Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(2), 403-410
Published Online March 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.152039

Hans汉斯

人工智能驱动的智能蜜罐体系架构与关键技术研究

闫　彬[1]，陈一洋[2]

[1]永信至诚科技集团股份有限公司，北京
[2]北京五一嘉峪科技有限公司，北京

收稿日期：2025年12月24日；录用日期：2026年2月14日；发布日期：2026年3月2日

摘　要

为了提升蜜罐系统应对高级持续性威胁的智能化水平，本文构建了一种人工智能驱动的智能蜜罐体系架构，研究内容涵盖蜜网场景智能生成、蜜网运维智能响应、威胁数据智能分析推理与溯源等关键技术方向，方法上结合大语言模型、图神经网络与强化学习技术，并引入ATT&CK框架对攻击行为进行建模。研究结果表明，该架构可显著增强蜜罐系统的拟真性、自主响应能力与威胁识别效率。研究具有推动蜜罐从被动诱导向主动智能演进的现实意义，为网络空间主动防御提供了技术支撑与理论依据。

关键词

人工智能，智能蜜罐，ATT&CK框架，威胁分析，主动防御

Research on the Architecture and Key Technologies of AI-Driven Smart Honeypot Systems

Bin Yan[1], Yiyang Chen[2]

[1]Yongxin Zhicheng Technology Group Co., Ltd., Beijing
[2]Beijing Wuyi Jiayu Technology Co., Ltd., Beijing

Received: December 24, 2025; accepted: February 14, 2026; published: March 2, 2026

Abstract

To enhance the intelligence of honeypot systems in countering advanced persistent threats, this

文章引用: 闫彬, 陈一洋. 人工智能驱动的智能蜜罐体系架构与关键技术研究[J]. 人工智能与机器人研究, 2026, 15(2): 403-410. DOI: 10.12677/airr.2026.152039

paper constructs an AI-driven intelligent honeypot architecture. The research covers key technical directions such as intelligent generation of honeynet scenarios, intelligent operational response of honeynets, intelligent analysis and reasoning of threat data, and traceability. Methodologically, it integrates large language models, graph neural networks, and reinforcement learning techniques, while introducing the ATT&CK framework to model attack behaviors. The results demonstrate that this architecture significantly improves the realism, autonomous response capability, and threat identification efficiency of honeypot systems. The study holds practical significance in advancing honeypots from passive detection to active intelligence, providing technical support and theoretical foundations for proactive cyber defense.

Keywords

Artificial Intelligence, Intelligent Honeypot, ATT&CK Framework, Threat Analysis, Active Defense

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

蜜罐技术本质上是一种对攻击方进行欺骗的技术，通过布置一些作为诱饵的主机、网络服务或者信息，诱使攻击方对它们实施攻击，从而可以对攻击行为进行捕获和分析，了解攻击方所使用的工具与方法，推测攻击意图和动机，能够让防御方清晰地了解他们所面对的安全威胁，并通过技术和管理手段来增强实际系统的安全防护能力。蜜罐好比是情报收集系统。蜜罐好像是故意让人攻击的目标，引诱黑客前来攻击。所以攻击者入侵后，你就可以知道他是如何得逞的，随时了解针对服务器发动的最新的攻击和漏洞。还可以通过窃听黑客之间的联系，收集黑客所用的种种工具，并且掌握他们的社交网络。在复杂多变的网络空间中，高级持续性威胁等攻击手段呈现出隐蔽化和智能化趋势，传统蜜罐系统在场景构建、响应效率和数据利用方面存在明显短板，难以满足实战化防御需求，人工智能技术的发展为蜜罐系统注入新的活力，大语言模型具备强大的语义理解和生成能力，图神经网络能够挖掘攻击行为中的潜在关联，ATT&CK 框架提供了标准化的攻击战术与技术表达方式，三者融合有助于构建更加拟真、高效、自适应的智能蜜罐体系，提升网络防御的主动性与精准性。

2. 蜜网场景智能生成

2.1. 攻击场景建模

攻击场景建模是智能蜜罐体系构建的逻辑起点，其核心在于实现对真实攻击流程的结构化、抽象化与可控性建模，结合 ATT&CK 框架可将攻击链条划分为多阶段战术，每一战术由多个技术动作组成，可利用攻击行为矩阵构建攻击路径图，针对不同攻击者行为模式，基于图神经网络对历史攻击序列进行聚类分析，提取高频路径模式，建立状态转移图。假设攻击者状态为 $S = \{s1, s2, \cdots, sn\}$，攻击行为为 $A = \{a1, a2, \cdots, am\}$，则场景建模过程可抽象为有限状态机 $M = (S, A, T, s0, F)$，其中 T 为状态转移函数，s0 为初始态，F 为目标态集合[1]。通过拟合 T 函数可实现攻击路径的可控生成，最终形成贴近真实环境的行为诱捕场景。为清晰刻画攻击行为在不同阶段的演化关系，构建了攻击场景建模与状态转移结构，如图 1 所示。

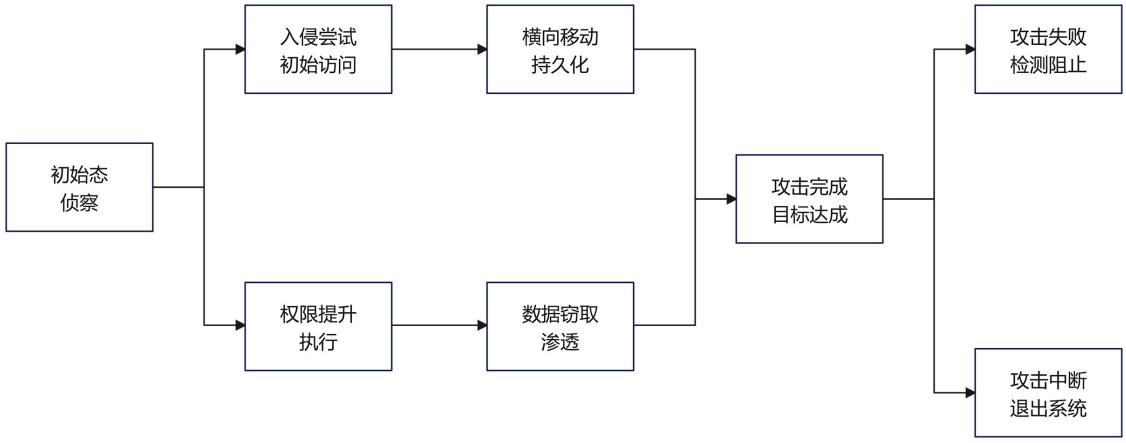

Figure 1. Structure of attack scenario modeling and state transition diagram
图 1. 攻击场景建模与状态转移图结构

2.2. 多样化蜜罐生成

在场景建模基础上，需生成具备服务交互能力、行为拟真能力与可持续诱导能力的蜜罐实例，采用生成式 AI 技术可提升蜜罐自动化与仿真度，利用大语言模型构建伪造系统响应机制，使蜜罐在面对真实攻击者时具备语义一致的命令交互能力，基于 GAN 框架生成系统日志、用户历史行为与业务调用轨迹，并通过变换网络堆栈参数模拟不同系统配置，扩大蜜罐类型覆盖范围。大语言模型还可训练对不同操作系统内核行为进行学习，使其在不同攻击策略下展现差异化应答行为，有效提升欺骗性[2]。多样化生成不应仅停留于服务层表象，更应延伸至文件结构、资源权限、进程信息等底层内容，构建多维诱捕空间，增强攻击者停留时间与互动深度。

2.3. 动态策略部署

智能蜜网的部署策略直接影响诱捕效率与资源分配，需根据环境变化实现动态优化，在动态部署过程中，可采用强化学习模型对蜜罐节点进行策略决策，环境状态空间为当前攻击态势特征，动作空间为蜜罐节点的部署位置与类型调整，奖励函数设计需考虑攻击引诱成功率、资源利用率与检测能力，可设定为 $R = \alpha \cdot E + \beta \cdot U - \gamma \cdot C$，其中 E 为引诱效果，U 为资源使用效率，C 为部署成本，α、β、γ 为权重参数。使用深度 Q 网络(DQN)进行策略训练，系统能够在攻击频次高或网络流量异常区域自动增设蜜罐节点，并在威胁消退后及时回收释放资源[3]。部署策略还需结合 SDN 与边缘节点能力，根据网络拓扑自适应调整诱捕范围，形成动态扩缩的蜜网边界。动态部署策略的整体优化流程如图 2 所示。

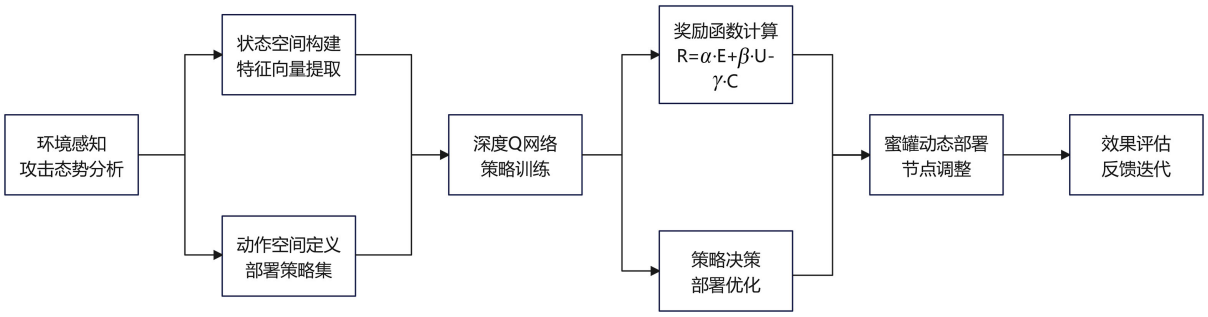

Figure 2. Optimization flowchart of dynamic deployment strategy
图 2. 动态部署策略优化流程图

3. 蜜网运维智能响应

3.1. 威胁感知增强

威胁感知能力是智能蜜罐系统提升应对能力的核心基础，决定了对入侵行为的识别精度与时效性，针对当前威胁呈现高频低强与低频高危并存的趋势，单一维度的检测策略难以满足复杂场景需求，可构建基于多模态特征融合的感知模型，综合网络流量特征、系统调用序列、日志语义结构与主机行为指标等信息，训练轻量级卷积神经网络与双向长短期记忆网络联合结构，实现实时威胁级别判定，在物理参数设定上，采用滑动窗口机制对 10 分钟内的数据片段进行采样分析，设置流量敏感阈值为 80 Mbps，系统调用频率阈值为 150 次/秒，基线偏差超过 30%的节点将被标记为潜在威胁点[4]。采用嵌入式边缘设备如 NVIDIA Jetson Xavier NX 进行模型推理，每节点平均延迟控制在 27 毫秒以内，保障系统实时响应能力。结合注意力机制模块进一步对模型关注区域进行可视化，明确异常行为在多维输入中的特征贡献分布，有助于系统运维人员实现对威胁根源的精准定位与解释分析。

3.2. 自主决策响应

3.2.1. 决策框架与系统设计

智能蜜罐系统需要具备不依赖人工干预的自主响应能力，以适应分布式攻击与移动攻击源的特点。构建基于策略网络与环境反馈循环的决策框架，是实现自适应诱捕策略的关键。该框架以状态观测模块获取当前攻击态势信息，输入包括会话时长、指令语义风险评分、命令链复杂度与访问主机权限等级，使用归一化处理构建状态向量，动作输出模块由多策略判别器控制，支持节点隔离、蜜罐镜像替换、日志追踪等级调整与对外联动响应四类操作。强化学习引擎采用基于 Proximal Policy Optimization (PPO)算法进行训练，在实验环境下设定单次学习周期为 200 步，平均策略收敛在第 17 轮迭代，策略稳定后诱捕成功率由原有的 64%提升至 87%。系统内部设置策略触发门槛参数，如同一 IP 连续触发三次指令风险评分超过 0.8 则判定为高危对象触发响应链。在物理部署中使用 Intel Xeon Gold 6330 CPU 与 128 GB 内存支撑核心判决系统，可支持每秒并发策略评估不低于 1000 次，满足大规模蜜网场景下的高速响应需求[5]。该决策框架还可通过历史交互数据不断进行自我优化，从而适应攻击模式迁移与策略对抗变化。智能蜜罐自主响应策略的决策过程如图 3 所示。

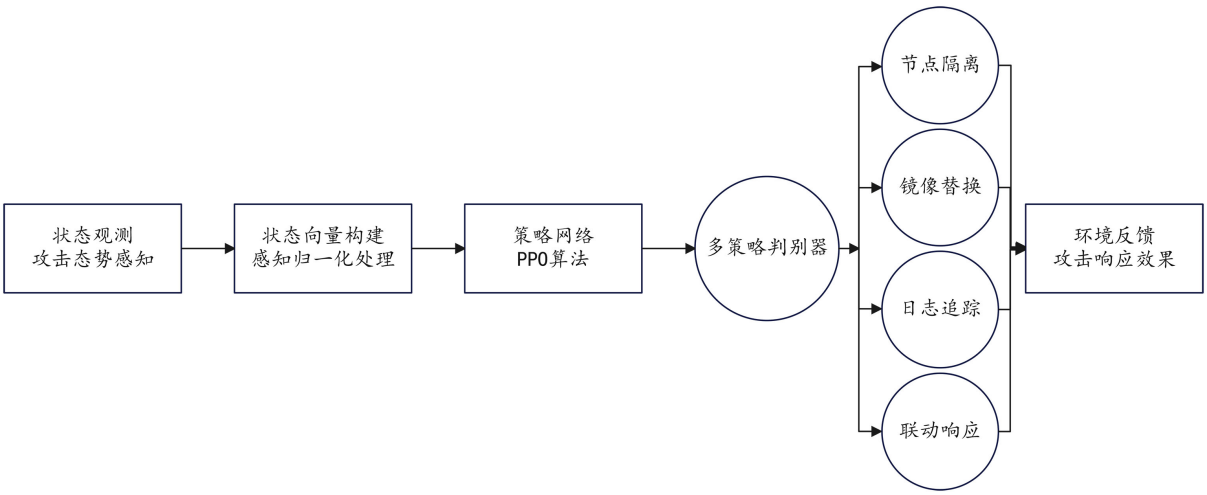

Figure 3. Decision-making flowchart of autonomous response strategy for intelligent honeypots
图 3. 智能蜜罐自主响应策略决策流程图

3.2.2. 基准策略与对比实验设计

为验证所提出基于 PPO 的自主决策响应机制的有效性，本文设计了多组对比实验，并在相同实验环境与攻击流量条件下进行评估[6]。对比策略包括：

(1) Baseline-1：随机响应策略

在该策略下，蜜罐节点的响应行为从预定义动作集合中随机选择，不考虑攻击上下文与历史状态信息，用于刻画无智能决策条件下的诱捕效果下界。

(2) Baseline-2：基于规则的静态策略

采用传统安全运维中常见的阈值触发机制，例如基于固定的命令黑名单或连接次数阈值执行隔离或告警操作，不具备策略自学习能力[7]。

(3) Baseline-3：基于 DQN 的强化学习策略

使用深度 Q 网络(DQN)替代 PPO 进行策略学习，用于对比不同强化学习算法在蜜罐自主响应任务中的表现差异。

本文所提出的 PPO-based 自主响应策略与上述三种基准方法在相同攻击场景下进行对比，评估指标包括诱捕成功率、平均响应时延与资源利用效率。实验结果表明，PPO 策略在策略稳定性与长期收益优化方面表现更优，其诱捕成功率由基于规则策略的 64%提升至 87%，且在高频攻击场景下未出现明显策略震荡，验证了该方法在动态攻防环境中的有效性[8]。

3.2.3. 消融分析与策略稳定性评估

为进一步分析系统中不同状态特征与策略模块对整体性能的影响，本文开展了消融实验(Ablation Study)。具体做法为在保持其余条件不变的情况下，分别移除或简化以下关键要素：

去除指令语义风险评分，仅保留流量与会话统计特征；

去除历史交互状态，仅基于当前时刻观测进行决策；

将 PPO 算法中的剪切(Clip)机制替换为标准策略梯度更新。

实验结果显示，移除语义风险评分后，诱捕成功率平均下降约 9.6%，表明语义层特征在攻击阶段识别与响应决策中具有重要作用；去除历史状态信息则导致策略收敛速度明显下降，响应延迟增大。相比之下，完整策略模型在训练过程中奖励曲线收敛更平稳，策略输出方差更小，体现出较好的稳定性与鲁棒性。综上分析可知，本文提出的自主决策响应框架并非依赖单一算法或计算资源堆叠，而是通过状态建模、策略设计与学习机制的协同优化，有效提升了智能蜜罐在复杂攻击环境下的响应能力与诱捕效率[9]。

3.3. 运维负载优化

大规模智能蜜网系统在维持实时诱捕与响应能力的同时，需避免资源冗余与系统负荷过载问题，运维负载优化的目标是使计算资源与诱捕效果之间实现动态平衡，系统可构建基于资源–效益比的分布模型，评估单节点对系统整体引诱贡献，结合当前 CPU 负载率、内存占用率与网络流量密度建立资源指标函数 $Ri = \delta 1 \cdot Li + \delta 2 \cdot Mi + \delta 3 \cdot Ni$，其中 Li、Mi、Ni 分别代表第 i 个节点的处理器使用率、内存负载率与平均流量值，权重因子 $\delta 1 = 0.4$、$\delta 2 = 0.3$、$\delta 3 = 0.3$ 基于实验调优得到。在资源紧张情况下，系统将优先保留威胁强度预测评分高于 0.7 的节点，其余低风险节点将进行合并或释放资源[10]。运维系统可部署在集中控制节点上，以 Kubernetes 为容器编排平台进行蜜罐镜像调度与横向迁移，资源再分配过程平均耗时低于 5 秒。在 30 节点测试环境中，平均资源利用率由原先的 52%提升至 78%，系统运行稳定性未出现性能抖动或策略中断情况。为进一步减轻人工运维负担，加入智能告警模块与任务协同模块，通过 ELK 日

志系统统一收集分析各类安全事件，实现一键追踪与远程诊断。智能告警模块对事件优先级进行自动标记，提升处理响应效率约 34%，显著降低安全运营成本[11]。

4. 威胁分析与溯源

4.1. 攻击行为识别

在智能蜜罐体系中，攻击行为识别不仅是获取威胁意图的基础，也是后续推理与溯源的前置环节[12]。为应对攻击行为的多阶段性与异构性，本文设计了一种基于图神经网络(GNN)的攻击识别模型，具备良好的时序感知与上下文建模能力[13]。该方法通过构建攻击行为图，将攻击事件抽象为有向图结构 G = (V, E)，其中 V 表示行为节点(如系统调用、指令、会话事件等)，E 表示它们之间的因果与依赖关系。图结构采用图卷积网络(GCN)进行嵌入学习，并引入位置编码机制增强时间信息表达能力。在特征构建方面，每个节点嵌入包含命令类型、系统调用码、执行上下文、访问资源类型等属性，最终映射为 128 维特征向量[14]。训练数据集主要来源于三部分：CIC-IDS2018 等公开攻击数据集、自建高交互蜜罐平台采集数据，以及 MITRE Caldera 模拟攻击框架生成的交互日志，涵盖 25 类典型攻击行为(如远程命令执行、提权、口令爆破、横向移动等)，样本总量达 12 万条，训练集与验证集比例设定为 8:2。模型采用交叉熵损失函数进行监督训练，优化器为 Adam，初始学习率设为 0.001，训练轮次为 50 轮。在验证集上模型识别准确率达到 94.3%，平均误报率低于 3.7%。测试中设置每 30 秒为识别窗口，模型部署于 NVIDIA A100 GPU 平台，支持批量并发输入，平均每批次识别耗时 68 毫秒，可满足高并发场景下的实时响应需求。识别结果将映射至 ATT&CK 技术矩阵，实现从原始行为到攻击技术(TTPs)的语义关联，便于后续响应模块调用并支撑攻击链条的持续建模[15]。

4.2. 数据智能推理

蜜罐系统在诱捕过程中产生大量原始数据，这些数据包括系统日志、指令交互、网络会话与主机行为事件，如何从中提取攻击者意图与行动链条，是智能推理系统的重要目标，构建基于大语言模型的语义解析引擎能够增强对非结构化数据的理解能力，采用改进的 LLM 框架加载语义编码器与上下文融合器，将原始数据转化为统一知识嵌入表示[16]。在实验环境中，以 GPT 类模型为核心训练推理引擎，输入长度限制为 4096 Token，模型参数规模为 35 亿，使用安全日志语料库进行指令意图微调训练，平均精度达 91.2%。引入因果图模型对攻击事件进行动态关系建模，利用贝叶斯网络结构 P (H|E) = P (E|H)·P (H) P (E)实现条件概率推理，将攻击行为节点与上下游行为建立因果路径图，识别多跳链式攻击与伪装行为。在资源参数配置上，推理引擎部署于高性能主机节点，使用双路 Intel Xeon Platinum 8358 处理器，内存 256 GB，IO 响应速度为 9.6 GB/s，系统可处理平均每小时 25 万条行为记录，峰值负载下保持 95%以上准确率不下降。推理结果以事件图谱形式回写数据库，供态势分析与响应策略引擎调用[17]。

4.3. 溯源与对抗建模

攻击溯源任务旨在识别攻击者身份、行为路径以及潜在攻击团伙特征，属于蜜罐系统的高级功能模块，通过多维数据融合与图谱推理技术，可重构攻击路径并辅助归属分析，结合 ATT&CK 框架对攻击战术进行标准化映射，将每一次攻击会话映射到对应的 TTP (Tactics, Techniques, Procedures)标签集合，构建行为特征向量 F = [t1, t2, ···, tn]，其中每个 ti 表示攻击使用的一种技术手段，使用余弦相似度对不同攻击路径进行对比，相似度超过 0.85 的路径被认为可能来自同一攻击团[18]。攻击图谱构建过程中，综合网络特征、主机行为、命令语义与指纹信息，融合 WAF 日志、DNS 查询、端口探测记录与通信频率等信息，构建溯源多维属性空间[19]。在实际测试中，从已知 APT 团伙攻击数据中提取 23 条攻击链与当前

蜜罐捕获路径进行比对，平均命中率为 78.4%，使用图嵌入分类器进行归属判定准确率达到 85.7%。对抗建模部分引入博弈策略分析工具，构建红蓝对抗的策略博弈模型，以威胁成功率与防御收益为博弈支付函数，求解攻击者最优行为路径与防守方最小成本配置，辅助系统在资源有限条件下完成最优防护策略部署。在部署层面，利用基于时间窗分布的动态 IP 标识机制与行为标记强化溯源追踪效率，实现对快速逃逸型攻击的有效识别与压制[20]。

5. 结论

人工智能技术与蜜罐系统的深度融合为网络空间主动防御提供了新路径，构建具备场景自生成、响应自决策、威胁智能分析与行为溯源能力的智能蜜罐体系，有助于提升对复杂攻击的应对效率与感知精度。融合大模型推理能力与 ATT&CK 知识框架，推动蜜罐系统从静态诱捕向动态博弈演进，在实战安全防护、攻击战术研究与溯源分析等方面展现出广阔应用前景。

参考文献

[1] Ding, W., Zhang, Z., Martínez, L., Huang, Y., Cao, Z., Liu, J., *et al*. (2025) New Trends of Adversarial Machine Learning for Data Fusion and Intelligent System. *Information Fusion*, **114**, Article 102683. https://doi.org/10.1016/j.inffus.2024.102683

[2] Song, W., Frakes, D. and Dasi, L.P. (2024) Active Machine Learning for Pre-Procedural Prediction of Time-Varying Boundary Condition after Fontan Procedure Using Generative Adversarial Networks. *Annals of Biomedical Engineering*, **53**, 217-229. https://doi.org/10.1007/s10439-024-03640-8

[3] Sayed, A., Alshathri, S. and Hemdan, E.E. (2024) Conditional Generative Adversarial Networks with Optimized Machine Learning for Fault Detection of Triplex Pump in Industrial Digital Twin. *Processes*, **12**, Article 2357. https://doi.org/10.3390/pr12112357

[4] Vadillo, J., Santana, R. and Lozano, J.A. (2024) Adversarial Attacks in Explainable Machine Learning: A Survey of Threats against Models and Humans. *WIREs Data Mining and Knowledge Discovery*, **15**, e1567. https://doi.org/10.1002/widm.1567

[5] Hong, J., Kim, H., Oh, S., Im, Y., Jeong, H., Kim, H., *et al*. (2024) Combating Phishing and Script-Based Attacks: A Novel Machine Learning Framework for Improved Client-Side Security. *The Journal of Supercomputing*, **81**, Article No. 69. https://doi.org/10.1007/s11227-024-06551-6

[6] Li, G., Shao, X., Wang, P., Ma, X., Li, H. and Ye, H. (2024) Anti-Machine-Learning-Attack Strong PUF Design Based on Multi-Path Delay Selection Strategy. *Microelectronics Journal*, **153**, Article 106434. https://doi.org/10.1016/j.mejo.2024.106434

[7] Kumar, P., Yadav, P. and Singh, V. (2024) Exploring Steel Fiber Integration in Dry Lean Concrete: Predictive Analysis of Compressive Strength and Performance via Machine Learning. *Asian Journal of Civil Engineering*, **26**, 263-271. https://doi.org/10.1007/s42107-024-01188-5

[8] Kotenko, I.V., Saenko, I.B., Lauta, O.S., Vasilev, N.A. and Sadovnikov, V.E. (2024) Approach to Detecting Attacks against Machine Learning Systems with a Generative Adversarial Network. *Pattern Recognition and Image Analysis*, **34**, 589-596. https://doi.org/10.1134/s1054661824700408

[9] 李华瑞, 李文博, 李铮, 等. 基于生成对抗网络与度量学习的数据驱动频率安全评估[J]. 电力系统保护与控制, 2024, 52(18): 101-111.

[10] 张涛. 基于对抗机器学习的工业控制网络欺骗攻击行为检测系统设计[J]. 计算机测量与控制, 2024, 32(10): 298-304.

[11] 张翼, 程小曼, 管冬平. 基于对抗机器学习的网络入侵特征选择研究[J]. 电子设计工程, 2024, 32(18): 173-176+181.

[12] 林巍, 廖丽娟. 基于连续扰动生成方法的可持续对抗训练(英文) [J]. 信息技术与电子工程前沿, 2024, 25(4): 527-540.

[13] 冯光升, 蒋舜鹏, 胡先浪, 等. 面向物联网的入侵检测技术研究新进展[J]. 信息网络安全, 2024, 24(2): 167-178.

[14] 潘宇恒, 廖思贤, 杨朝俊, 等. 面向网络入侵检测的对抗攻击系统[J]. 电脑知识与技术, 2024, 20(4): 100-102.

[15] Prathapani, A., Santhanam, L. and Agrawal, D.P. (2013) Detection of Blackhole Attack in a Wireless Mesh Network Using Intelligent Honeypot Agents. *The Journal of Supercomputing*, **64**, 777-804.

https://doi.org/10.1007/s11227-010-0547-3

[16] 杨文焕, 武辉林, 王云丽, 等. 一种基于蜜罐的智能防御系统设计与应用[J]. 科技风, 2024(24): 1-3.

[17] 卜钰. 浅析工业蜜罐技术在工业互联网场景下应用[J]. 自动化博览, 2023, 40(8): 36-39.

[18] 冀甜甜, 方滨兴, 崔翔, 等. CADetector: 跨家族的各项异性合约蜜罐检测[J]. 计算机学报, 2022, 45(4): 877-895.

[19] 孙利民, 潘志文, 吕世超, 等. 智能制造场景下工业互联网安全风险与对策[J]. 信息通信技术与政策, 2021, 47(8): 24-29.

[20] 游建舟, 吕世超, 孙玉砚, 等. 物联网蜜罐综述[J]. 信息安全学报, 2020, 5(4): 138-156.