Authentic Learning Approach for Artificial Intelligence Systems Security and Privacy

1st Mst Shapna Akter
Dept. of Computer Science
Kennesaw State University
Kennesaw, USA
Email: makter2@students.kennesaw.edu

2nd Hossain Shahriar
Dept. of Information Technology
Kennesaw State University
Kennesaw, USA
Email: hshahria@kennesaw.edu

3rd Dan Lo
Dept. Computer Science
Kennesaw State University
Kennesaw State University, USA
Email: dlo2@kennesaw.edu

4th Nazmus Sakib
Dept. of Computer Science
Kennesaw State University
Kennesaw, USA
Email: nsakib1@kennesaw.edu

5th Kai Qian
Dept. of Computer Science
Kennesaw State University
Kennesaw, USA
Email: kqian@kennesaw.edu

6th Michael Whitman
Dept. of Information Systems and Security
Kennesaw State University
Kennesaw, USA
Email: mwhitman@kennesaw.edu

7th Fan Wu
Dept of Computer Science
Tuskegee University
Tuskegee, USA
Email: fwu@tuskegee.edu

***Abstract*—The main objective of authentic learning is to offer students an exciting and stimulating educational setting that provides practical experiences in tackling real-world security issues. Each educational theme is composed of pre-lab, lab, and post-lab activities. Through the application of authentic learning, we create and produce portable lab equipment for AI Security and Privacy on Google CoLab. This enables students to access and practice these hands-on labs conveniently and without the need for time-consuming installations and configurations. As a result, students can concentrate more on learning concepts and gain more experience in hands-on problem-solving abilities.**

***Index Terms*—Authentic learning, ML/DL algorithm, Adversarial attack, Security, Privacy, Education.**

I. Introduction

Authentic learning is a hands-on approach to education that aims to provide students with the skills and knowledge they need to tackle real-world problems. In the context of cybersecurity, authentic learning can help students develop the skills they need to combat the increasing risk of adversarial attacks on machine learning systems. To achieve this, authentic learning typically involves a series of pre-lab, lab, and post-lab activities, where students learn key concepts, practice problem-solving, and reflect on their solutions. As machine learning becomes more widespread, the risk of adversarial attacks and other security threats also increases. Adversarial attacks can bypass conventional cybersecurity defenses and cause significant damage, such as stealing sensitive data or injecting malicious code. Apart from adversarial attacks, there are several security threats to AI systems [1], such as AI Trojan [2], model inversion [3], and other types of cyberattacks. To combat these threats effectively, there is a need for cybersecurity curricula to incorporate authentic learning of attacks and defense on machine learning systems. However, there is currently a shortage of teaching and learning materials, open-source portable hands-on labware, and dedicated staff and faculty in this field. To address these challenges, we propose an open-sourced, portable, and modular approach to enhance AI Security and Privacy. This approach involves developing online, portable hands-on labware consisting of multiple modules covering various topics such as Getting Started , Adversarial Example attack and defense, AI Trojan attack and defense, Model Inversion attack and defense, Dataset Poisoning attack and defense, Algorithm Poisoning attack and defense, Model Poisoning attack and defense, Privacy: Parameter Inference (model extraction, model theft) attack and defense, Privacy: Membership Inference attack and defense, Privacy: Sensitive test data protection, Backdoor Injection attack and defense, Securing AI development and training. Through this approach, students can gain hands-on problem-solving experience in preventing and predicting suspicious security attacks and threats using suitable machine learning and deep learning techniques. By developing the skills and knowledge needed to combat adversarial attacks and other security threats, students can become better equipped to tackle real-world cybersecurity challenges.

II. Authentic Labware Design

The portable labware is created by designing, developing, and deploying it on the open-source Google CoLaboratory (CoLab) environment. This enables learners to access, share,

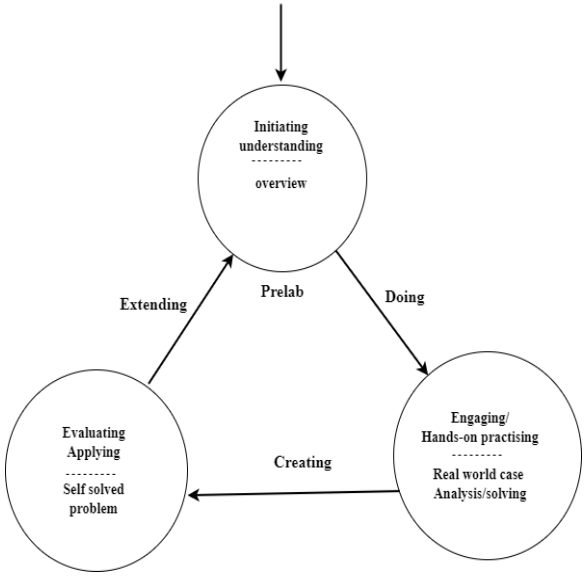

Fig. 1: Case-study based Learning Model

and practice all labs interactively with browsers anywhere and anytime without requiring tedious installation and configuration. Each module in the case study-based portable hands-on labware is designed around a specific real-world ransomware case and comprises three components: a pre-lab for conceptualization and getting started with a "Hello World" example, a hands-on lab activity with concrete real-world data sets, and a post-add-on lab with additional real-world data sets.

A. Pre-Lab for conceptualization and getting started

The Pre-Lab module introduces a specific security case study with a focus on the root of security threats, attack strategies, and their consequences. It provides an overview of ML solutions for such security issues, including prevention and detection. A simplified "hello world" example for the Adversarial Machine Learning attack case and its corresponding ML solution are demonstrated. This allows students to watch, observe and gain perspective insight into the processing, preparing them with a specific security case for conceptual understanding and getting started experience with ML solutions. It helps students build a basic understanding of why these security issues need to be fixed by using the machine learning algorithm. Fig. 2 shows a screenshot of Pre-lab in Module 1: ML for adversarial example attack and defense.

B. Hands-on activity lab for doing with concrete hands-hands experience

The hands-on activity labs are created and implemented using the Google CoLab collaboration platform, which is an online browser-based environment that offers free Google cloud services. With just a Google account, students can access CoLab and run the lab on any mobile device or laptop, anytime and anywhere. Completing the hands-on activity lab provides

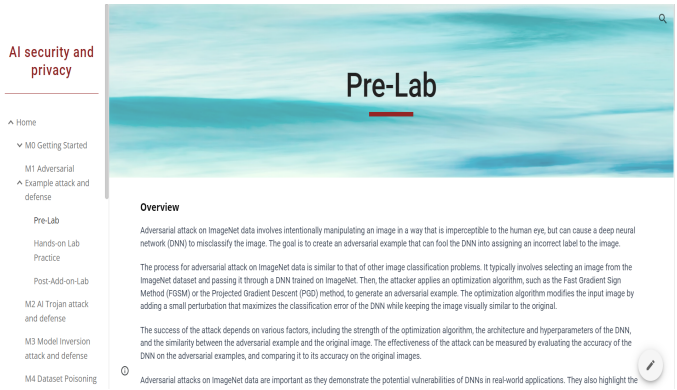

Fig. 2: Screenshot of Module 1 Pre-Lab

students with practical experience in problem-solving. The step-by-step screenshots assist students in practicing with more direction, offering visual cues to enhance learning.

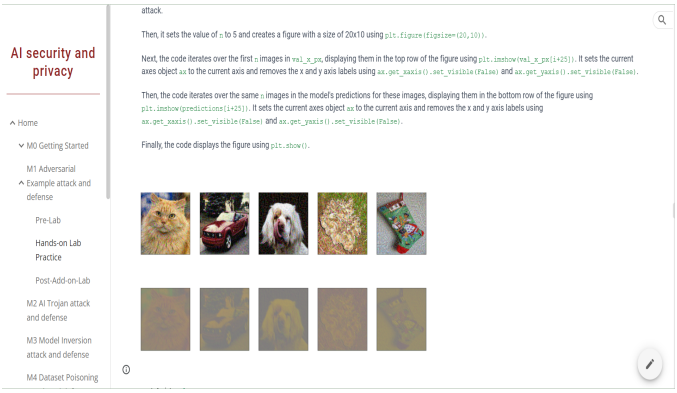

Fig. 3: Screenshot of Module 1 Lab consists of plotting the images added with perturbation

C. Post add-on lab for creative enhancement

The Post add-on lab encourages students to engage in reflective thinking on the given case, as well as hands-on experimentation to enhance problem-solving abilities. Through the lab, students can improve their prediction and detection accuracy rates by exploring new and creative ideas and conducting active testing and experiments. Learners are motivated to find more effective ML/DL algorithms for attack detection prevention, and can share their creative work with others on Colab. The post-add-on lab aims to promote active learning and problem-solving among studentsThe main objective of this project is to tackle the needs and difficulties of learning about ML/DL for security, including attacks, by providing authentic hands-on practice and addressing the lack of educational materials. The initial feedback from students regarding the selected learning modules has been positive.

ACKNOWLEDGEMENT

The work is supported by the National Science Foundation under NSF Award #2100134, #2100115, #2209638,

#2209637, #1663350. Any opinions, findings, recommendations, expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

III. CONCLUSION

The main objective of this project is to tackle the needs and difficulties of learning about ML/DL for security, including attacks, by providing authentic hands-on practice and addressing the lack of educational materials. The initial feedback from students regarding the selected learning modules has been positive.

REFERENCES

[1] D. Jeong, "Artificial intelligence security threat, crime, and forensics: Taxonomy and open issues," *IEEE Access*, vol. 8, pp. 184560–184574, 2020.

[2] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting ai trojans using meta neural analysis," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 103–120, IEEE, 2021.

[3] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.