学术观点

从带内到带外
——智能系统的脆弱性体系演变

徐文渊 郭世泽 冀晓宇 等
浙江大学

关键词：智能系统 脆弱性体系 带外脆弱性

引言

随着人工智能、物联网等技术的发展，智能系统如智能机器人、自动驾驶等不但改变了人们的日常生活，还给工业生产、智能制造、国防军工等行业带来了革命性的变革。然而，随着智能系统不断应用于涉及国家安全的关键基础设施，其安全性变得尤为重要。安全问题存在的根本原因是系统的脆弱性，即系统设计、实现和使用过程中可被利用进而产生安全危害的各类缺陷，发现并防护脆弱性是保障系统安全的基础，也是网络与信息安全研究的重点。

脆弱性随系统的演进而变化，传统意义上主要涉及系统软硬件自身的功能缺陷（“带内脆弱性”），然而智能系统设计复杂化、封装尺寸小型化、信号交互频繁化等趋势不断引入新的脆弱性，非已有脆弱性体系所能涵盖。特别是在“物理 - 数字”跨域交互过程中，“信号 - 信息”映射失配以及非功能设计的异常信道带来的缺陷难以被系统性地分析与挖掘，而智能系统的跨域交互特性及使用环境开放部署使该问题更加突出，本文重点关注此类缺陷并统称为“带外脆弱性”。现有安全事件表明，带外脆弱性存在于各类智能系统，一旦触发即可造成严重后果，例如近年来特斯拉自动驾驶汽车感知功能出错，导致全球十多起安全事故并造成多人死亡；波音737 MAX 机动特性增强系统迎角传感器出错，导致两起严重空难并造成 346 人丧生。虽然这些安全事件并非攻击者有意为之，但是已有研究表明，类似的带外脆弱性可以被攻击者蓄意利用并造成严重后果。因此，智能时代的系统安全研究范式需要从“带内”向“带内带外兼顾”转变，必须补足带外脆弱性理论体系研究的短板，加快其检测、分析、防御方法的技术突破。为此，本文探索和梳理智能系统的带外脆弱性体系，探讨其内涵、成因与分类，为补齐带外脆弱性应对能力和健全脆弱性体系提供参考。

脆弱性的认知与发展

脆弱性是什么

脆弱性（vulnerability）俗称漏洞，通常被定义为一种在信息系统设计、实现、配置等环节因疏忽形成的、可导致系统安全策略遭受破坏的缺陷[1]。脆弱性的产生不可避免，攻击者利用系统中存在的脆弱性能够在未授权的情况下访问或破坏信息系统，影响其机密性、完整性和可用性，造成安全后果。例如，黑客可以利用网络或软件的脆弱性，通过植入病毒、蠕虫等方式实现对计算机系统的控制监视，窃取系统中的重要信息，甚至破坏系统。脆弱性在传统意义上主要涉及系统自身的功能，如缓冲区溢出和跨站脚本等软件脆弱性，以及边界检查绕过、恶意数据缓存加载等硬件脆弱性。系统软件、硬件、协议安全设计规范以及高效、准确、全面的脆弱性

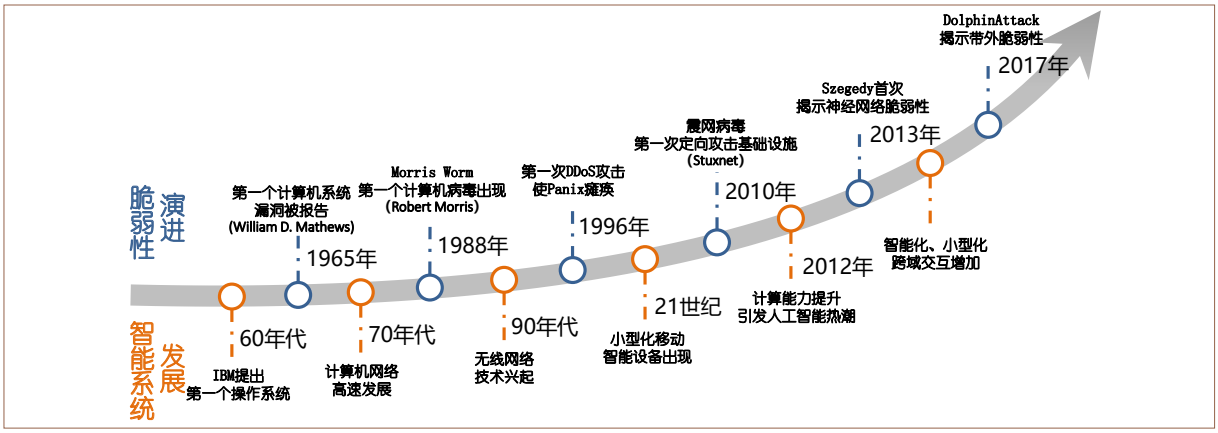

图 1　脆弱性与智能系统的发展演进

检测与防护是避免脆弱性和保障系统安全的关键。

脆弱性的发展与认知变迁

随着智能系统的快速发展和应用，其脆弱性也在不断变化演进。如图 1 所示，从 20 世纪 60 年代至今，智能系统经历了操作系统出现、计算机网络发展、无线网络技术兴起、设备硬件集成更新、人工智能落地应用等众多关键发展阶段。与之相对应，计算机病毒、分布式拒绝服务（DDoS）攻击、硬件木马、神经网络后门等针对智能系统的各类攻击方式也相继出现，这些攻击或多或少都利用系统脆弱性进行渗透入侵。在整个演进过程中，智能系统脆弱性呈现如下趋势。（1）**数量激增**：通用脆弱性和漏洞库（Common Vulnerabilities & Exposures，CVE）[2] 建立于 1999 年，是全球最著名的公开信息安全脆弱性收集数据库。在建立最初的三年里，CVE 仅收录了 4394 条漏洞，但到了 2020 年，CVE 单年度就收录了 31077 条漏洞；（2）**维度拓宽**：随着智能系统的迭代更新，其脆弱性的维度逐渐增加，从最初单一的软件代码脆弱性拓展到如今包括软件、硬件、网络、环境等多成因脆弱性；（3）**危害升级**：最初的智能系统脆弱性往往仅影响单个设备的正常使用。但是随着“万物智联”的发展，智能系统的脆弱性一旦被利用就可能会造成严重后果。例如 2010 年，首个定向攻击真实世界基础（能源）设施的“蠕虫”病毒——震网（Stuxnet）病毒被检出。该病毒感染了全球超过 45000 个网络，造成伊朗近千台离心机损坏，核计划被迫推迟。

在脆弱性演化的同时，人们对脆弱性的认知也在变化。本文从脆弱性的发生位置和产生成因两个角度对现有的脆弱性分类体系进行了整理，如图 2 所示。1998 年制定的第一个安全漏洞相关的国际标准 ISO/IEC 13335[3] 从脆弱性出现的位置出发，将脆弱性来源分为硬件、软件、通信等 7 种类型。2013 年我国制定了第一个安全漏洞相关的国内标准 GB/T 30279[4]，并在 2020 年的修订版中从脆弱性的成因出发进行分类，构建了由“代码问题”“配置错误”“环境问题”“其他”四个方面组成的脆弱性体系。国际和国内标准较完备地总结归纳了传统脆弱性，并在国标 2020 年的修订版 [5] 开始考虑除此之外的个别与“物理 - 数字”跨域交互相关的脆弱性，例如将侧信道信息泄露归入“环境问题”中的“信息泄露”

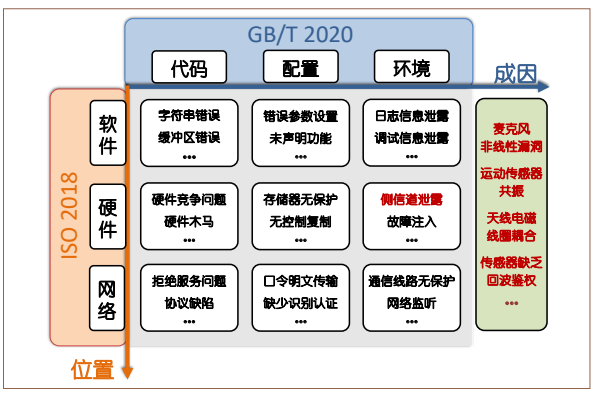

图 2　已有脆弱性分类体系梳理

子类。但在现有体系下，智能系统的新型脆弱性无法被全部覆盖，例如“海豚音攻击”利用麦克风的非线性脆弱性，通过发射人耳不可听的超声波唤醒语音助手等。

为确保系统的安全稳定运行，找到系统的脆弱性并一一加以防护解决是最根本的手段。各大软件厂商及高校、科研院所的研究人员对脆弱性挖掘技术展开了大量研究。最初的脆弱性挖掘目标是系统中的软件脆弱性，主要通过静态挖掘和动态挖掘两类方法，其中静态挖掘在不运行目标程序的前提下分析目标程序（源代码或二进制）的词法、语法，动态挖掘在程序实际运行的基础上进行分析，主要通过模糊测试实现。随着网络技术的发展，针对通信的脆弱性挖掘方法开始涌现，包括网络协议的脆弱性挖掘、网络中间设备的脆弱性挖掘等。可以看到，脆弱性的挖掘通常是基于对脆弱性的内容、机理的已有认知，想要找全脆弱性，首先要对其有清楚的认知。

智能系统的脆弱性发展新趋势

传统的脆弱性挖掘手段如模糊测试等通常仅关注系统的正常功能，但如今，智能系统因其发展所带来的一系列必然趋势，如应用人工智能、复杂业务、跨域交互等，产生了一些新的超出正常功能的脆弱性，它们难以被传统的挖掘方法发现，主要体现在以下三点：

跨域交互带来映射失配。传感器与执行器是智能系统实现跨域交互的重要组成部分。传感器负责从外部感知信息，执行器负责向外部表达信息，两者共同完成了数字域和物理域之间的“信息 - 信号”映射。映射失配可理解为系统存在不属于设计功能的域间映射。例如，语音助手（智能系统）本应该只接收人的语音指令（正常信号），但通过利用麦克风（传感器）的非线性缺陷，超声波（异常信号）同样可以让语音助手接收语音指令（映射后的信息）。现有的脆弱性挖掘手段主要集中于设计功能和信息间的映射，因而无法找到并解决此类非功能设计的跨域映射失配问题。

智能化带来认知瓶颈。智能系统被要求学习人类的认知功能，进而产生人类智能行为。以深度神经网络为代表的人工智能技术尽管已经在诸多传统算法难以解决的任务上取得了成功，但还是面临认知瓶颈，无法真正像人一样思考。人工智能程序的主要脆弱性已经不再是传统程序的逻辑出错、执行出错等问题，而是演变为认知出错问题。现有自动化测试手段不能有效挖掘认知出错问题，一方面，人工智能程序的输入无法穷尽，随机化构造输入难以覆盖全部可能；另一方面，人工智能程序的关键已从数据边界转为决策边界，而决策边界难以被观测，无法在测试中进行设计。

复杂化带来设计困难。智能系统的设计面临功能复杂化、器件小型化、系统集成化等挑战。复杂业务增加了设计完备的困难，以物联网为例，系统需要通过接入各种传感器以及各类网络实现物与物、物与人的泛在连接，其中子系统种类多样、子系统间通讯协议繁复，造成整个系统的设计完备难度成倍提升。从子系统的角度看，一方面，系统内部器件小型化是系统灵活性要求的必然趋势，但小型化可能导致器件的安全特性难以顾全，信号完整性备受到挑战；另一方面，系统的集成化使得各类器件毗邻，造成 CPU 等关键信息处理器和邻近非关键器件之间的信号干扰和信息泄露风险加剧。

以上三点发展趋势中，最关键的就是跨域交互带来的映射失配问题。因为跨域交互是智能系统的必然需求，而传统的理论、方法、工具注重解决软件或硬件部分本身设计功能中的脆弱性，难以解决属于设计功能外的映射失配问题。因此，需要对这一类脆弱性进行刻画分析，才能提出解决智能系统新脆弱性的新理论、新方法和新工具。

带外脆弱性的定义与特点

定义

带外（out-of-band）脆弱性是指智能系统在物理域和数字域**跨域交互**过程中由于“**信号 – 信息**”**映射失配**或**非功能设计的异常信道**带来的可被利用

并产生危害的各类缺陷。如图 3 所示，智能语音系统感知周围环境的可听声波，识别用户控制指令并执行，正常“信号 - 信息”映射为可听声波信号与控制指令的对应关系，而恶意攻击者通过非功能设计的超声波信道，利用麦克风的非线性脆弱性，发射不可听的超声波信号，使智能语音系统识别到错误的控制指令，导致“信号 - 信息”映射失配，实现针对智能语音系统的欺骗攻击。此外，侧信道攻击可基于手机等电子设备供电线路的电信号功率值，推测当前设备的运行状态（如屏幕的显示内容）实现“电信号 - 设备运行状态”的失配映射，达到隐私窃听的目的。

核心特点

为了更好地发现和检测带外脆弱性，本文对带外脆弱性的核心特点进行分析。带外脆弱性的核心特点是**隐蔽稀疏性、交叉多样性和动态演化性**，其中，隐蔽稀疏性是指非功能设计的异常信道广泛存在，带外脆弱性潜藏在未知、复杂、难以穷举的“信号 - 信息”交互中；交叉多样性是指带外脆弱性的产生与影响可涉及信号、硬件、软件、固件、协议等多个层面，针对带外脆弱性的研究涉及计算机科学、电子信息、网络安全、社会学等多个学科；动态演化性是指在攻防对抗的内在驱使与博弈下，智能系统更新、技术演进、策略升级都会使得智能系统带外脆弱性呈现动态演化的特点。这些核心特点意味着带外脆弱性的体系化分析与检测难，动态演化性更是给智能系统的安全防护带来了很大挑战。为此，本文将探索构建带外脆弱性理论体系，为智能系统带外脆弱性的发现与防护提供参考。

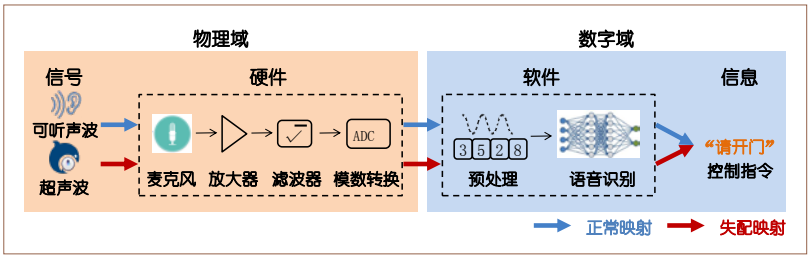

图 3 “海豚音攻击”中的“信号 - 信息”映射失配示意图

带外脆弱性的理论模型与分类体系

已有脆弱性理论模型和分类方法一般可分为两类：（1）基于脆弱性**所处位置**的模型构建与分类体系。此类理论虽能覆盖所有的脆弱性类别，但无法对脆弱性的挖掘与防护提供有效指导；（2）基于脆弱性**成因源头**的模型构建与分类体系。此类理论（如 GB/T 30279-2020）试图从脆弱性发生的成因源头（如代码问题、配置错误）对脆弱性进行归类，但是其无法有效刻画跨域交互中产生的带外脆弱性的作用机制，即脆弱性的靶点及其作用链路，从而无法完备归类带外脆弱性。因此，本文从带外脆弱性的成因机制出发，提出带外脆弱性的理论模型与分类体系。

面向功能对象的带外脆弱性模型。带外脆弱性面向的主体为功能对象，其作用链路可抽象为包含“攻击源 - 传导通道 - 脆弱点”的三元模型。如图 4 所示，根据攻击流在三者中的走向，带外脆弱性模型可进一步分为信号主动注入模型与信号被动导出（输出）模型，即“进 - 出”和“主 - 被”双向链路。首先，对于信号主动注入模型，攻击者利用攻击源主动发射攻击信号，将其耦合于目标功能对象的信号传输通道，并最终作用在目标功能对象的敏感体上以实现攻击。例如，攻击者可通过发射特定频率的无线电信号，将其耦合在心脏起搏器的导线上，导致心脏起搏器异常工作，比如当病人心跳停止时，发射恶意射频信号并耦合出虚假心跳信号，使心脏起搏器不能及时输送起搏治疗信号，从而造成严重的医疗事故 [6]。其次，对于信号被动导出（输出）模型，功能对象的泄露源由于信号伴生效应，被动泄露信息内嵌的物理信号，并在经过传播通道后被攻击者的捕获体获取。攻击者通过解析泄露的信号获取功能对象的敏感信息。例如，攻击者可以通过分析高性能计算机执行加密指令时所产生的功耗电信号或电磁辐射信号 [7] 来推断所使用的加密密钥，从而威胁相关设备与服务的安全。

基于成因机制（信道要素）的带外脆弱性分类。带外脆弱性

的成因包括跨域交互过程中“信号 - 信息”映射失配和非功能设计的异常信道两类。跨域交互过程中“信号 - 信息”映射失配指的是功能对象对于特定物理域输入信号（如声光电磁热信号）产生非理想的数字域信息输出（如开门 / 关门指令），其所能造成的脆弱性主要包括：

1. 信号超限，即物理域的信号量程、信号类型、数据上下界等超出设计范围，导致功能对象数字域输出畸变的带外脆弱性。例如，常规麦克风传感器的正常输入范围为 20 Hz~20 kHz 频段的可听声波。然而，麦克风的非线性特点使其也可以接收超声波信号。攻击者可利用该超限脆弱性实现隐蔽语音控制 [8]。

2. 奇异构造，即物理域输入的特定构造，造成数字域分析识别结果产生错误的带外脆弱性。例如，攻击者基于智能语音识别模型的带外脆弱性，通过在原始音频上叠加微小扰动，构造人耳不易察觉的对抗样本，实现对智能语音识别系统的定向控制，比如将“你好”识别为“开门”[9]。

非功能设计的异常信道指的是功能对象原有正常功能之外的“信号 - 信息”传输信道，其所能造成的脆弱性主要包括：

1. 物理旁路，即利用物理系统正常运行过程中产生的衍生信号，如电磁辐射、功耗、时间、内存状态等，获取敏感信息的带外脆弱性。例如，攻击者可以利用运算单元执行加密操作或机器学习模型执行识别推理时所产生的功耗或电磁辐射信号，推断其所使用的加密密钥 [10] 或者模型参数 [11] 等敏感信息。

2. 逻辑陷门，即在特定输入下，启动预先设置的隐蔽功能，并按预定功能执行操作的带外脆弱性。例如，非法芯片设计厂商在芯片设计阶段故意植入特殊的功能电路，使其在给定的输入激励下，芯片功能改变或失效 [12]。

3. 预设暗道，即利用预先设置的隐蔽信道，在特定介质中向外界传输信息的带外脆弱性。例如，攻击者通过供应链攻击、社会工程学攻击等手段使物理隔离网络内的设备感染恶意软件，然后利用被感染设备中的硬盘灯向外发送隔离设备内的敏感数据，并最终通过摄像头传感器、光传感器等接收信息 [13]。其中利用的光通信信道即为预设暗道。

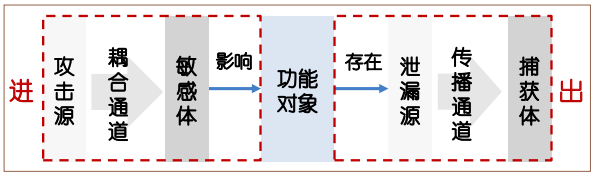

图 4　面向功能对象的带外脆弱性模型示意图

表 1　基于脆弱成因的带外脆弱性分类表

脆弱成因	带外脆弱性	典型攻击案例
跨域交互过程中信号 - 信息映射失配	信号超限	海豚音攻击、硬盘窃听攻击
	奇异构造	图像或语音对抗样本攻击
非功能设计的异常信道	物理旁路	基于侧信道的密钥或模型推断
	逻辑陷门	硬件木马攻击、软件后门攻击
	预设暗道	基于隐蔽信道的敏感信息泄露

已有脆弱性体系主要从结构、协议、软件、硬件等方面总结归纳带内脆弱性。基于以上分析，本文提出“带内带外兼顾”的脆弱性体系，如图 5 所示，为构建更加健全的脆弱性体系提供参考。

展望与思考

现有针对智能系统的带外脆弱性研究还处于早期阶段，本文建议设立专项研究课题，引导学术界和工业界联合资源和力量，在带外脆弱性理论体系研究、量化评估标准建设以及检测产品设计三方面整体布局，增强忧患意识，做到“知其因，看得清；

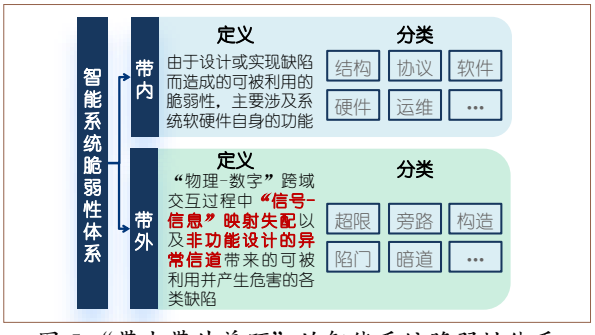

图 5　“带内带外兼顾”的智能系统脆弱性体系

明其害，灭得快”。

1. **加快推进智能系统带外脆弱性研究，建立健全带外脆弱性理论体系**。相比计算机和网络安全领域常见的传统脆弱性，智能系统的带外脆弱性在设计时难以发现或避免，隐蔽性高、危害严重。学术界应聚焦智能系统在物理域和数字域之间的跨域交互过程，根据“攻击源 - 传导通道 - 脆弱点”三元模型，结合智能系统自身“有感有控、AI 驱动”等技术特点进行全面系统的研究与分析；构建智能系统带外脆弱性理论体系，全方位剖析带外脆弱性，为智能系统的安全设计、带外脆弱性的检测与防护提供指导，保障智能系统“感知可信、计算可验、执行可靠”。

2. **推动脆弱性量化评估标准建设，将带外脆弱性测试纳入智能系统出厂安全合规**。虽然智能系统遵循严格的安全设计和检测流程，但系统与物理世界“强交互”的特点导致智能系统的带外脆弱性由于多种内生和外生原因仍然不可避免。在带外脆弱性理论体系的指导下，参考《网络安全等级保护制度》制定设计智能系统“带外脆弱性量化评估标准”，对智能系统出厂进行带外脆弱性测试。大力推动带外脆弱性量化评估标准在物联网、基础设施等重要行业的应用，强化智能系统安全审查力度，做到防患于未然、处置于未萌，始终处于主动地位。

3. **推进带外脆弱性检测工具的设计落地，构建保障智能系统安全的关键防线**。尽管“带外脆弱性量化评估标准”能够在一定程度上减少带外脆弱性的安全隐患，但是智能系统在频繁的跨域交互过程中存在不可预测性和环境复杂性，带外安全问题仍然无法避免。因此，为保障智能系统在数字域和物理域的总体安全，带外脆弱性检测工具将成为保障智能系统总体安全的关键防线。如何设计、完善检测工具并推动相关检测标准体系的建设，是学术界和工业界未来需要重点关注和研究的方向。

总结

带外脆弱性是随着智能系统的发展出现的一类新型脆弱性，主要体现为物理域和数字域跨域交互过程中由于“信号 - 信息”映射失配或非功能设计的异常信道带来的可被利用并产生危害的各类缺陷。相比传统的带内脆弱性，带外脆弱性研究尚处于起步阶段，对其攻击可直接在物理世界造成严重后果，对其形成系统性认知具有重要的研究意义。本文基于对已有脆弱性体系的梳理，认为智能时代的脆弱性体系应该从“带内”向“带内带外兼顾”转变。为此，我们对带外脆弱性体系进行了探索和讨论，并对未来的工作进行了展望。我们坚信，在领域同行的共同努力下，带外脆弱性将会得到有效的认知、研究和解决，助力形成更为健全的系统安全体系与生态。 ■

徐文渊

CCF 高级会员。浙江大学教授，国家杰出青年基金获得者。主要研究方向为物联网安全、语音安全、智能电网安全等。
wyxu@zju.edu.cn

郭世泽

浙江大学求是特聘教授。主要研究方向为信息系统脆弱性分析和网络空间防御。

冀晓宇

CCF 专业会员。浙江大学副教授。主要研究方向为物联网安全、智能电网安全、移动感知安全及隐私等。
xji@zju.edu.cn

其他作者：闫　琛

（本文责任编委：**刘云浩**）

参考文献

[1] 全国科学技术名词审定委员会 . 计算机科学技术名词（第三版）.2018.

[2] Common Vulnerabilities & Exposures[DB/OL]. https://cve.mitre.org.

[3] ISO/IEC TR 13335:1998 Information technology — Guidelines for the management of IT Security[S]. 1998.

[4] GB/T 30279-2013, 信息安全技术 网络安全漏洞分类分级指南 [S]. 2013.

[5] GB/T 30279-2020, 信息安全技术 网络安全漏洞分类分级指南 [S],.2020.

[6] Kune D F, Backes J,Clark S S, et al. Ghost talk: Mitigating EMI signal injection attacks against analog sensors [C]// 2013 IEEE Symposium on Security *and Privacy*. IEEE, 2013:145-159.

[7] Agrawal D, Archambeault B, Rao J R, et al. The EM Side-Channel(s)[C]// Revised Papers from the International Workshop on Cryptographic Hardware & Embedded Systems. Springer-Verlag, 2002.

[8] Zhang G, et al. Dolphinattack: Inaudible voice commands[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.

[9] Nicholas C, and Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[J]. 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.

[10]Giovanni C, et al. Screaming channels: When electromagnetic side channels meet radio transceivers[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018.

[11]Batina L, Bhasin S, Jap D, et al. CSINN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel[C]//28th USENIX Security Symposium (USENIX Security 19). 2019.

[12]Bhunia S, Hsiao M S, Banga M, et al. Hardware Trojan Attacks: Threat Analysis and Countermeasures[J]. Proceedings of the IEEE, 2014, 102(8):1229-1247.

[13]Guri M, Zadov B, Atias E, et al. LED-it-GO: Leaking (a lot of) Data from Air-Gapped Computers via the (small) Hard Drive LED[C]// International Conference on Detecton of Intrrusions and Madware,and Uulnerability Assessmont Springer, Cham, 2017:161-184.