Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 411-419
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152040

Hans汉斯

人工智能技术在网络安全评估中的应用与影响综述

范佳誉

西京学院计算机学院，陕西 西安

收稿日期：2025年12月25日；录用日期：2026年2月14日；发布日期：2026年3月2日

摘　要

本文综述了人工智能技术在网络安全评估领域的应用现状。其核心应用涵盖漏洞检测与管理、入侵检测与防御、风险评估与态势感知等关键方向。在系统梳理各研究领域进展的基础上，详细阐释了人工智能技术在网络安全评估中所发挥的关键作用，包括显著提升检测精度与效率、具备强大的自适应学习能力，以及能够针对研究对象展开全面且深入的分析等。然而，人工智能在网络安全评估实践中仍面临多重挑战：高质量标注数据短缺与隐私泄露风险、模型可解释性不足，以及抗攻击能力薄弱。基于上述现状，本文梳理了人工智能与人协同工作的现有实践模式，总结了联邦学习在网络安全领域的应用进展与拓展潜力，并探讨了量子计算环境下网络安全评估方法的变革趋势。期望通过以上分析，为读者全方位呈现人工智能在网络安全评价中的应用图景，同时为后续相关研究提供参考。

关键词

人工智能，网络安全评估，漏洞检测，入侵防御

A Review of Artificial Intelligence Technology: Applications and Impacts in Cybersecurity Evaluation

Jiayu Fan

School of Computer Science, Xijing University, Xi'an Shaanxi

Received: December 25, 2025; accepted: February 14, 2026; published: March 2, 2026

Abstract

This paper provides a comprehensive review of the application status of artificial intelligence

文章引用: 范佳誉. 人工智能技术在网络安全评估中的应用与影响综述[J]. 人工智能与机器人研究, 2026, 15(2): 411-419. DOI: 10.12677/airr.2026.152040

technology in the field of cybersecurity assessment. Its core applications cover key directions such as vulnerability detection and management, intrusion detection and prevention, as well as risk assessment and situational awareness. Based on a systematic collation of research progress across various fields, this paper elaborates on the pivotal roles played by AI technology in cybersecurity assessment, including significantly improving detection accuracy and efficiency, possessing robust adaptive learning capabilities, and enabling comprehensive and in-depth analysis of assessment objects. However, AI still faces multiple challenges in practical cybersecurity assessment: the shortage of high-quality labeled data, risks of privacy leakage, insufficient model interpretability, and weak adversarial attack resistance. In light of the aforementioned status quo, this paper summarizes the existing practical models of human-AI collaborative work, synthesizes the application progress and expansion potential of federated learning in the cybersecurity domain, and discusses the evolutionary trends of cybersecurity assessment methodologies in the quantum computing era. It is anticipated that through the aforementioned analysis, this paper will comprehensively present the application landscape of AI in cybersecurity assessment and provide references for subsequent related research.

Keywords

Artificial Intelligence, Cybersecurity Assessment, Vulnerability Detection, Intrusion Defense

1. 引言

数字化时代，网络复杂度与规模持续扩大。网络在为用户带来便捷高效体验的同时，也带来了日益严峻的网络安全问题。传统网络安全评估方法存在固有局限且模式相对固定，难以应对当前多样化、智能化的网络威胁与攻击手段。在此背景下，人工智能凭借数据处理和模式识别能力以及持续迭代的自学习能力，正逐步崛起为网络安全评估领域的核心技术[1]。

将人工智能应用于网络安全评估中，不仅弥补了许多传统防护手段的不足，还借助机器学习等先进技术，做到了对网络流量的实时监测和分析，能够快速识别异常行为和潜在安全风险。除此之外，人工智能还能协助安全人员建立起更准确、可信的风险评估模型[2]，进而对潜在网络攻击实现有效预警。随着人工智能技术的持续创新，其在网络安全方面的作用会与日俱增，为我们的数字世界提供更智能、高效和全面的安全保护方案。

2. 人工智能在网络安全评估中的应用

2.1. 漏洞检测与管理

2.1.1. 基于机器学习的漏洞扫描器

基于有监督学习的漏洞扫描策略，以已明确标注的脆弱性特征为训练基础，构建精准的漏洞识别框架。相较于传统基于规则库的方法，该方案不仅能识别特定危险代码片段、网络端口异常等特征[3]，更在检测未知逻辑漏洞和适应漏洞演变方面展现出本质优势。为清晰展示其效能，当前部分模型与传统模型的量化指标对比如表 1 中所示。

与传统的基于静态特征库的漏洞扫描算法(如图 1 所示)相比，机器学习算法展现出了更为强大的适应性与灵活性。它能够更好地契合软件漏洞不断演变、进化的复杂过程，成功发现一些运用传统手段极

难探测到的新型软件漏洞。尤为重要的是，由于机器学习模型具备持续学习、不断优化的能力，能够对自身的检测算法进行持续更新与改进，从而更为有效地适应不断变化的新威胁环境。正因如此，该技术在安全领域的应用正日益广泛，备受关注[4] [5]。

Table 1. Performance comparison table of partial machine learning models and traditional models
表 1. 部分机器学习模型与传统模型性能对比表

模型名称	采用技术	准确率(Accuracy)	误报率(FPR)	F1 值
SySeVR-BGRU	程序语义切片 + BGRU	93.5%	4.2%	0.89
Devign	图神经网络(GNN) + 代码属性图	89.0%	7.8%	0.85
LineVul	CodeBERT + Transformer + 行级粒度	91.0%	6.5%	0.91
GraphCodeBERT	Transformer + 图结构代码	97.5%	4.8%	0.95
SonarQube	传统静态代码分析 + 规则库匹配	75.0%	22.0%	0.68

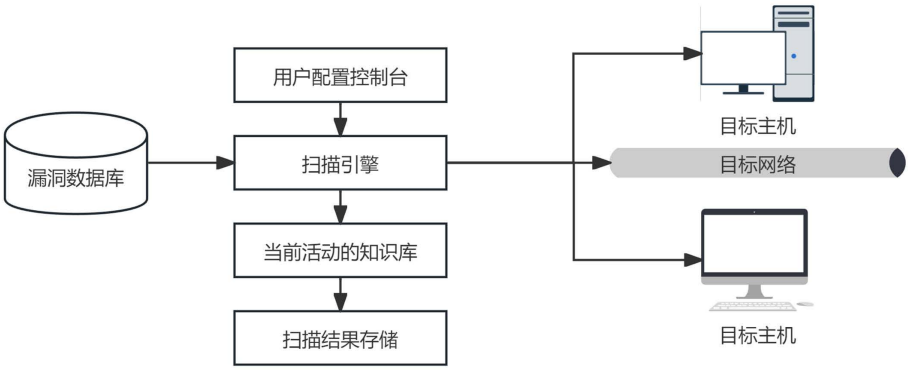

Figure 1. Traditional vulnerability scanning
图 1. 漏洞扫描

2.1.2. 智能漏洞优先级排序

随着数字化转型的深入，网络安全威胁呈现多维度演进特征。当安全团队面临海量的漏洞检测结果时，人工智能通过动态分析与建模，能够超越传统静态评分系统。研究表明，通过分析攻击者对漏洞参数的可控域，可以构建更精细的量化模型，从而有效区分漏洞的实际可利用性差异[6]。

基于多维度的量化分析，智能系统能实现上下文感知的优先级排序。列如，ILLATION 模型融合神经网络与逻辑推理，通过学习特定网络环境下的风险模式和对抗者行为，为管理员提供网络定制化的修复优先级列表，极大提升了修复资源的配置效率[7]，切实提升工作的针对性与有效性。

这种优先级划分是构建主动防御体系的核心。研究指出，未来的优先级排序模型必须是动态、可解释且能集成实时威胁情报。将预测性指标与环境因素结合，是实现精准、自动化威胁响应与修复决策的关键趋势[8]。

2.2. 入侵检测与防御

2.2.1. 入侵检测

无监督学习作为人工智能的重要分支，其中的聚类分析、自编码器等算法在网络异常检测领域已被广泛应用。通过对网络中的正常业务数据进行学习和建模，该方法可以准确地识别出异常流量，这对于是否能够有效检测出网络中潜藏的攻击行为具有重要的意义[9] [10]。

该方法最突出的优势在于其对未知攻击(如零日攻击)的检测能力。传统的基于规则或签名库的入侵检测系统在面对新型攻击时，常因规则更新滞后而失效。而无监督学习算法无需依赖预定义的攻击特征，能够直接从数据分布中识别异常。研究证实，此类方法能够有效提升对未知攻击的检出率，从而系统性地弥补传统方法的缺陷，增强安全防护体系[9]。

2.2.2. 基于深度学习的入侵防御

以卷积神经网络(CNN)、循环神经网络(RNN)及其变体(如 LSTM)为代表的深度学习模型，凭借其强大的高阶特征自动提取与复杂模式学习能力，在网络流量深度分析领域展现出显著优势。与依赖人工规则的传统方法不同，这些模型能够从原始或低层网络数据中自动学习到与攻击行为相关的深层表征，实现对入侵行为的精准识别与分类[11] [12]。当前主流基于深度学习的入侵检测模型性能对比如表 2 所示：

Table 2. Performance comparison table of partial mainstream intrusion detection models and traditional models
表 2. 部分主流入侵检测模型与传统模型性能对比表

模型/技术类型	代表模型技术	准确率(Accuracy)	误报率(FPR)	核心特点
基于 CNN	多层感知机(MLP)	99.44%	0.52%	作为基础深度学习模型，在 NSL-KDD 数据集上达到了极高的精度与极低的误报
基于 RNN/LSTM	循环神经网络(RNN)	98.02%	2.21%	在 NSL-KDD 数据集上处理序列数据的能力较好
混合深度学习	CNN-双向长短期记忆网络(CNN-BiLSTM)	99.89%	未明确报告	综合了 CNN 的空间特征提取和 BiLSTM 的时序建模优势
传统模型	基于规则的 IDS，经典机器学习	约 85%~92%	常>5%	依赖专家规则或浅层特征，对新型、复杂攻击泛化能力弱，但可解释性好，部署简单。

如表 2 所示，混合深度学习架构(如 CNN-BiLSTM)通过结合不同网络的优点，在公开基准数据集上取得了接近 99.9%的检测准确率，显著超越了传统方法。即便是相对基础的多层感知机(MLP)，在精心调优后也能实现超过 99.4%的准确率与低于 0.6%的误报率，这充分证明了深度学习模型从复杂数据中自动学习有效特征、实现精准分类的强大能力。

2.3. 风险评估与态势感知

2.3.1. 动态风险评估模型

通过人工智能技术的强大实时数据分析能力，能够动态追踪和监测网络中各种复杂的变化。如系统配置的升级优化、新应用的部署上线以及外部威胁的演变转化等多个方面。在此基础上，构建起基于实时数据的风险评价模型[13]。根据评估结果和反馈，企业能够灵活科学地调整自己安全策略，来适应复杂多变的网络威胁情况，保证网络安全稳定的运行，为企业的正常运营提供坚实保障。

2.3.2. 态势感知平台

在高校智慧化校园建设不断深入推进、高校信息化水平持续提升的背景下，有学者开展了基于态势感知的高校网络安全工作实践，为高校网络安全建设提供了极具价值的参考范例[14]。

利用先进的人工智能技术，能够高效、有序地集成海量数据源，包括防火墙日志、IDS 报警信息、服务器性能指标等关键数据。通过计算机可视化技术手段，实现对网络安全态势的直观、清晰可视化展示。同时，运用机器学习技术对大数据进行深度挖掘与细致分析，进而对潜在的安全威胁进行科学预测。

3. 人工智能为网络安全评估带来的优势

3.1. 提高检测准确性与效率

人工智能算法凭借其强大的运算能力，在分析与处理海量、复杂且多样化的数据时能够做到高效且精准。基于深度学习的网络安全检测技术，能够在极短时间内就发现网络中隐藏着的安全问题。与传统的检测方法相比，该技术在使检测速度大幅度提升的同时，大幅减少了漏报误报情况。

人工智能在提升安全分析效率的同时，还极大地增强了对未知威胁的预警与防范能力，提高了发现预防未知威胁的能力，为网络安全的可靠稳定打下基础。借助此技术，我们可以更好的看懂数据里隐藏的规律和未来走向，提前做好预防准备。另外，人工智能算法有较强的自主学习适应能力，能在实际使用中持续改进，以更好地满足安全领域日益增长的复杂需求[15]。

3.2. 自适应学习能力

为了有效处理复杂多变的网络安全威胁，人工智能必须具备自主学习和快速适应新攻击方式与脆弱性特征的能力。为了实现该目标目前一系列新型智能算法应运而生，这些算法具备自主学习、最优解自寻的特性，可以持续对当前网络安全问题进行动态调整和改进。

网络攻击者常常使用病毒或木马、钓鱼网站、DDoS 攻击等手段，这些手段对网络系统的运行与数据的安全构成了严重的威胁。传统网络安全防御方法在碰到此类情况时，很难做到有效的处理。而人工智能凭借其优秀的自适应学习能力，可针对性地对威胁进行处理，在网络安全领域发挥出了巨大的潜力。

3.3. 提供全面深度的分析

通过对不同来源的数据进行综合使用和整合分析，人工智能能够全方面对网络安全态势进行剖析。这种方式不但能有效找出那些浅显的问题，还能运用先进技术手段深入挖掘来找到隐藏在大量数据里的复杂风险联系。

人工智能给制定精准高效的安全策略提供了巨大的帮助，为相关机构及企业防范网络威胁提供了有效的助力，通过机器学习、模式识别等前沿技术还可以将这一卓越能力不断提升，以更好地应对网络中层出不穷的攻击方式[16] [17]。

4. 人工智能在网络安全评估中面临的挑战

4.1. 数据质量与隐私问题

4.1.1. 高质量标注数据的短缺

人工智能模型的训练效能与标注数据质量直接相关。网络安全领域因攻击数据的敏感性特征，导致行业级标注数据集存在显著缺口：一方面，受攻击数据涉及系统脆弱性信息，机构普遍存在共享顾虑；另一方面，攻击模式的多态性使人工标注面临语义歧义挑战，标注一致性误差率可达 18.7%，这无疑进一步加剧了构建有效模型的难度[18]。

4.1.2. 数据不平衡问题严重

在真实的网络环境中，网络安全数据往往呈现明显的类别不平衡现象，即正常数据流量远多于异常或攻击数据流量。这种数据分布不均衡会导致模型在训练过程中过度偏向多数类样本，从而降低对少数类(如攻击行为)的识别能力，影响入侵检测系统的实际应用效果。结合改进 SMOTE 与 GA-XGBoost 的方法被提出来解决这一问题，通过对少数类样本进行局部离群因子引导的过采样，并对多数类样本进行随机欠采样，有效实现了数据再平衡，提升了模型对异常流量的分类性能[19]。实验结果表明，该方法在

UNSW-NB15 数据集上取得了 97.40%的准确率与 70.2%的平均召回率，验证了其在不平衡网络安全数据分类中的有效性。

4.1.3. 数据隐私风险

人工智能在网络安全评估中需处理大量敏感数据，包括用户网络行为日志、系统漏洞信息、攻击流量特征等，这些数据往往涉及个人隐私、商业机密等信息[20]。

在数据采集、传输、存储和模型训练过程中，若缺乏有效的隐私保护机制，极易引发数据泄露和合规风险。传统的匿名化与数据脱敏技术面临严峻挑战，攻击者可通过模型逆向攻击、成员推断攻击等方式，从训练好的 AI 模型中反推出部分原始训练数据，造成再泄露[21]。部分针对入侵检测模型的梯度泄露攻击可成功恢复特定攻击流量特征，对关键基础设施的隐蔽威胁检测构成严重隐患[22]。

针对上述隐私挑战，当前学术界主要通过隐私增强计算和去中心化学习框架提供解决思路。基于同态加密的推理框架允许直接在密文上执行检测，从根本上防止数据泄露[23]。差分隐私机制通过添加校准噪声，在保护个体记录的同时维持模型性能[24]。联邦学习作为主流分布式架构，通过本地训练与参数聚合实现“数据不动模型动”，有效减少原始数据暴露[25]。同时，针对联邦学习中的恶意参与者问题，可验证的参与方选择机制被提出以保障训练安全[26]。此外，采用对抗训练与知识蒸馏等方法，能有效防御模型逆向与成员推断攻击，增强模型的隐私鲁棒性[27] [28]。

4.2. 模型可解释性差

以深度学习为典型代表的多种高级人工智能技术，其决策过程犹如一个难以洞悉的“黑箱”，内部运行机理至今尚不清晰，难以从理论层面直接、清晰地揭示其内在逻辑与运作机制。

近年来，针对这一“黑箱”问题，学术界的研究主要围绕开发和应用可解释人工智能(XAI)方法展开，旨在为模型的决策提供足够的可读性。Kazi Fatema 等人[29]在本地模型层面集成人工神经网络与 SHAP 解释方法，在保证隐私和可扩展性的同时，提供公平、可解释的决策过程。Matteo Brosolo 等人[30]综合使用遮挡图、HiResCAM 等多种 XAI 工具可视化 CNN 的决策依据，同时还提出了提升模型抗混淆鲁棒性策略。任昊等人[2]则提出了一种隐私保护的深度神经模糊推理系统(PrivDNFIS)，通过将模糊逻辑的可解释规则与深度学习相融合，构建了原生可解释的入侵检测模型。

5. 未来发展趋势

5.1. 人工智能与人类专家协作

在未来，人工智能和人类专家合作将作为一个主要的发展方向[31]。在大数据背景下，人工智能凭借强大的计算能力，能够快速和全面地进行检测大量数据。而具有较深专业知识和丰富经验的安全专家，可以对复杂和细微的安全问题，进行更深入的分析判断[32]。

安全人员需要仔细核对人工智能给出的建议，确保其准确可靠。在此基础上构建人和机器协同合作的新方法，让双方在实际应用中发挥各自长处，形成协同共进、优势互补的工作格局，为网络安全提供坚实保障[33]。

5.2. 联邦学习在网络安全中的应用拓展

自 2016 年 Google 的 AI 团队提出在移动互联网手机终端使用联邦学习算法框架用来隐私保护[34]起，联邦学习开始在机器学习领域兴起。对等的邦联学习框架如图 2 所示，它能够应对与隐私保护、通信成本及可扩展性相关的挑战。该技术仅传输模型更新参数，从而保护设备生成的敏感信息安全[35]。这种分布式的学习方法，不仅能够减少对集中存储数据的需要，还可实现跨领域、平台合作，吸引各类企

业和机构参与，提升网络安全防护能力[36]。

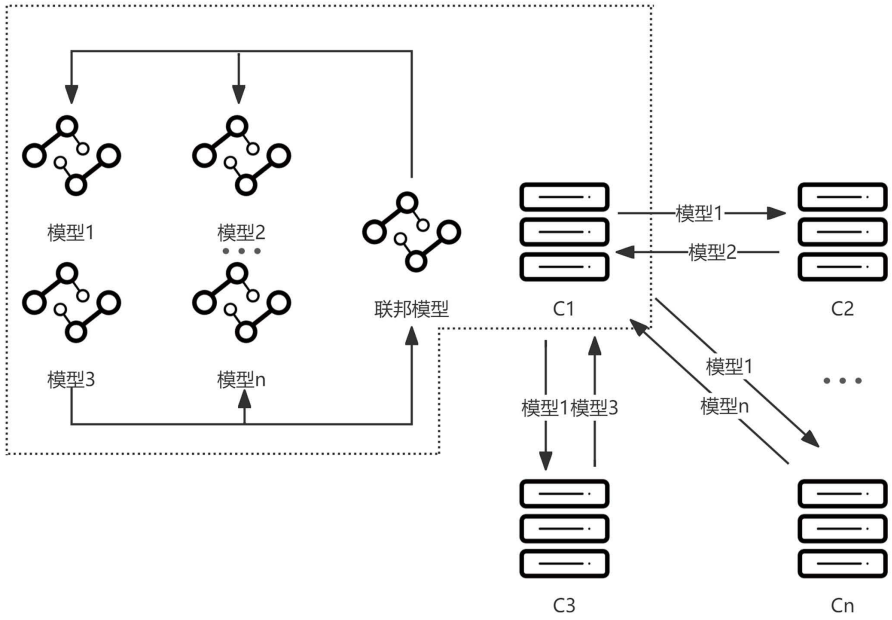

Figure 2. Peer-to-peer federated learning framework
图 2. 对等的联邦学习框架

在相关研究成果方面，Jiang 等[37]针对长尾数据与拜占庭节点攻击共存的联邦学习场景，提出含“智囊团”角色的双层聚合方法，通过过滤层与投票层双重验证实现恶意模型剔除与长尾数据价值保留。实验显示，该方法在 MNIST 长尾数据集上较 multi-Krum 算法准确率提升 9%以上，且能有效抵御拜占庭节点攻击。余锋等[38]针对联邦学习中的梯度泄露问题，提出基于差分隐私生成对抗网络(DP-enabled GAN)的隐私增强方案，相比 DP-SGD 方法，在 MNIST、Fashion-MNIST 数据集上分类准确率显著提升，实现了隐私保护与模型实用性的平衡。Li 等[39]提出可学习聚合权重的联邦学习方法 FEDLAW，突破权重归一化惯例，通过全局权重收缩与客户端一致性优化提升模型泛化能力。该方法在多数据集和模型上表现优异，且对代理数据集偏移、恶意客户端等场景具有强鲁棒性。

随着技术发展还有应用的普及，联邦学习有望成为安全领域重要支撑力量，为构建可靠且稳定的数字世界打下良好的基础。

5.3. 量子计算时代的网络安全评估变革

在量子计算技术从理论研究逐步走向实际应用这一背景下，人工智能在量子安全评估中的作用日益凸显，它不仅能够辅助设计抗量子攻击的新型密码算法，还能给量子计算环境下的入侵检测和风险分析带来新途径[40] [41]。

随着量子计算技术的不断发展，通过量子信息展开攻击将会越来越普遍，这不仅要求我们对现有的加密技术进行全面升级与改进，还需要对已有的安安全协议进行彻底的评估和调整。在这一过程中，利用机器学习或深度学习等技术，对可能存在的量子威胁进行预测和模拟[42] [43]，Schuld 等[44]针对传统核方法在高维数据场景下计算复杂度高、核函数估计成本高的问题，提出将量子态空间用作特征空间的监督学习框架，含量子变分分类器与量子核估计器两种算法。该框架借助量子纠缠与干涉特性实现高效数据编码，为噪声中尺度量子设备的机器学习应用提供了可行路径。

6. 结论

网络安全评估因人工智能技术的发展正经历着一场深刻的变革。它凭借强大的数据挖掘和分析能力，能深度分析海量复杂的数据，精准识别网络中的异常行为和存在的漏洞，快速给出有效的解决措施。将人工智能与网络安全相融合，将显著提升事故响应速度，降低对人工的依赖，使网络安全防护更加智能与高效。

必须承认，当前人工智能应用在网络安全上还有很多棘手的问题。数据收集和处理上，数据获取敏感性高，来源渠道少，质量起伏较大，要有效利用十分困难；建立模型方面，以深度学习为代表的模型决策过程像“黑箱”，解释性弱，给安全人员理解和应用模型带来挑战；同时，黑客针对人工智能系统的攻击手段层出不穷，使得人工智能系统的抗攻击能力成为亟待攻克的短板。

随着人工智能技术不断突破和尝试新的应用方式，它在网络安全领域的发展前景依然十分广阔。它将与人类智能深度融合，形成优势互补的人机协作模式，打造更可靠的网络安全防御体系。随着科技成熟和应用范围的拓展，人工智能成为网络安全领域的核心支柱只是时间问题，在未来它将为保护个人、企业隐私安全及国家安全注入强大动力，让网络安全进入新的阶段。

参考文献

[1] Arreche, O. and Abdallah, M. (2024) A Comparative Analysis of DNN-Based White-Box Explainable AI Methods in Network Security. https://arxiv.org/abs/2501.07801

[2] Ren, H., Lan, X., Tang, R. and Chen, X. (2025) PrivDNFIS: Privacy-Preserving and Efficient Deep Neuro-Fuzzy Inference System. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 20174-20182. https://doi.org/10.1609/aaai.v39i19.34222

[3] 陈朗, 王春玲. 基于机器学习的 Android 系统漏洞扫描处理系统设计[J]. 电脑知识与技术, 2019, 15(25): 20-22.

[4] 张锦蓉, 刘伟民. 基于漏洞扫描的网络安全维护策略探讨[J]. 信息与电脑(理论版), 2024, 36(22): 89-91.

[5] 曹文, 胡志锋, 代飞. 基于 Python 的通信网络安全漏洞扫描技术研究与实现[J]. 电脑编程技巧与维护, 2024(11): 171-173.

[6] Lacombe, G. and Sébastien, B. (2025) Attacker Control and Bug Prioritization. https://arxiv.org/abs/2501.17740

[7] Zeng, Z., Huang, D., Xue, G., Deng, Y., Vadnere, N. and Xie, L. (2024) ILLATION: Improving Vulnerability Risk Prioritization by Learning from Network. *IEEE Transactions on Dependable and Secure Computing*, **21**, 1890-1901. https://doi.org/10.1109/tdsc.2023.3294433

[8] Jiang, Y., Oo, N., Meng, Q., *et al*. (2025) A Survey on Vulnerability Prioritization: Taxonomy, Metrics, and Research Challenges. https://arxiv.org/abs/2502.11070

[9] 尹梓诺, 陈鸿昶, 马海龙, 等. 无监督自适应抽样与改进孪生网络结合的网络流量异常检测方法[J]. 电子与信息学报, 2025, 47(7): 2211-2224.

[10] 池彬, 胡辉, 周天宇, 等. 一种改进自编码器与流特征结合的入侵检测方法[J]. 重庆理工大学学报(自然科学), 2025, 39(7): 119-126.

[11] 史承斌. 基于深度学习的网络入侵检测与防御机制[J]. 无线互联科技, 2024, 21(14): 123-125.

[12] 陈智勇. 基于深度学习的网络入侵检测与防御研究[J]. 无线互联科技, 2023, 20(19): 152-154.

[13] 廖天颖, 杨斯博, 窦润亮. 基于贝叶斯网络的大数据安全动态风险评估模型研究[J]. 网络空间安全, 2023, 14(1): 60-68.

[14] 张小雷. 基于态势感知的高校网络安全实践探索[J]. 网络安全技术与应用, 2024(11): 64-66.

[15] 施雪清. 基于人工智能技术的计算机网络安全风险评估系统设计[J]. 信息与电脑(理论版), 2023, 35(23): 199-202.

[16] 亓文法. 基于人工智能的网络安全态势评估技术综述及展望[J]. 保密科学技术, 2024(10): 42-48.

[17] 王书义. 人工智能驱动的网络安全威胁检测与防御策略[J]. 信息记录材料, 2025, 26(8): 40-42.

[18] Goldschmidt, P. and Chudá, D. (2025) Network Intrusion Datasets: A Survey, Limitations, and Recommendations. https://arxiv.org/abs/2502.06688

[19] 韩凤董, 宗学军, 何戡, 等. 面向网络安全不平衡数据的特征学习和分类研究应用[J]. 科学技术与工程, 2023, 23(3): 1130-1137.

[20] Du, M., Li, F., Zheng, G. and Srikumar, V. (2017) DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. *Proceedings of the* 2017 *ACM SIGSAC Conference on Computer and Communications Security*, Dallas, 30 October-3 November 2017, 1285-1298. https://doi.org/10.1145/3133956.3134015

[21] Chen, H. and Babar, M.A. (2024) Security for Machine Learning-Based Software Systems: A Survey of Threats, Practices, and Challenges. *ACM Computing Surveys*, **56**, 1-38. https://doi.org/10.1145/3638531

[22] 王子帆. 基于梯度反演的模型隐私攻击及防御方法研究[D]: [硕士学位论文]. 贵阳: 贵州大学, 2023.

[23] 周炜, 王超, 徐剑, 胡克勇, 王金龙. 基于区块链的隐私保护去中心化联邦学习模型[J]. 计算机研究与发展, 2022, 59(11): 2423-2436.

[24] Hu, W. and Fang, H. (2024) Towards Differential Privacy in Sequential Recommendation: A Noisy Graph Neural Network Approach. *ACM Transactions on Knowledge Discovery from Data*, **18**, 1-21. https://doi.org/10.1145/3643821

[25] Kairouz, P. and McMahan, H.B. (2021) Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, **14**, 1-210. https://doi.org/10.1561/2200000083

[26] Weng, J., Weng, J., Zhang, J., *et al*. (2021) DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-Based Incentive. *IEEE Transactions on Dependable and Secure Computing*, **18**, 2568-2582.

[27] Nasr, M., Songi, S., Thakurta, A., Papernot, N. and Carlin, N. (2021) Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. 2021 *IEEE Symposium on Security and Privacy* (*SP*), San Francisco, 24-27 May 2021, 866-882. https://doi.org/10.1109/sp40001.2021.00069

[28] Mora, A., Tenison, I., Bellavista, P., *et al*. (2022) Knowledge Distillation for Federated Learning: A Practical Guide. https://arxiv.org/abs/2211.04742

[29] Fatema, K., Anannya, M., Dey, S.K., Su, C. and Mazumder, R. (2024) Securing Networks: A Deep Learning Approach with Explainable AI (XAI) and Federated Learning for Intrusion Detection. In: *Lecture Notes in Computer Science*, Springer, 260-275. https://doi.org/10.1007/978-981-97-8540-7_16

[30] Brosolo, M., Puthuvath, V. and Conti, M. (2025) The Road Less Traveled: Investigating Robustness and Explainability in CNN Malware Detection. https://arxiv.org/abs/2503.01391

[31] 马春来, 王群, 孙中豪, 等. 基于人机协作迭代分析的网络协议逆向方法[J]. 信息对抗技术, 2024, 3(5): 84-96.

[32] Singh, N. (2025) Enhancing Search and Discovery: The Synergistic Collaboration between Humans and AI. *European Journal of Computer Science and Information Technology*, **13**, 112-123. https://doi.org/10.37745/ejcsit.2013/vol13n10112123

[33] Arker, I.H., Janicke, H., Mohammad, N., *et al*. (2023) AI Potentiality and Awareness: A Position Paper from the Perspective of Human-AI Teaming in Cybersecurity. https://arxiv.org/abs/2310.12162

[34] Mcmahan, H.B., Moore, E., Ramage, D., *et al*. (2016) Communication-Efficient Learning of Deep Networks from Decentralized Data. https://arxiv.org/abs/1602.05629

[35] Pachar, S., Dhabhai, A., Vali, S.M., Sharma, D., Yadav, S. and Khatoon, A. (2024) A Survey of Federated Learning for Internet of Things: Recent Advances, Research Problems and Solutions. 2024 International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation (*ARIIA*), Manipal, 20-21 December 2024, 1-4.

[36] 康海燕, 张聪明. 基于联邦学习的自适应网络攻击分析方法研究[J]. 信息安全研究, 2024, 10(12): 1091-1099.

[37] Jiang, Y., Ma, B., Wang, X., *et al*. (2023) A Secure Aggregation for Federated Learning on Long-Tailed Data. https://arxiv.org/abs/2307.08324

[38] 余锋, 林庆新, 林晖, 等. 基于生成对抗网络的隐私增强联邦学习方案[J]. 网络与信息安全学报, 2023, 9(3): 113-122.

[39] Li, Z., Lin, T., Shang, X., *et al*. (2023) Revisiting Weighted Aggregation in Federated Learning with Neural Networks. https://arxiv.org/abs/2302.10911

[40] Preskill, J. (2018) Quantum Computing in the NISQ Era and beyond. *Quantum*, **2**, Article 79. https://doi.org/10.22331/q-2018-08-06-79

[41] 王宝楠, 胡风, 张焕国, 等. 从演化密码到量子人工智能密码综述[J]. 计算机研究与发展, 2019, 56(10): 2112-2134.

[42] 张梓钧. 量子通信对现代电信网络安全的影响分析[J]. 集成电路应用, 2025, 42(1): 126-127.

[43] 张燕. 量子技术在通信网络安全方面的应用[J]. 电气自动化, 2022, 44(2): 72-74+77.

[44] Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., *et al*. (2019) Supervised Learning with Quantum-Enhanced Feature Spaces. *Nature*, **567**, 209-212. https://doi.org/10.1038/s41586-019-0980-2