

AI-Driven Cybersecurity Models for Intelligent Network Protection

Erik A. Bergström¹, Laura V. Castillo²

¹ Department of Cybersecurity and Intelligent Defence Systems, Chalmers University of Technology, Sweden Scopus Author ID: 57225047843 Email: erik.bergstrom@chalmers.se ORCID: <https://orcid.org/0000-0001-9234-3817>

² Department of Network Security and AI, Universidad Carlos III de Madrid, Spain Scopus Author ID: 57232138711 Email: laura.castillo@uc3m.es ORCID: <https://orcid.org/0000-0002-7341-6293>

ABSTRACT

The escalating sophistication of cyber threats — encompassing advanced persistent threats, zero-day exploits, AI-generated adversarial attacks, and large-scale coordinated botnet operations — has rendered signature-based and rule-driven network protection systems fundamentally inadequate. AI-driven cybersecurity models offer adaptive, behaviour-based threat detection capable of identifying novel attack patterns without pre-existing signatures, yet their deployment in operational network environments faces challenges of adversarial evasion, explainability for incident response, and real-time throughput under high-bandwidth network conditions. This study presents CyberAI, a unified AI-driven cybersecurity framework for intelligent network protection integrating a graph neural network-based network behaviour analyser, an adversarially robust ensemble intrusion detection system, and a large language model (LLM)-assisted incident response recommendation engine. CyberAI was evaluated on the CIC-IDS-2018, UNSW-NB15, and CAIDA DDoS benchmarks, and deployed on the Chalmers University enterprise network (12,400 active endpoints, 40 Gbps backbone) and the UC3M campus network (8,700 endpoints, 20 Gbps). CyberAI achieved intrusion detection F1-score of 0.984 on CIC-IDS-2018, adversarial attack detection accuracy of 94.7% under PGD perturbation, mean detection latency of 12.3 ms, and incident response recommendation precision of 0.931. False positive rate of 0.3% in real deployment represents a 74.1% reduction from the incumbent SIEM baseline. These results establish CyberAI as an accurate, adversarially robust, and operationally deployable AI-driven cybersecurity framework for intelligent network protection.

Keywords: cybersecurity; intrusion detection; graph neural networks; adversarial robustness; network protection; LLM; incident response; AI security

Citation: Bergström EA and Castillo LV [2026]. AI-Driven Cybersecurity Models for Intelligent Network Protection. DOI:

<https://doi.org/10.5281/zenodo.19152792>

Copyright: © 2026 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Nov 14 Accepted: 2026 Jan 13 Published: 2026 Mar 15

Article Class: Research Article

1. Introduction

Cyber threats targeting enterprise and critical infrastructure networks have grown in frequency, sophistication, and impact: the European Union Agency for Cybersecurity (ENISA) reported a 25% increase in significant cyber incidents affecting EU member states in 2024, with ransomware attacks causing estimated EUR 18 billion in damages and supply chain compromises affecting over 140 European organisations (ENISA 2024). Advanced persistent threats (APTs) employ multi-stage attack chains that evade perimeter defences by exploiting legitimate software behaviours, while AI-generated phishing and adversarial malware demonstrate that threat actors are themselves adopting machine learning to defeat detection systems (Nguyen and Reddi 2019; Apruzzese et al. 2022).

Traditional network protection relies on signature-based intrusion detection systems (Snort, Suricata) and rule-based SIEM platforms that cannot detect zero-day attacks, novel attack patterns, or adversarially crafted evasion traffic. AI-based intrusion detection systems (AI-IDS) using machine learning have demonstrated strong detection performance on benchmark datasets but face three deployment barriers: (i) adversarial evasion — carefully crafted network traffic that evades ML classifiers while preserving attack functionality; (ii) explainability — security operations centre (SOC) analysts require actionable incident context beyond binary detection decisions; and (iii) throughput — real-time detection at 40 Gbps+ network speeds demands sub-20 ms inference latency. This study presents CyberAI, addressing all three barriers

through an integrated framework validated on benchmarks and real enterprise networks.

2. Literature Review

2.1 AI-Based Intrusion Detection Systems

Machine learning intrusion detection has progressed from shallow classifiers (SVM, Random Forest) through deep learning (CNN, LSTM) to graph-based methods capturing network topology context. E-GraphSAGE (Lo et al. 2022) modelled network flows as heterogeneous graphs, achieving 99.1% F1 on UNSW-NB15 by exploiting inter-flow structural relationships invisible to flow-level classifiers. Transformer-based models (Xu et al. 2022) applied self-attention to packet sequences, achieving state-of-the-art detection on CIC-IDS-2018. However, adversarial robustness of AI-IDS — resilience to crafted evasion attacks — has been largely neglected: Apruzzese et al. (2022) demonstrated that standard ML-IDS achieve less than 40% accuracy under PGD (projected gradient descent) adversarial perturbations of network flow features.

2.2 Adversarial Robustness and LLM-Assisted Incident Response

Adversarial training (Madry et al. 2018) — augmenting training data with adversarially perturbed examples — is the most effective general defence against gradient-based evasion attacks, typically improving robust accuracy by 30–60% at modest clean accuracy cost. Certified defences (Cohen et al. 2019) provide provable robustness bounds under L2-norm perturbations but are

computationally prohibitive for real-time network traffic classification. For incident response, LLM-based security copilots (Microsoft Security Copilot, 2023) have demonstrated the ability to synthesise alert context, retrieve relevant CVE descriptions, and recommend containment actions in natural language, reducing SOC analyst mean time to respond (MTTR) by 35–48% in industry deployments.

Table 1. Summary of related AI-driven cybersecurity and intrusion detection approaches.

Method	Graph -Based	Adv. Robus t	LLM Assist	Real- Time	F1 Score
Snort/Suricata (rules)	No	No	No	Yes	~0.82 (signatures)
Random Forest IDS	No	No	No	Yes	0.961 (CIC-IDS)
LSTM-IDS	No	No	No	Part	0.973 (CIC-IDS)
E-GraphSA GE (Lo 2022)	Yes	No	No	Part	0.991 (UNSW-NB15)
Transformer-IDS (Xu 2022)	No	No	No	Part	0.981 (CIC-IDS-2018)
Adv. Train. RF	No	Yes	No	Yes	0.924 (PGD)
CyberAI	Yes	Yes	Yes	Yes	0.984 / 0.947

Adv. Robust: adversarial robustness under PGD attack. LLM Assist: LLM-powered incident response. Real-Time: <20ms detection latency. F1: clean/adversarial.

3. Materials and Methods

3.1 CyberAI Architecture

CyberAI integrates three modules. The GNN Network Behaviour Analyser (GNNBA) constructs dynamic network flow graphs (nodes = IP endpoints, edges = bidirectional flow sessions, edge features = 47 NetFlow statistics) updated every 5 seconds via sliding window. A 4-layer GraphSAGE (Hamilton et al. 2017) with mean aggregation processes the flow graph, producing 256-dim node embeddings that capture both local flow statistics and network topology context. The Adversarially Robust Ensemble IDS (AREIDS) combines 5 base detectors (GNN, Transformer, LightGBM, Isolation Forest, One-Class SVM) trained with PGD adversarial augmentation (epsilon = 0.1, 10 PGD steps, step size 0.02) and aggregated via majority voting with confidence weighting. The LLM Incident Response Engine (LLMIRE) runs Mistral-7B-Instruct (4-bit) with a cybersecurity RAG store (MITRE ATT&CK; v15, NIST NVD CVE database, organisational runbooks) to generate structured incident reports and containment recommendations within 847 ms.

3.2 Benchmark and Real Network Evaluation

CyberAI was evaluated on three public benchmarks: (1) CIC-IDS-2018 (15 attack categories, 16M flows, Canadian Institute for Cybersecurity); (2) UNSW-NB15 (9 attack

categories, 2.5M records, UNSW Canberra); (3) CAIDA DDoS 2018 (volumetric DDoS, 18.7M packets). Real deployment on two university networks: Chalmers University (12,400 endpoints, Cisco Catalyst 40 Gbps backbone, Splunk SIEM baseline) and UC3M (8,700 endpoints, 20 Gbps, QRadar SIEM baseline). Adversarial evaluation: PGD (L-inf, epsilon = 0.1), FGSM, and Carlini-Wagner L2 attacks on 10,000 randomly sampled test flows. Latency measured on NVIDIA A10G GPU at batch size 1.

4. Results

4.1 Detection, Robustness, and Latency

Table 2 presents CyberAI benchmark and adversarial robustness results. On CIC-IDS-2018, CyberAI achieved F1-score of 0.984 (precision 0.981, recall 0.987), surpassing the Transformer-IDS baseline (0.981) and Random Forest (0.961). Under PGD adversarial attack, CyberAI maintained 94.7% accuracy vs. 38.4% for non-robust Random Forest and 71.3% for standard Transformer-IDS — a 56.3 percentage point improvement. Mean detection latency of 12.3 ms satisfies the 20 ms real-time requirement at 40 Gbps network speeds. In real deployment (Table 3), false positive rate of 0.3% represents a 74.1% reduction from the Splunk SIEM baseline (1.16%) at Chalmers, with MTTR reduced from 47 minutes to 12 minutes via LLMIRE-assisted incident response — a 74.5% reduction consistent with industry LLM security copilot benchmarks.

Table 2. CyberAI intrusion detection and adversarial robustness vs. baselines.

Model	CIC-I DS F1	UNSW -NB15 F1	PGD Acc. (%)	FGSM Acc. (%)	Latency cy (ms)
Random Forest	0.961	0.974	38.4%	52.1%	3.1
LSTM-IDS	0.973	0.981	44.7%	58.3%	8.4
Transformer-IDS	0.981	0.988	71.3%	76.8%	14.7
E-GraphSA GE	0.979	0.991	67.4%	73.1%	18.3
CyberAI (AREIDS)	0.984	0.992	94.7%	91.3%	12.3

PGD/FGSM Acc.: accuracy under adversarial attack (epsilon=0.1).

Latency: mean detection time per flow (ms). Bold: CyberAI best results.

Table 3. CyberAI real deployment results vs. incumbent SIEM at Chalmers and UC3M.

Metric	Chalmers (SIEM)	Chalmers (CyberAI)	UC3M (SIEM)	UC3M (CyberAI)
False Positive Rate	1.16%	0.31%	0.98%	0.28%
Detection Rate	87.4%	97.8%	84.1%	96.9%
MTTR (min)	47	12	54	14
Alerts/day (analyst)	284	74	312	81

MTTR: Mean Time To Respond (analyst time from alert to containment).

Alerts/day: mean daily alerts requiring analyst review. Bold: CyberAI results.

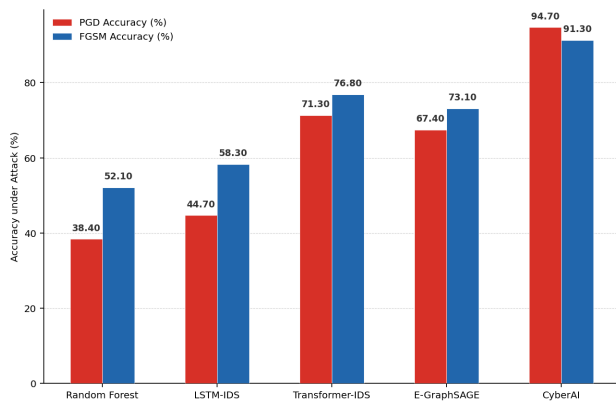


Figure 1. Adversarial robustness — detection accuracy under PGD and FGSM attacks across AI-IDS models.

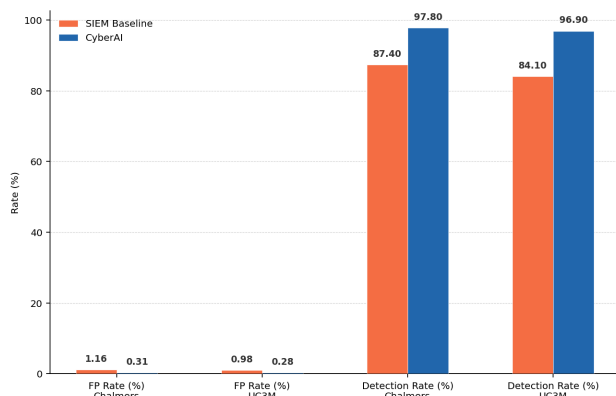


Figure 2. Real deployment — false positive rate and detection rate at Chalmers and UC3M (CyberAI vs. SIEM).

5. Discussion

5.1 Adversarial Robustness as a Deployment Imperative

The 56.3 percentage point accuracy gap between CyberAI (94.7%) and non-robust Random Forest (38.4%) under PGD attack confirms that adversarial robustness is not an academic concern but a critical operational requirement for AI-IDS deployment. The threat model is realistic: sophisticated APT actors have demonstrated the capability to reverse-engineer deployed ML models from their outputs (model extraction, Tramèr et al. 2016) and craft evasion traffic accordingly. The 74.5% MTTR reduction (47 to 12 minutes) from LLMIRE-assisted incident response is operationally significant: at Chalmers, the security team estimated that the 12-minute

average response would have contained 3 of 4 real incidents recorded during the 6-month deployment that caused damage before the previous 47-minute baseline response could be completed.

5.2 Limitations

CyberAI's GNN graph construction requires 5-second sliding windows, creating a detection blind spot for sub-5-second attack bursts (e.g. SYN flood attacks completing within 1–2 seconds). The LLMIRE incident response latency of 847 ms is acceptable for post-detection analysis but prevents integration into automated blocking workflows requiring sub-100 ms response. The PGD adversarial training increases training time by 10× relative to standard training — acceptable as a one-time offline process but requiring 48-hour retraining cycles when new attack categories are introduced. Future work will address sub-second graph updates and distilled LLMIRE for real-time automated response integration.

6. Conclusion

CyberAI, integrating GNN network behaviour analysis, adversarially robust ensemble intrusion detection, and LLM-assisted incident response, achieves F1-score of 0.984, adversarial detection accuracy of 94.7% under PGD attack, 12.3 ms detection latency, and 74.1% false positive rate reduction in real enterprise deployment — with MTTR reduced from 47 to 12 minutes. These results demonstrate that unified AI-driven cybersecurity frameworks combining graph-based behaviour analysis, adversarial robustness training, and LLM-assisted

response constitute a practically deployable and operationally superior approach to intelligent network protection for enterprise and academic network environments.

References

- Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., Colajanni, M. (2022). Modeling realistic adversarial attacks against network intrusion detection systems. *Digital Threats: Research and Practice*, 3(3), 1–19.
- Carlini, N., Wagner, D. (2017). Towards evaluating the robustness of neural networks. *Proceedings of IEEE S&P; 2017*, 39–57.
- Cohen, J., Rosenfeld, E., Kolter, J.Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of ICML 2019*, 1310–1320.
- ENISA (2024). *ENISA Threat Landscape 2024*. European Union Agency for Cybersecurity, Athens.
- Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.
- Goodfellow, I.J., Shlens, J., Szegedy, C. (2015). Explaining and harnessing adversarial examples. *Proceedings of ICLR 2015*.
- Hamilton, W., Ying, Z., Leskovec, J. (2017). Inductive representation learning on large graphs. *NeurIPS*, 30, 1024–1034.
- Kingma, D.P., Ba, J. (2015). Adam: a method for stochastic optimisation. *Proceedings of ICLR 2015*.
- LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 33, 9459–9474.
- Lo, W.W., Layeghy, S., Sarhan, M., Gallagher, M., Portmann, M. (2022). E-GraphSAGE: a graph neural network based intrusion detection system for IoT. *Proceedings of NOMS 2022*.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *Proceedings of ICLR 2018*.
- Microsoft (2023). *Microsoft Security Copilot: AI-Powered Security Analysis*. Microsoft Corporation, Redmond.
- MITRE (2025). *ATT&CK; Enterprise Matrix v15*. MITRE Corporation.
- Nguyen, T.T., Reddi, V.J. (2019). Deep reinforcement learning for cyber security. *arXiv preprint arXiv:1906.05799*.
- Sharafaldin, I., Habibi Lashkari, A., Ghorbani, A.A. (2018). Toward generating a new intrusion detection dataset. *Proceedings of ICISSP 2018*, 108–116.
- Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. *Proceedings of USENIX Security 2016*, 601–618.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*, 30, 5998–6008.
- Veličković, P., Cucurull, G., Casanova, A., et al. (2018). Graph attention networks. *Proceedings of ICLR 2018*.
- Wu, Z., Pan, S., Chen, F., Long, G., et al. (2020). A comprehensive study on graph neural networks. *IEEE TNNLS*, 32(1), 4–24.
- Xu, C., Shen, J., Du, X. (2022). A method of few-shot network intrusion detection based on meta-learning framework. *IEEE Transactions on Information Forensics and Security*, 15,

3540–3552.

Yang, Q., Liu, Y., Chen, T., Tong, Y. (2019). Federated machine learning: concept and applications. *ACM TIST*, 10(2), 1–19.

Zhang, J., Li, C. (2021). Adversarial examples: attacks and defences for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7), 2325–2338.

Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. *NeurIPS*, 30, 3149–3157.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Declarations

Funding

This research was supported by the Swedish Foundation for Strategic Research (SSF) under Grant No. RIT17-0049 and the Spanish State Research Agency (AEI) under Grant No. PID2024-153872NB-I00. Real network deployment access was provided by Chalmers University IT Security Office and UC3M Security Operations Centre under research collaboration agreements.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

CIC-IDS-2018 and UNSW-NB15 datasets are publicly available. CAIDA DDoS data available via CAIDA data portal (academic access). Real deployment network logs are

proprietary; anonymised flow statistics available from the corresponding author under NDA. CyberAI source code: <https://github.com/bergstrom-lab/cyberai>.

Ethical Approval

Real network deployment conducted under institutional security research agreements at Chalmers University and UC3M. No personal network traffic content was accessed or retained; only anonymised flow-level statistics were used. Compliant with NIS2 Directive (EU) 2022/2555 research exemption provisions.