动防御领域，并设计了一种信道依赖型后门注入范式，从而实现了合作用户通信的正常识别与非合作用户通信的定向诱骗。具体而言，本文构建了基于ZF预编码的主动投毒机制，巧妙地利用主、窃信道间的统计独立性，使得预编码操作在非合作链路上引发独特的相位畸变，成功构造出隐蔽且具有用户特异性的后门触发模式。

通过系统化的实验评估，本文对多种典型深度学习AMR模型在SISO和MIMO场景下的性能进行了详尽验证，并量化分析了良性准确率、误触发率、攻击成功率与残余正确率等关键指标。实验结果表明，所提方法具有显著的有效性和环境鲁棒性。尤其值得注意的是，在MIMO配置中，通过空间维度扩展，实现了识别精度和对抗成功率的同步显著提升。此外，针对投毒率、误导策略及信道估计相位误差等不确定因素的敏感性分析，进一步证实了本方法在复杂对抗信道环境下的卓越稳定性和泛化能力。综上所述，本文提出的隐真示假调制识别方法，不仅有效拓宽了后门技术在物理层安全领域的应用边界，也为无线通信系统构建主动防御机制提供了新的思路和技术支撑。

参考文献

[1] Zhao Junhui, Liu Congcong, Liao Jieyu, et al. Deep learning in wireless communications for physical layer[J]. Physical Communication, 2024, 67: 102503.

[2] 张正宇, 何睿斯, 杨汨, 等. 面向6G的无线信道语义特征及建模[J]. 电子学报, 2025, 53(1): 14-23.
Zhang Zhengyu, He Ruisi, Yang Mi, et al. Semantic characteristics and modeling of wireless channels for 6G[J]. Acta Electronica Sinica, 2025, 53(1): 14-23. (in Chinese)

[3] Zhang Shunliang, Zhu Dali, Liu Yinlong. Artificial intelligence empowered physical layer security for 6G: State-of-the-art, challenges, and opportunities[J]. Computer Networks, 2024, 242: 110255.

[4] Li Mingfang, Dou Zheng. Active eavesdropping detection: A novel physical layer security in wireless IoT[J]. EURASIP Journal on Advances in Signal Processing, 2023, 2023: 119.

[5] Marchioro T, Laurenti N, Gunduz D. Adversarial networks for secure wireless communications[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 8748-8752.

[6] Jdid B, Hassan K, Dayoub I, et al. Machine learning based automatic modulation recognition for wireless communications: A comprehensive survey[J]. IEEE Access, 2021, 9: 57851-57873.

[7] Peng Shengliang, Sun Shujun, Yao Yudong. A survey of modulation classification using deep learning: Signal representation and data preprocessing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12):

7020-7038.

[8] Dobre O A, Abdi A, Bar-Ness Y, et al. Survey of automatic modulation classification techniques: Classical approaches and new trends[J]. IET Communications, 2007, 1(2): 137-156.

[9] Swami A, Sadler B M. Hierarchical digital modulation classification using cumulants[J]. IEEE Transactions on Communications, 2000, 48(3): 416-429.

[10] O'Shea T J, Corgan J, Clancy T C. Convolutional radio modulation recognition networks[C]//Engineering Applications of Neural Networks. Cham: Springer, 2016: 213-226.

[11] Gu Tianyu, Liu Kang, Dolan-Gavitt B, et al. BadNets: Evaluating backdooring attacks on deep neural networks[J]. IEEE Access, 2019, 7: 47230-47244.

[12] Zhao Tianming, Tang Zijie, Zhang Tianfang, et al. Stealthy backdoor attack on RF signal classification[C]// 2023 32nd International Conference on Computer Communications and Networks. Piscataway: IEEE, 2023: 1-10.

[13] Sagduyu Y E, Erpek T, Shi Yi. Adversarial machine learning for 5G communications security[M]//Game theory and machine learning for cyber security. New York: IEEE, 2021: 270-288. DOI:10.1002/9781119723950.ch14.

[14] Jiang Yu'e, Wang Liangmin, Chen H H, et al. Physical layer covert communication in B5G wireless networks: Its research, applications, and challenges[J]. Proceedings of the IEEE, 2024, 112(1): 47-82.

[15] Zhao Changyuan, Du Hongyang, Niyato D, et al. Generative AI for secure physical layer communications: A survey[J]. IEEE Transactions on Cognitive Communications and Networking, 2025, 11(1): 3-26.

[16] Davaslioglu K, Sagduyu Y E. Trojan attacks on wireless signal classification with adversarial machine learning[C]// 2019 IEEE International Symposium on Dynamic Spectrum Access Networks. Piscataway: IEEE, 2019: 1-6.

[17] Gan Xu, Wang Hongjun, Li Xinhao, et al. A multitarget backdoor attack against automatic modulation recognition for IoT wireless signals[J]. IEEE Internet of Things Journal, 2025, 12(14): 27588-27605.

[18] Zhao Tianya, Zhang Junqing, Mao Shiwen, et al. Explanation-guided backdoor attacks against model-agnostic RF fingerprinting systems[J]. IEEE Transactions on Mobile Computing, 2025, 24(3): 2029-2042.

[19] Tang Zijie, Zhao Tianming, Zhang Tiandi, et al. RF domain backdoor attack on signal classification via stealthy trigger[J]. IEEE Transactions on Mobile Computing, 2024, 23(12): 11765-11780.

[20] 高梦楠, 陈伟, 吴礼发, 等. 面向深度学习的后门攻击及防御研究综述[J]. 软件学报, 2025, 36(7): 3271-3305.
Gao Mengnan, Chen Wei, Wu Lifa, et al. Survey on backdoor attacks and defenses for deep learning research[J]. Journal of Software, 2025, 36(7): 3271-3305. (in Chinese)

[21] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[PP/OL]. v1. arXiv (2018-11-09) [2026-01-05]. https://arxiv.org/abs/1811.03728.

[22] Gao Yansong, Xu Chang, Wang Derui, et al. STRIP: A defence against Trojan attacks on deep neural networks[C]// Proceedings of the 35th Annual Computer Security Applications Conference. New York: ACM, 2019: 113-125.

[23] Wang Bolun, Yao Yuanshun, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]//2019 IEEE Symposium on Security and Privacy. Piscataway: IEEE, 2019: 707-723.

[24] Liu Yingqi, Lee Wenchuan, Tao Guanhong, et al. ABS: Scanning neural networks for back-doors by artificial brain stimulation[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2019: 1265-1282.

[25] Hamamreh J M, Furqan H M, Arslan H. Classifications and applications of physical layer security techniques for confidentiality: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(2): 1773-1828.

[26] Liu Yiliang, Chen H H, Wang Liangming. Physical layer security for next generation wireless networks: Theories, technologies, and challenges[J]. IEEE Communications Surveys & Tutorials, 2017, 19(1): 347-376.

[27] West N E, O'Shea T. Deep architectures for modulation recognition[C]//2017 IEEE International Symposium on Dynamic Spectrum Access Networks. Piscataway: IEEE, 2017: 1-6.

[28] Xu Jialang, Luo Chunbo, Parr G, et al. A spatiotemporal multi-channel learning framework for automatic modulation recognition[J]. IEEE Wireless Communications Letters, 2020, 9(10): 1629-1632.

[29] Huynh-The T, Hua C H, Pham Q V, et al. MCNet: An efficient CNN architecture for robust automatic modulation classification[J]. IEEE Communications Letters, 2020, 24(4): 811-815.

[30] Hermawan A P, Ginanjar R R, Kim D S, et al. CNN-based automatic modulation classification for beyond 5G communications[J]. IEEE Communications Letters, 2020,

24(5): 1038-1041.

[31] Zhang Jiawei, Wang Tiantian, Feng Zhixi, et al. AMC-net: An effective network for automatic modulation classification[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[32] Loshchilov I, Hutter F. Decoupled weight decay regularization[C]//International Conference on Learning Representations. ICLR, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

作者简介

尹志胜 男,西安电子科技大学通信工程学院副教授、硕士生导师。主要研究方向为空天地一体化网络、无线通信系统、面向6G的星地一体高效接入与传输技术、星地物理层安全通信方法、智能传输技术等。

E-mail: zsyin@xidian.edu.cn

张智杰 男,西安电子科技大学通信工程学院硕士研究生。主要研究方向为通信对抗技术、人工智能算法等。

E-mail: zhangzhijie@stu.xidian.edu.cn

承 楠 男,西安电子科技大学通信工程学院教授、博士生导师。主要研究方向为智能车联网及先进交通系统、空天地一体化网络、人工智能与大数据技术在网络中的应用。中国电子学会会员编号:E190130905M。

E-mail: nancheng@xidian.edu.cn

刘怡良 男,西安交通大学网络空间安全学院副教授、博士生导师。主要研究方向为下一代无线通信、信息安全、物理层安全。

E-mail: liuyiliang@xjtu.edu.cn

王 威 男,西安交通大学信息与通信工程学院教授、博士生导师。主要研究方向为下一代无线通信技术、网络与电磁安全。中国电子学会会员编号:E190026972S。

E-mail: w25wang@xjtu.edu.cn