Minitrack Introduction: Cybersecurity in the Age of Artificial Intelligence, AI for Cybersecurity, and Cybersecurity for AI

Mark Patton
University of Arizona
mpatton@email.arizona.edu

Sagar Samtani
Indiana University
ssamtani@iu.edu

Hongyi Zhu
University of Texas at San Antonio
hongyi.zhu@utsa.edu

Hsinchun Chen
University of Arizona
hchen@eller.arizona.edu

Abstract

Artificial Intelligence (AI) and cybersecurity intersect in ways that offer significant opportunities and lead to considerable threats. Recognizing the critical importance of this convergence, major entities like the National Science Foundation (NSF), the National Science Technology Council (NSTC), and the National Academies of Sciences (NAS) have prioritized enhancing AI capabilities for cybersecurity. This minitrack focuses on AI and cybersecurity within broad domains, including collaborative inter-organizational environments and shared technologies. The five papers in this minitrack can potentially contribute to advancing the development of AI for cybersecurity, addressing critical cybersecurity challenges for AI, and proposing innovative solutions.

Keywords: Cybersecurity, Artificial Intelligence, Machine Learning, Deep Learning, Pretrained Language Models, Collaboration, Data Analytics, Cybersecurity for Artificial Intelligence.

1. Introduction

The intersection of Artificial Intelligence (AI) and cybersecurity offers significant promise and poses considerable threats. This convergence spans multiple domains, from internal organizational solutions identifying network threats to broad, collaborative efforts addressing threats across platforms. Major entities like the National Science Foundation (NSF), the National Science Technology Council (NSTC), and the National Academies of Sciences (NAS) have recognized enhancing AI capabilities for cybersecurity as a key national priority. While AI implementation in cybersecurity can be both internal and collaborative, the full potential of using AI (e.g., machine learning (ML), deep learning (DL), and Pretrained Language Models (PTM) for cybersecurity and improving the security of AI systems remains relatively understudied yet critically important. Exploring these areas is essential for addressing emerging threats and advancing national security.

2. Minitrack Goals and Focus

This minitrack centers on AI and Cybersecurity within broad domains such as collaborative inter-organizational environments, shared domains, and collaborative technologies. The threats being addressed with and/or to AI have significant societal impacts. Topics and research areas include, but are not limited to:

- Novel applications of AI (e.g., ML/DL/PTM) in Cybersecurity for multi-user/multi-organizational collaborative domains and/or systems.
- Adversarial AI/ML Applications in Cybersecurity that collaboratively span organizations or apply to collaborative systems (i.e., malware, phishing, or any applicable threat/identification domain).
- Protecting AI that is used collaboratively (i.e., shared data sets, models, applications) or spans collaborative domains from cybersecurity threats (e.g., adversarial examples, model inversion).
- Using AI to protect AI in any appropriate wide-reaching setting.
- Novel collaboration approaches to leveraging and protecting AI in the cybersecurity domain.
- Disseminating tools, techniques, and applications of AI in Cybersecurity and Cybersecurity for AI.

3. Papers

After a rigorous peer review process with careful management of conflicts of interest, we accepted five papers for this minitrack. Below, we present a brief introduction to each accepted paper. We encourage interested readers to access the full papers in the proceedings.

Paper 1 Title: Towards Attribution in Network Attacks: A Deep Learning-Based Robust Framework for Intrusion Detection and Adversarial Toolchain Identification.

URI: https://hdl.handle.net/10125/108881
978-0-9981331-8-8

This study proposes a novel defense framework for network intrusion detection systems (NIDS) to address challenges like susceptibility to evasion attacks and high false positives/negatives in traditional ML models. By integrating supervised and unsupervised learning, the framework enhances NIDS capabilities to accurately identify known and novel attacks and detect adversarial attacks and their toolchains. The framework includes a toolchain detection component that attributes attacks, understands adversary motivations, and improves threat intelligence and incident response in cybersecurity operations centers (CSOCs).

Paper 2 Title: BERT-Cuckoo15: A Comprehensive Framework for Malware Detection Using 15 Dynamic Feature Types.

Malware detection is challenged by the need to select features from diverse data sources, affecting model accuracy. Existing methods often rely on single feature types or extensive feature engineering, which may fail to capture complex feature relationships and often assume feature independence—a false premise for sophisticated malware behavior. To overcome these limitations, this study introduces BERT-Cuckoo15, a malware detection model using Bidirectional Encoder Representations from Transformers (BERT) to analyze relationships among diverse features from dynamic analysis in the Cuckoo sandbox. Evaluated on 36,770 samples across nine malware types, BERT-Cuckoo15 achieves 97.61% accuracy, showing its effectiveness in capturing complex feature interdependencies and improving detection accuracy.

Paper 3 Title: Collecting, Linking, and Assessing Machine Learning Open-Source Software: A Large Scale Collection and Vulnerability Assessment Pipeline.

Machine Learning Open-Source Software (MLOSS) platforms such as GitHub and Hugging Face have significantly contributed to the recent rapid advancement of AI by enabling developers to share and collaborate on AI projects. However, these platforms also host assets containing vulnerabilities that can impact high-stake applications utilizing them (in finance or healthcare industries). This paper developed an MLOSS Collection Pipeline to map the MLOSS landscape and understand these vulnerabilities. This pipeline collected 373,634 models from Hugging Face and 39,115 repositories from GitHub, identifying 6,751,739 vulnerabilities. The results offer promising directions for future research, including vulnerability linking analysis and cross-platform vulnerability propagation identification.

Paper 4 Title: Depressive Behavior Detection Using Sensor Signal Data: An Attention-based Privacy-Preserving Approach.

Using personally identifiable information (PII) introduces notable privacy concerns in sensor signal-based depression detection. This study introduces a novel attention-based, privacy-preserving model that mitigates these concerns by weighting non-PII sensors more heavily and high-risk sensors less, using differential privacy (DP) principles. The model achieves high performance—recall, precision, and F1 score of 0.889, and an AUC of 0.9—without relying on PII-releasing sensors, demonstrating effective depression detection while preserving privacy. This approach has significant implications for unobtrusive mental healthcare and secure deployment in digital health applications.

Paper 5 Title: Improving the Adversarial Robustness of Machine Learning-based Phishing Website Detectors: An Autoencoder-based Auxiliary Approach.

Anti-phishing research involves defensive efforts to develop ML-based phishing detectors and offensive efforts to understand adversarial threats. Adversaries manipulate phishing websites into adversarial examples, misleading detectors into misclassification. Enhancing adversarial robustness often compromises performance on clean data. This study proposes a novel approach using a Graph Convolutional Autoencoder as an auxiliary model to collaborate with the original detector, distinguishing evasive phishing websites from legitimate ones. The approach achieves high adversarial robustness and maintains higher performance on clean data than retrained and fine-tuned benchmarks.

4. Acknowledgements

We express our gratitude to all the authors for their contributions and acknowledge the reviewers for their diligent work on the submissions to this minitrack. This minitrack is based upon work funded by the U.S. Military Academy (USMA) under Cooperative Agreement No. W911NF-22-2-0045, the U.S. Army Combat Capabilities Development Command C5ISR Center under Support Agreement No. USMA21056, DGE-2038483 (SaTC-EDU), DGE-1946537 (SFS), OAC-1917117 (CICI), OAC-2319325 (CICI), and CNS-2338479 (CAREER).