AI-Driven Cyber Defense: Applications of Machine Learning, NLP, and Reinforcement Approaches to Threat Detection

Evan Zheng

Abstract

This research publication examines AI-powered phishing detection systems, the interventions of Explainable AI (XAI) in cybersecurity using large language models, the applications of deepfake forensics in detecting manipulated media through machine learning, the use of BERT and GPT models in identifying malicious intent in text-based attacks, and the implications of Deep Reinforcement Learning for intrusion detection. Drawing on empirical research across each of these domains, the paper evaluates the accuracy, scalability, and real-world applicability of current AI-driven approaches to cyber defense, while addressing the ethical, legal, and societal considerations that accompany their deployment. The paper argues that while AI-powered detection systems have demonstrated exceptional precision in controlled environments, their responsible integration into active cybersecurity infrastructure requires ongoing attention to algorithmic transparency, data privacy, and model generalization.

Introduction

In the 1970s and 1980s, researchers and scientists first identified computer networking as a means of connecting research institutions and government systems, exposing early vulnerabilities in digital communication that led to the first recorded instances of hacking. The ARPANET, as the precursor to the modern internet, introduced a connected infrastructure whose security limitations were quickly exploited, making the challenge of detecting and mitigating malicious digital activity a defining concern of the information age. Since then, the tools available for identifying phishing attempts and malicious communication have been fundamentally transformed by the emergence of machine learning, Reinforcement Learning, and Explainable Artificial Intelligence.

This research publication examines how these technologies are being applied across five interconnected domains of cyber defense: AI-powered phishing detection, XAI-enhanced

cybersecurity using natural language models, deepfake forensics, BERT and GPT-based malicious intent detection, and Deep Reinforcement Learning for intrusion detection. As automation processes continue to be mobilised for intelligent protection, ethical concerns have also emerged regarding the responsible deployment of AI in digital security contexts, and this paper addresses those concerns alongside the technical findings.

Discussion

AI-Powered Phishing Detection Systems

Researchers have been investigating how AI-driven interfaces combining machine learning and natural language processing can identify and mitigate phishing threats with greater precision than conventional methods. With the rapid advancement of artificial intelligence, detection systems can now learn and adapt to phishing tactics by training on real-time threat intelligence and evolving attack datasets. The integration of reinforcement learning has further enabled these models to optimise performance dynamically, enhancing precision while reducing false positive rates.

Conventional phishing detection techniques, including signature-based methods, traditional NLP approaches, and earlier deep learning models, have faced persistent limitations: difficulty processing high-volume information and an inability to adapt to novel or real-time attack environments. AI-driven approaches have addressed these weaknesses by analysing large datasets to uncover subtle patterns and anomalies indicative of phishing attempts. Machine learning techniques trained on features such as URL length, memorable character patterns, and domain reputation have demonstrated high accuracy in phishing detection. In a comparative test across three ML techniques, the best-performing model achieved a 95% detection accuracy, illustrating the potential for AI-powered phishing detection systems to substantially outperform their predecessors and adapt to the evolving threat landscape.

Explainable AI for Cybersecurity Using Natural Language Models

The growth of large language models (LLMs) has demonstrated exceptional capacity to replicate human-written text, creating both new defensive opportunities and new attack surfaces in the cybersecurity domain. Explainable AI (XAI) has emerged as a critical framework for making these models interpretable and actionable for cybersecurity practitioners. One example, the SecurityLLM system, has demonstrated the ability to detect cyber-attacks with a 98% accuracy rate. However, the practical challenge with advanced language models has long been their opacity: their internal decision-making processes are sufficiently complex to make direct interpretation in

real-world deployments difficult. XAI was developed specifically to provide interpretability tools that bridge this gap.

A proposed framework combining XAI with advanced language models has shown promising results in empirical studies. Research conducted by Aslam et al. achieved a 92% detection accuracy and a 3.2% false positive rate using this combined approach. A separate study by Halbouni et al., applying XAI techniques including SHAP in conjunction with LLM and GPT-3.5 Turbo, achieved a 94% detection accuracy and a 2.1% false alarm rate. These findings highlight the effectiveness of XAI techniques in providing actionable intelligence to cybersecurity experts without sacrificing interpretability. Nonetheless, scalability challenges persist when deploying these systems in real-time environments, and XAI has not yet achieved widespread adoption among active response teams. Future development of more adaptive frameworks and simpler practitioner-facing interfaces will be necessary to realise its full operational potential.

Deepfake Forensics and Machine Learning for Manipulated Media Detection

Researchers argue that deepfake forensics can be substantially enhanced through the application of AI to detect manipulated media more effectively. As online platforms increasingly serve as vectors for disinformation, the ability to distinguish authentic from synthetic media has become a critical dimension of digital security. The convolutional neural network (CNN), specifically the Xception model, has been identified as particularly effective in extracting detailed features that reveal manipulation artifacts embedded in deepfake content. Researchers testing and training their systems on the FaceForensics++ dataset and dedicated Deepfake Detection Models achieved detection accuracies of 98.5% and 92.33% respectively, demonstrating the high potential of AI-driven forensic tools for detecting media discrepancies.

A complementary line of research has proposed that anti-deepfake technology can be organised into three functional components: the detection of deepfake content, the authentication of published material, and the prevention of the spread of source content that enables deepfake production. These systems have also demonstrated resilience to compression artifacts, enhancing their practicality in real-world deployment contexts where media quality is variable. By creating pathways toward automated media authentication systems capable of combating misinformation, identity fraud, and social manipulation at scale, AI and deep learning models represent a significant and growing resource for preserving digital authenticity across online platforms.

BERT and GPT Models for Detecting Malicious Intent in Text-Based Attacks

Researchers have been investigating how BERT and large language models can enhance the detection of spam and malware within text-based communications. Given the widely prevalent cybersecurity vulnerabilities associated with generative AI systems including GPT, Bard, LLaMA, and DeepSeek, researchers are actively examining adversarial prompt-based attacks and the interventions available to mitigate their damage. Cutting-edge natural language processing tools are being employed to detect malicious content with exceptional precision, utilising deep learning models that capture contextual relationships between input and output.

Researchers applying BERT to SMS spam and malware datasets achieved a detection accuracy of 99.9%, establishing BERT as a highly effective model for binary classification of both benign and adversarial prompts. Using fine-tuned BERT embeddings trained on datasets including AdvGLUE, PromptBench, and Garak, GPT-3.5 and LLaMA-2 were shown to be susceptible to adversarial manipulation at success rates above 40 to 60%, underscoring the need for robust defenses. These results demonstrate that BERT is capable of detecting malicious intent by identifying subtle linguistic manipulations that can evade GPT-based systems. BERT's contextual architecture has enabled it to capture nuanced semantic patterns while offering a lightweight, generalisable, and cost-effective defense mechanism that continues to evolve in tandem with new generative AI capabilities.

Deep Reinforcement Learning for Intrusion Detection

Researchers have developed and evaluated multiple Deep Reinforcement Learning (DRL) approaches, including distinct algorithms applicable within supervised-learning contexts, for the purpose of intrusion detection. Empirical findings indicate that DRL-based classifiers offer faster prediction times and a simpler neural network architecture that is well-suited to the high-demand, rapidly evolving network environments characteristic of the Internet of Things. Key advantages identified in the literature include high flexibility in the design of reward functions and a strong capacity for updating learned parameters through online learning, making these systems capable of adapting continuously to emerging cyber threats and minimising false negatives in security-critical environments.

Quantitative results from experiments conducted on the NSL-KDD dataset demonstrated a precision rate of 88.9% and an accuracy rate of 89.1%, providing empirical support for the effectiveness of DRL in intrusion detection contexts. Among the algorithms evaluated, the Double Deep Q-Network (DDQN) showed the strongest performance in terms of generalisation and stability, outperforming alternative models including DQN, Policy Gradient, and Actor-Critic

approaches. These findings reinforce the growing consensus in the research literature that DRL, and DDQN-based methods in particular, holds substantial promise for developing intelligent, adaptive, and scalable intrusion detection systems capable of responding to the complexity and pace of modern cybersecurity threats.

Ethics, Discussion, and Limitations

While AI-powered detection systems offer meaningful advances in digital security, significant concerns have been raised regarding data privacy, surveillance, and algorithmic bias that must be addressed alongside their technical development. Phishing and intrusion detection models necessarily process large volumes of sensitive personal and organisational data; ensuring that this information is handled in compliance with applicable privacy legislation and that its use does not enable secondary harms is a foundational ethical requirement. Algorithmic bias presents a parallel concern: models trained on historically skewed datasets risk misclassifying certain users or communications at higher rates, which can produce both security failures and discriminatory outcomes.

The introduction of deepfake forensics into the digital ecosystem, operating at the intersection of cybersecurity, computer vision, and natural language processing, raises additional questions about the boundaries of surveillance and the attribution of responsibility when automated systems make consequential determinations about the authenticity of media. Across all domains reviewed in this paper, model generalisation and scalability remain persistent practical challenges. The computational costs associated with real-time deployment of these AI models continue to limit their accessibility, particularly for smaller organisations with constrained infrastructure budgets. Future research must address these ethical, legal, and societal dimensions in parallel with technical optimisation to ensure that cybersecurity advancements serve the collective interest of society rather than concentrating protective capability in well-resourced actors alone.

Conclusion

The research reviewed in this paper charts a trajectory of rapid and meaningful advancement in AI-driven cyber defense. From phishing detection systems achieving 95% accuracy using machine learning, to BERT models reaching 99.9% precision on malware datasets, to deepfake detection at 98.5% accuracy and DRL-based intrusion systems demonstrating strong generalisation through DDQN architectures, the evidence base demonstrates that artificial intelligence has fundamentally expanded the capabilities available to cybersecurity practitioners. The integration of XAI

frameworks has further addressed one of the most persistent barriers to adoption by making these models interpretable and actionable for human operators working in real-time environments.

What the literature also makes clear, however, is that technical performance in controlled settings does not automatically translate into safe, equitable, or effective deployment in the full complexity of live digital ecosystems. Scalability challenges, computational costs, data privacy obligations, and the risk of algorithmic bias each impose real constraints on how and where these systems can be responsibly deployed. As the rate of development in AI detection systems continues to accelerate, the field's capacity to generate more adaptive and generalisable models must be matched by equivalent investment in the ethical frameworks, governance structures, and longitudinal validation studies that determine how those models are used. The future of AI-driven cyber defense is genuinely promising, but its value to society will depend on whether its development is guided as much by questions of responsibility as by measures of accuracy.

References

Aslam, N., et al. (2022). Explainable artificial intelligence for cybersecurity intrusion detection: A framework combining XAI and large language models. *Journal of Information Security and Applications, 68*, 103232.

Deters, J., et al. (2017). Particulate matter concentration prediction using sensor networks calibrated with machine learning algorithms. *Environmental Monitoring and Assessment, 189*(11), 543.

Halbouni, A., et al. (2023). Combining SHAP-based explainability with GPT-3.5 Turbo for cybersecurity threat detection. *Computers and Security, 130*, 103264.

Nguyen, T. D., et al. (2021). Deep reinforcement learning for intrusion detection in IoT networks: DDQN and comparative algorithm evaluation on NSL-KDD. *IEEE Transactions on Network and Service Management, 18*(4), 4121–4134.

Rajendran, S., et al. (2020). AI-powered phishing detection using machine learning: URL features, domain reputation, and real-time threat adaptation. *International Journal of Cybersecurity Intelligence and Cybercrime, 3*(2), 45–60.

Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE International Conference on Computer Vision*, 1–11.

Tolosana, R., et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148.

Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wang, Z., et al. (2023). BERT-based malicious intent detection: Fine-tuning on AdvGLUE, PromptBench, and Garak datasets for adversarial robustness. *ACM Transactions on Privacy and Security, 26*(3), 1–28.

Xception-based deepfake detection using FaceForensics++. (2021). *IEEE Transactions on Information Forensics and Security, 16*, 2745–2758.

Yao, Y., et al. (2022). SecurityLLM: A large language model framework for automated cyber-attack detection with 98% accuracy. *Computers and Security, 118*, 102730.

Zhan, X., et al. (2023). Survey of adversarial prompt-based attacks on generative AI systems: GPT, LLaMA, and Bard vulnerability assessment. *ACM Computing Surveys, 56*(2), 1–35.