Cybersecurity Education in the Age of Artificial Intelligence: A Novel Proactive and Collaborative Learning Paradigm

Jin Wei-Kocsis, *Member, IEEE,*, Moein Sabounchi, Gihan J. Mendis, Praveen Fernando, Baijian Yang, *Member, IEEE,*, and Tonglin Zhang

***Abstract*—*Contribution:* A novel proactive and collaborative learning paradigm was proposed to engage learners with different backgrounds and enable effective retention and transfer of the multidisciplinary AI-cybersecurity knowledge. Specifically, the proposed learning paradigm contains: 1) an immersive learning environment to motivate the students for exploring artificial intelligence (AI)/machine learning (ML) development in the context of real-world cybersecurity scenarios by constructing learning models with tangible objects, and 2) a proactive education paradigm designed with the use of collaborative learning activities based on game-based learning and social constructivism.**

***Background:* Increasing evidence shows that AI techniques can be manipulated, evaded, and misled, which can result in new and profound security implications. There is an education and training gap to foster a qualified cyber-workforce that understands the usefulness, limitations, and best practices of AI technologies in cybersecurity domain. Efforts have been made to incorporate a comprehensive curriculum to meet the demand. There still remain essential challenges for effectively educating students on the interaction of AI and cybersecurity.**

***Intended Outcomes:* A novel proactive and collaborative learning paradigm is proposed to educate and train a qualified cyber-workforce in this new era where security breaches, privacy violations, and AI have become commonplace.**

***Application Design:* The development of this learning paradigm is grounded in the pedagogical approaches of technology-mediated learning and social constructivism.**

***Findings:* Although the research work is still ongoing, the prototype learning paradigm has shown encouraging results in promoting the learners' engagement in applied AI learning.**

***Index Terms*—Artificial intelligence, machine learning, cybersecurity, education.**

I. Introduction

THE phenomenal growth of AI techniques, especially ML, has great impacts on all aspects of human life. Examples include autonomous and semi-autonomous security systems that are demonstrating impressive promises for increasing awareness, reacting in real-time, and improving the overall effectiveness [1]–[3]. According to VynZ Research, the global AI in cybersecurity market reached 12 billion USD in 2020 and will grow to 30.5 billion USD in 2025 [4]. However, increasing evidence shows that AI techniques can be manipulated, evaded, and misled, which can result in new and profound security implications [5], [6]. While prominent research progress has been made in understanding the trust and security of AI techniques [7], there is an education and training gap to foster a qualified cyber-workforce that understands the usefulness, limitations, and best practices of AI technologies in cybersecurity domain. Recent reports also indicate that such an education gap will throttle aspirations in the advance of AI and intensify the shortage problem in cybersecurity workforce [8]–[10].

Efforts have been made to incorporate a comprehensive curriculum to meet the demand. However, there still contain essential challenges for effectively educating students on the interaction of AI and cybersecurity. Examples include: (1) due to the emerging and growing features of AI technologies and zero-day exploits, the integration of AI and cybersecurity technologies are rapidly and dynamically evolving; (2) students may have very diverse knowledge backgrounds, ranging from conventional information technology to data science, inducing various needs for inspiring skill and interaction engagement; and (3) while significant studies have been developed in understanding AI/ML-specific threats, most of the existing research focuses on computer vision domain with limited efforts made in the cybersecurity domain that is complex and rife with adversaries. In this research, to address these challenges, the researchers aim to educate and train a qualified cyber-workforce in this new era where security breaches, privacy violations, and AI have become commonplace. In the current state of work, the researchers have developed a Minecraft-based immersive learning environment for AI/ML development in cyber security domain. The researchers have also designed team-based tasks via the proposed Minecraft-based immersive learning environment by exploring game-based learning and social constructivism theories, which is a critical step for developing the game-based proactive and collaborative learning paradigm with hackathon activities.

The rest of the paper is organized as follows. Section II introduces the background and related work. The system design is elaborated in detail in Section III. The system deployment results are presented in Section IV. The paper is concluded in Section V with future work highlighted.

II. Background and Related Work

Statistical ML algorithms have been extensively used in the field of cybersecurity, such as spam detection [11], malware detection [12], and network intrusion detection [13]. The rise of deep learning (DL) approaches offers a promising direction in discovering sophisticated and unseen attack patterns [14]. While the benefits of DL are immense, the black-box nature of DL casts doubts in the decision-making process. More

importantly, when the new intelligent component is applied to an existing cybersecurity system, it increases the attack surface and is subjected to additional attacks. A typical ML workflow starts with data collection, followed by data pre-processing to clean up the noises, normalize the scales, and manage the missing data. Feature selection and feature engineering are often needed before feeding the training data to various ML algorithms. Since there is no single model that fits all, inference models are evaluated on the validation data and the best model will be selected to predict unseen new data. When the performance deteriorates, the model update will be triggered to retrain the model. Every step of the ML process is subjected to attacks, as detailed in the next paragraph.

Recently, adversarial attacks on ML techniques, especially DL systems, received much attention from both the AI and the security communities [5], [15], [16]. For data acquisition and prepossessing, various attacks were proposed to mislead the classifier using data poisoning attacks [17], [18]. Likewise, feature selections could be reduced to impair the accuracy of ML algorithms [19]. A significant portion of the work in adversarial attacks was targeting the ML algorithms themselves. Adversarial perturbations were generated in the direction of the gradient to attack the models, such as FGSM [20] and JSMA [21]. Attacks could also be launched to degrade the confidence scores of the classifiers, such as the ZOO attack presented in [22]. Decisions could also be confused if the attackers feed the system with adversarial examples crafted with Generative Adversarial Network (GAN) techniques [23]. The above-described attacks are well discussed in the context of computer vision and natural language processing. More studies are needed to understand how these attacks impact the security of a system. Given that cybersecurity problems are often manifested in the form of binary classifiers, attacks on AI/ML-driven security systems will be even more dangerous because attacks on one model can be easily transferred to another model in the case of binary classifications [24].

From an education perspective, new technologies are increasingly adopted to innovate teaching and learning. With recent developments in the visualization and virtual reality (VR) possess great potential in education and training [25], [26]. Various studies have been developed to apply VR technologies to develop immersive learning environments in different fields, including medicine, engineering, and construction. All of them demonstrated valuable practices of applying VR in education [27], [28]. However, the established work mainly focused on teaching procedural, practical knowledge, and declarative knowledge that can be benefited from the realistic surrounding function of VR. In contrast, the AI-cybersecurity conjecture has many abstract concepts and theories. Additionally, existing work in VR often overlooked collaborative learning and social connectivism, which are essential for staying current in rapidly evolving information ecology [29].

In summary, the current research trend on AI and cybersecurity usually focuses on two different themes: the trustworthiness of AI systems and the application of AI in the cybersecurity domain. While it is wise for the researchers to have their own concentrations, it will be ill-advised to treat AI and cybersecurity as two distinct subjects to educate the next generation of scholars. In addition, existing curricula on integrating AI and cybersecurity are often reactive in nature: lectures and labs are created in response to known attacks. A transformative approach is needed to educate the learners to become proactive thinkers and practitioners. Although innovative technologies, have been adopted in the education domain, it is not yet clear how complex, dynamic, and abstract disciplines can benefit from this innovative pedagogy.

III. System Design

The overview of the proposed virtual, proactive, and collaborative learning paradigm, which enables the innovative integrated cybersecurity and AI/ML curriculum, is illustrated in Fig. 1. As shown in Fig. 1, the proposed paradigm

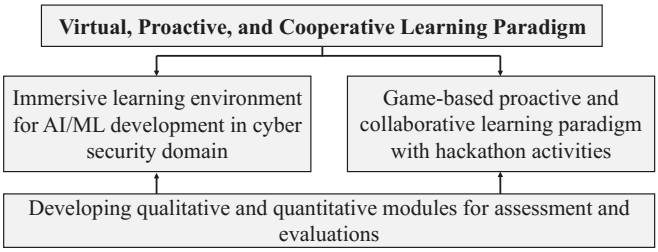


Fig. 1. Overview of the proposed virtual, proactive, and collaborative learning paradigm.

mainly consists of two main components: (1) a cost-effective immersive learning environment that motivates the students to explore AI/ML development in the context of real-world cybersecurity scenarios by constructing learning models with tangible objects; and (2) a game-based proactive education paradigm with the use of hackathon activities that engages students with a diverse background to collaboratively formulate AI/ML-specific threats and develop trustworthy and robust AI/ML solutions in cybersecurity domain. In the current stage of research, the researchers have been leveraging a cost-effective Minecraft platform to develop an immersive learning environment where the learners are able to proactively investigate the insights of the emerging AI/ML concepts by constructing related learning modules via interacting with tangible AI/ML building blocks in different cybersecurity application scenarios. The development achievement for the proposed immersive learning environment mainly consists of ML tangible module development and cybersecurity scenario module design. Additionally, the researchers have designed team-based tasks for deploying the proposed immersive learning environment via game-based activities, which is the early-stage development for the realization of the proposed game-based proactive and collaborative learning paradigm with hackathon activities.

A. ML Tangible Module Development

In the proposed ML tangible module development, the building blocks are developed to represent primitive units with different degrees of granularity. The interaction between the building blocks throughout the AI/ML workflow is also visualized to the learners for their individual or collaborative

investigation. In the current version of the immersive learning environment, the individual and interactive investigation is supported and the learning modules include 1) a logistic regression module, 2) different neural network modules including fully-connected neural network, convolutional neural network, recurrent neural network, autoencoder, and generative adversarial networks, and 3) deep reinforcement learning module.

1) Logistic Regression Module: As illustrated in Fig. 2, in the proposed immersive learning environment that is developed by leveraging MCPI API [30] and PythonTool Mod [31], the learners are able to select a dataset from the multiple available datasets and develop a logistic regression model with tangible building blocks by customizing the hyperparameters of the building blocks and the settings of the learning procedure via the interactive text-based user interface. The selected dataset can be visualized in the proposed learning environment. Additionally, the critical parameters of the logistic regression model during training and testing procedures, such as weights and bias, are also visualized via a color scale. Further, to enhance the transparency of the training and testing procedure, the decision boundary during the training and testing can also be visualized in real time. By using the decision boundary visualization, the learners are able to achieve more insights about the impact of different parameters and hyperparameters on the performance of the realized learning model.

Fig. 2. Screenshot of immersive learning environment for developing logistic regression model and explanations.

2) Neural Network (NN) Modules: In the learning environment, the researchers also integrate the development of different NN modules that have been widely adopted for different applications in cybersecurity domain.

a) Fully-Connected Neural network (FCNN): As illustrated in Fig. 3, while integrating the development of FCNN model in the learning environment, the researchers also explore another type of user interface where the learners are able to customize building blocks and ML procedure by setting the switches on a control panel instead of typing any text. While enabling the development of FCNN, the researchers noticed that the immersive learning platform has limitations on supporting multiple learners for collaborative investigation and on supporting the development of ML modules with high complexity and scalability. To address the limitations, the researchers explore alternative strategies to realize an immersive learning environment by leveraging flask server, Bukkit server, and Spigot API. The two screenshots of the immersive learning environment using this new strategy with different angles are shown in Fig. 4. In this environment,

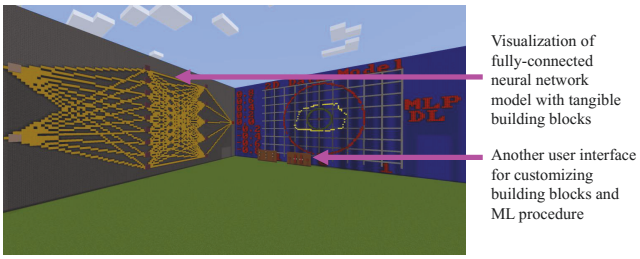


Fig. 3. Screenshot of immersive learning environment for developing FCNN model and explanations.

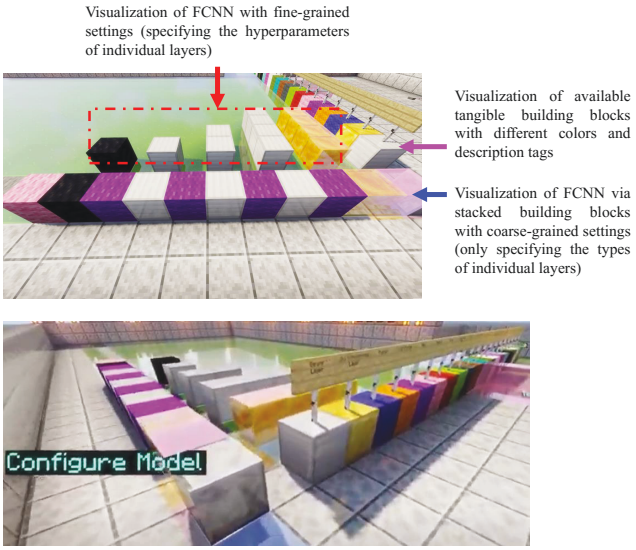


Fig. 4. Screenshots of an immersive learning environment with a new strategy for developing FCNN model via two different angles.

there are tangible building blocks whose concepts are provided via the associated colors and tags. For example, a building block can be used for realizing a Dense layer with Sigmoid or Softmax activation functions. In this environment, the learners are able to: (1) select the tangible building blocks based on the colors and the tags associated with the available building blocks; (2) interact with the selected tangible building blocks with different granularities; and (3) develop the ML models by stacking the building blocks.

b) Convolutional Neural Network (CNN): As shown in Fig. 5, the representation of an immersive learning environment for developing a CNN model is very similar to the screenshots in Fig. 4. The main difference is that the researchers integrate the additional visualization mechanism for enhancing the learners' experience on CNN development. The additional visualization mechanism includes visualizing the kernel weights and feature maps associated with the convolutional layers. Figure 6 shows a screenshot by zooming in on the outputs of the visualization mechanism in Fig. 5.

c) Recurrent Neural Network (RNN): The screenshot of the immersive learning environment for developing a Long short-term memory (LSTM)-based RNN model is shown in Fig. 7. The visualization mechanism is designed for visualizing the gate outputs and the hidden states.

Fig. 5. Screenshot of immersive learning environment for developing a CNN model with the outputs of visualization mechanism on the top.

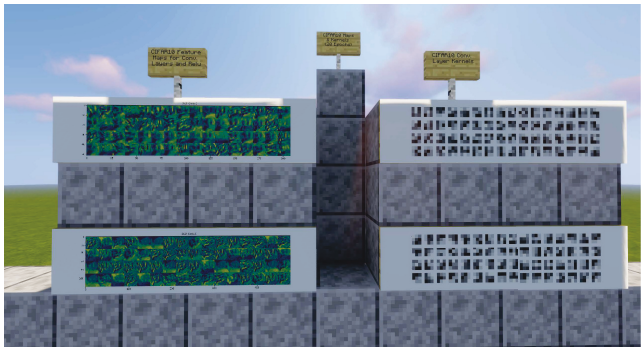
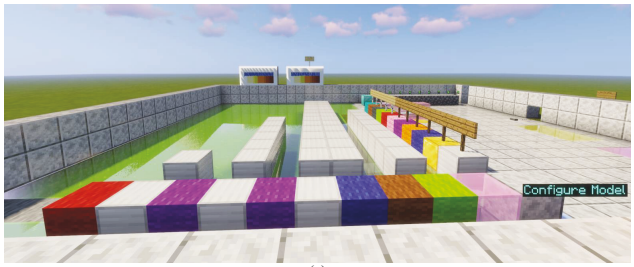
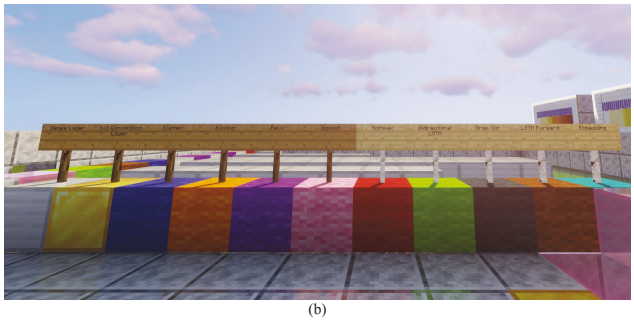


Fig. 6. Screenshot of the outputs of the visualization mechanism of developing a CNN for a classification task: Left: visualization of feature maps after given epochs, and Right: visualization of kernel weights after given epochs.



(a)



(b)

Fig. 7. Screenshots with different viewpoints of immersive learning environment for developing an LSTM-based RNN model.

d) Autoencoder: The screenshot of the immersive learning environment for developing an autoencoder model is shown in Fig. 8. The outputs of the visualization mechanism, as shown in Fig. 8, include the reconstructed data with different hyperparameters and parameters.

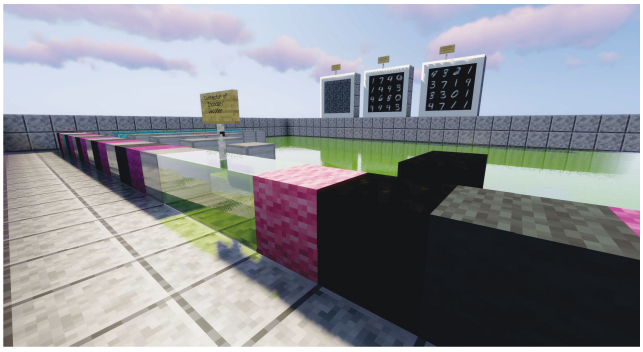


Fig. 8. Screenshot of immersive learning environment for developing an autoencoder model with the outputs of visualization mechanism on the top.

e) Generative Adversarial Network (GAN): The screenshot of the immersive learning environment for developing a GAN model is shown in Fig. 9. As shown in Fig. 9, the GAN

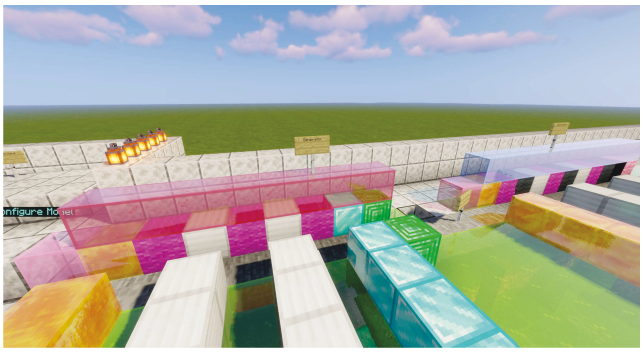


Fig. 9. Screenshot of immersive learning environment for developing a GAN model.

model consists of a generator associated with the pink glass block and a discriminator associated with the blue glass block. The training of GAN is realized by triggering the generator and discriminator blocks interactively. As shown in Fig. 10, the outputs of the visualization mechanism include the generated data with different hyperparameters and parameters.

3) Deep Reinforcement Learning (DRL): Considering that DRL technologies have been widely applied for automating cybersecurity-related operation systems, the researchers have also been working on integrating DRL modules in the immersive learning environment. To reduce the cognitive load of the learners to work on this advanced topic and to further enhance the learners' learning experience, the researchers have been exploiting imitation learning and interactive reinforcement learning techniques [32]–[36] as well as game-based learning pedagogical principle to develop interactive and practical application environment where the learners are able to advance their understanding of DRL performance by proactively interacting with the learning model. To realize this goal, in the current development stage, the DRL application

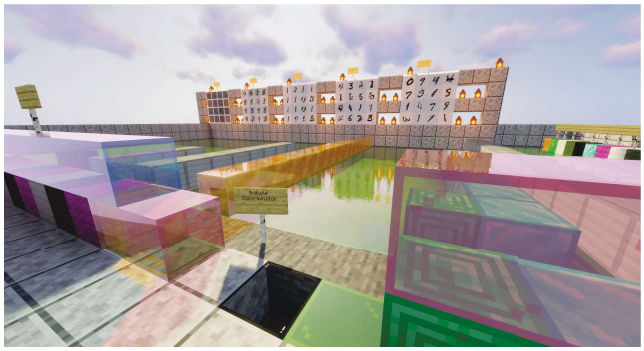


Fig. 10. Screenshot of immersive learning environment for developing a GAN model with another angle, where the outputs of the visualization mechanism are shown on the top.

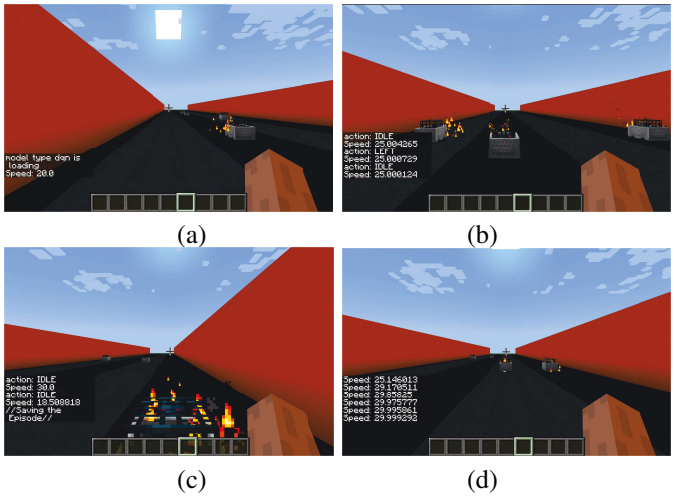


Fig. 11. Screenshots of the proposed immersive learning environment for proactively learning the DRL model in an interactive self-driving car game: (a) applying the DRL model without any leaner input; (b) and (c) learner interacts with the DRL model via demonstrating the end-to-end learning trajectories; and (d) applying the learning model taking the learner demonstration as input.

environment and the DRL development with tangible building blocks are realized in two different Minecraft/Unity platforms that communicate with each other by leveraging the Flask server. For realizing the DRL application environment in Minecraft, the researchers leverage Spigot plugin to introduce new commands to Minecraft to generate DRL-related visualizations. The Spigot plugin is developed in JAVA programming language using Bukkit API [37] to realize functionalities on Minecraft and using OKHttp3 API [38] to realize Web API client functionalities. In the following, the researchers demonstrate the functionalities of the proposed learning environment for DRL learning in one of the game-based learning scenarios, a self-driving car game. In this simplified self-driving car game, the objective of the DRL module development is to enable the car to run as fast as possible while avoiding collisions. The action space consists of idle (i.e. keeping the current speed and direction), accelerating, braking, turning left, and turning right. The state space is currently designed to include the absolute coordinates of the associated car and the relative coordinates of the four neighboring cars. In this game, the learners are encouraged to proactively explore different DRL models, including deep Q-learning (DQN), by interacting with the DRL models via different interaction approaches. In one of the interaction approaches, the learners are able to demonstrate the end-to-end learning trajectories, which will be used as inputs of the DRL model. Some screenshots of the environment are shown in Fig. 11.

B. Cybersecurity Scenario Module Design

To inspire learners to proactively explore the development of emerging AI/ML technologies in the context of the cyber security domain, the researchers have been working on developing a cybersecurity-scenario module to generate and visualize emerging real-world security threats by creatively integrating real-world datasets and the datasets resembling real-world applications with immersive scenes. One of the essential challenges is that cybersecurity-related datasets are usually associated with complex and/or large-scale systems that are difficult to be visualized in Minecraft. To address this challenge, in the current development stage, the researchers provide the learners with the application scenario information via: 1) visualizing the cybersecurity-related datasets and/or 2) visualizing the critical information of the associated cybersecurity-related systems such as topology of the systems; and/or 3) visualizing DL procedure while interacting with the systems. The researchers are still working on investigating other practical solutions to address the challenge and to provide the learners with more immersive cybersecurity scenarios. In the following, the researchers demonstrate the development achievement via two developed cybersecurity scenarios where the learners are able to proactively investigate the usefulness, limitations, and best practices of the DL models in solving practical cybersecurity problems.

1) Scenario 1: This scenario is developed to support the learners to proactively explore the application of different deep neural network (DNN) models to address different cyber attacks by analyzing related benchmark datasets. To achieve this goal, the researchers currently included CIC-IDS2017 which is an intrusion detection evaluation dataset [39], [40], CIC-MalMem-2022 [41], [42] which is an obfuscated malware dataset, and CIRA-CIC-DoHBrw-2020 which is a dataset consisting of benign and malicious Domain Name System (DNS) over HTTPS (DoH) traffic as well as non-DoH traffic [43], [44]. The screenshot of the proposed immersive learning environment with this scenario is shown in Fig. 12. As illustrated

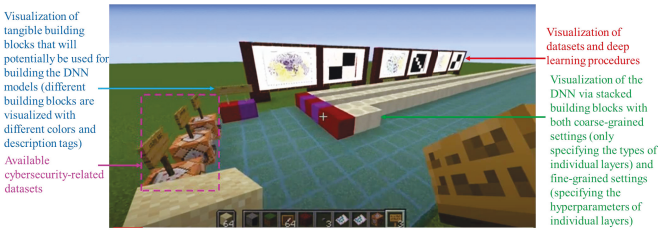


Fig. 12. A screenshot of the proposed immersive learning environment with the first scenario.

in Fig. 12, in this scenario, the learners are able to selectively visualize the tangible building blocks that will potentially

be used for building the DNN models. The specifications of the tangible building blocks are visualized with different colors and description tags. Additionally, the learners can select the available cybersecurity-related datasets by turning on the associated switch. Based on the selected cybersecurity-related dataset, the learners can proactively use the tangible building blocks to construct the DNNs for the given task such as attack detection and identification. The constructed DNN is visualized in the environment with both coarse-grained settings (i.e., only specifying the types of individual DNN layers) and fine-grained settings (i.e., specifying the details of each layer including hyperparameters). Furthermore, as illustrated in Fig. 12, the application scenario is visualized by visualizing datasets by employing t-SNE [45] and visualizing the DL procedure by plotting confusion matrices. The visualizations are realized in the Minecraft platform by leveraging ImageOnMap which is a Bukkit plugin [46].

To further enhance the visualization of the cybersecurity scenarios and improve the immersive learning experiences of the learners, the researchers are investigating extending the 2D visualization of the cybersecurity dataset to the immersive 3D visualization in the Minecraft-powered learning environment. To achieve this goal, the researchers are exploiting the Spigot plugin, Minecraft entities, and other evaluation tools suitable for the Minecraft environment. The screenshots of the achievement on immersive 3D visualization for the CIC-IDS2017 dataset are shown in Fig. 13. As shown in Fig. 13, the

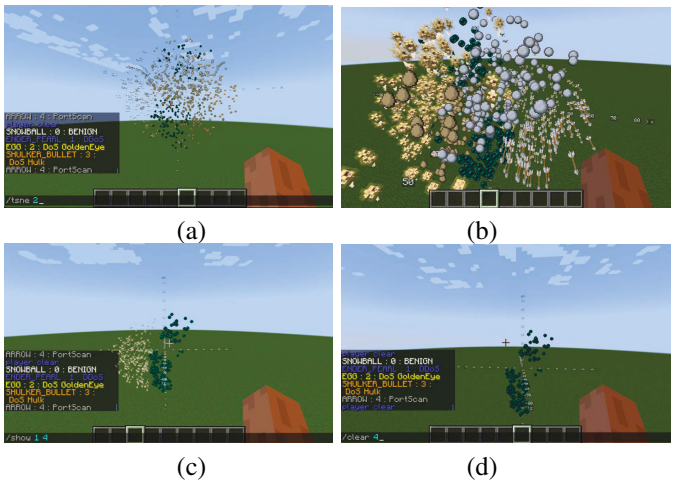


Fig. 13. Screenshots of the immersive 3D visualization of the CIC-IDS2017 dataset in the Minecraft-powered immersive learning environment: (a) the learners are able to select dimension-reduction technique to reduce the high-dimensional data samples into 3D data samples and visualize the data, (b) the learners are able to closely explore the correlations between the data samples from different perspectives and angles, (c) the learners are able to conveniently visualize the data belonging to selected classes, and (d) the learners are able to conveniently clear certain data from the visualization.

learners are able to select dimension reduction techniques such as t-SNE to reduce the high-dimension data in CIC-IDS2017 to 3D data as shown in Fig. 13(a). Then the 3D data belonging to different classes such as Benign, DDoS, DoS GoldenEye, DoS Hulk, and PortScan, are visualized by different types of Minecraft entities. With the proposed 3D data visualization, the learners are able to closely interact with the targeted data samples and have more perspectives to observe and investigate the correlation between the data samples belonging to the same and different classes as illustrated in Fig. 13(b). The proposed visualization module also provides the learners with convenient commands to activate or clear the visualization of the 3D data belonging to certain classes as illustrated in Fig. 13(c) and (d), respectively.

2) Scenario 2: In this cybersecurity scenario, the learners are able to proactively explore the application of DRL techniques in mitigating distributed denial of service (DDoS) attacks. To achieve this goal, the researchers exploited the dataset CIC-DDoS2019 which consists of benign and up-to-date common DDoS attacks [47], [48] and carefully designed the DDoS detection and mitigation scenario as shown in Fig. 14. In the designed DDoS scenario, the high-level

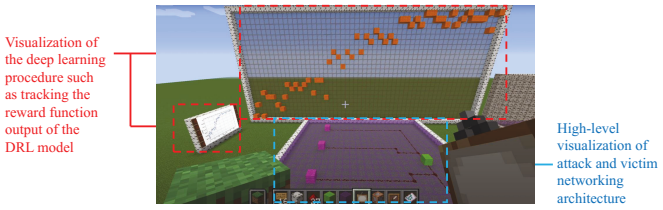


Fig. 14. A screenshot of the DDoS detection and mitigation scenario in the proposed immersive learning environment.

topology of the attack and victim networking architecture is visualized based on geographical regions. The collaborative geographical regions of the attack network are illustrated via pink blocks and the geographical region of the victim network is illustrated via a green block. In this scenario, the geographical regions of the attack network collaborate with each other to launch the attack to compromise the victim region. To be stealthy, at each time step, only one region in the attack network launches the attack and the other regions transmit the benign information to the victim region. The task is to develop a DRL model to detect the region launching an attack and disconnect the communication between this region and the victim region. Therefore, to develop the DRL algorithm, the action space can be formulated to include the identification of the malicious region index and the decision on whether to disconnect the communication between a given region and the victim region. The state space can be formulated to include received information and the previously identified malicious region. Additionally, in this scenario, the researchers provide the learners with the capability to visualize the deep learning procedure while interacting with the systems. As shown in Fig. 14, the learners can select to visualize the procedure by tracking the output of the reward function of DRL in two different ways.

IV. System Deployment Results

As illustrated in Fig. 1, to enhance the engagement of learners with diverse backgrounds in the learning process, the researchers have been exploring game-based learning and social constructivism theories and working on designing the strategy for deploying the proposed immersive learning environment via game-based teamwork activities. One essential

challenge of strategy development is to design motivating and practical tasks where the learners are inspired to explore the usefulness, limitations, and best practices of AI/ML technologies in the cybersecurity domain. The researchers have designed several team-based tasks via the proposed Minecraft-based immersive learning environment. In the ongoing work, the researchers are refining the team-based tasks, including the suitable deployment strategies, and developing more team-based tasks. After this, the researchers will leverage the well-designed team-based tasks to apply the game-based proactive and collaborative learning paradigm with hackathon activities in the classroom.

In the following, the researchers will demonstrate the achievements by using a team-based task related to adversarial attack. In this task, one team plays the role of operator and defender on a DNN-related task. In this team, some learners focus on getting an in-depth understanding of the associated dataset by exploring visualization techniques, some learners focus on developing DNN models to complete the task, and other learners focus on detecting the potential adversarial attacks compromising the developed DNN models. The other team plays the role of an attacker by identifying suitable adversarial attack techniques for compromising the DNN models and launching identified adversarial attacks. Some screenshots are shown in Figs. 15 to 18. As illustrated in Fig. 15, the

Fig. 15. Screenshots of fetching the selected data based on the given task.

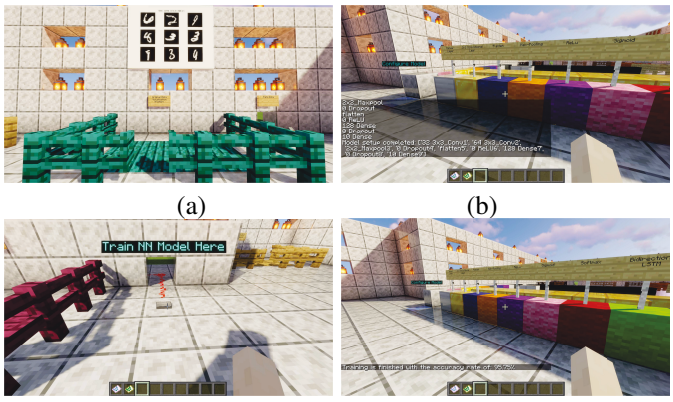


Fig. 16. Screenshots of exploring visualization techniques for advancing the understanding of the fetched data for given task (shown in (a)) and constructing and training the DNN model based on the given task and the associated data (shown in (b) to (d)).

learners can first fetch the data for the task by selecting the dataset provided by the platform. The learners in the operator/defender team are able to explore the visualization techniques to advance the understanding of the data samples as illustrated in Fig. 16(a). Based on the goal of the given task and the fetched data, the learners in the operator/defender team are able to further construct the DNN model by stacking the tangible building blocks and train the constructed DNN model as illustrated in Fig. 16(b) to (d).The learners in the attacker

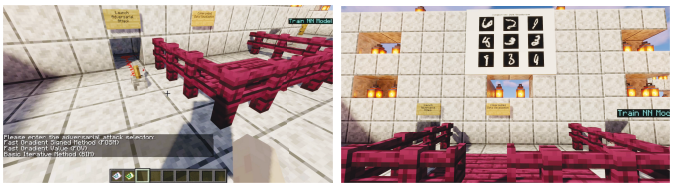


Fig. 17. Screenshots of investigating and launching an adversarial attack to compromise the constructed DNN model.

team are able to investigate, determine, and launch the adversarial attacks to compromise the DNN model as illustrated in Fig. 17(a). The learners in the attacker team can also examine the effect of the adversarial attacks, such as visualizing the compromised data to check the stealthiness of the attack as illustrated in Fig. 17(b). The learners in the operator/defender

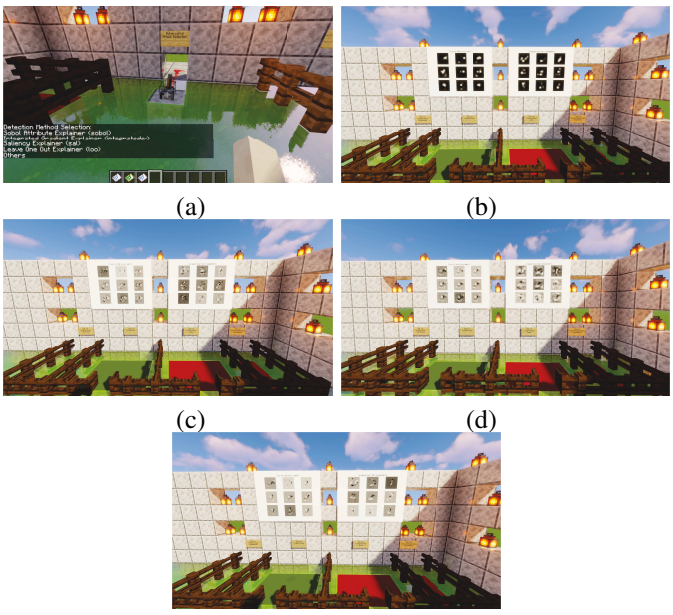


Fig. 18. Screenshots of investigating detection techniques for detecting the potential adversarial attack and developing the practical detection solution based on the investigation.

team are responsible for ensuring the trustworthiness of the DNN model by detecting the potential adversarial attacks as illustrated in Fig. 18(a). As shown in Fig. 18(a), for detecting potential adversarial attacks, the learners can either use the detection techniques available in the platform or develop their own detection solutions. In the current development stage, four eXplainable-AI (XAI)-based detection techniques are provided by the platform, including Sobol-attribute-based, Integrated gradients-based, Saliency-based, and Leave-One-Out (LOO)-based methods [49]–[52]. The learners are able to investigate the available detection techniques and visualize their effectiveness when considering different potential adversarial attacks

given the targeted operation tasks and the associated data, as illustrated in Fig. 18(b) to (e). Based on the investigation, the learners are able to determine a practical detection solution to mitigate the adversarial attacks launched by the attacker team.

V. Conclusions

The overarching goal of the research is to address a critical need to foster a qualified cybersecurity workforce that understands the usefulness, limitations, and best practices of AI/ML technologies in the cybersecurity domain. To achieve this goal, the researchers proposed a novel proactive and collaborative learning paradigm to engage learners with different backgrounds and enable effective retention and transfer of multidisciplinary AI-cybersecurity knowledge. Specifically, in the proposed learning paradigm: 1) an immersive learning environment is developed to motivate the learners to explore AI/ML development in the context of real-world cybersecurity scenarios by constructing learning models with tangible objects, and 2) a proactive education paradigm is designed with the use of collaborative learning activities based on game-based learning and social constructivism. At the current research stage, the researchers have been leveraging Minecraft to develop an immersive learning environment where the learners are able to investigate the insights of emerging AI/ML concepts by constructing related learning modules via interacting with tangible AI/ML building blocks in different cybersecurity scenarios. The researchers have realized the early-stage achievement on implementing the proposed game-based proactive education paradigm by designing the team-based tasks for deploying the proposed immersive learning environment via game-based activities.

For the future work, the researchers plan to first complete the development of the game-based proactive and collaborative learning paradigm. Then, the researchers will determine suitable strategies to apply the proposed proactive and cooperative learning paradigm to transform different courses with the multidisciplinary curriculum of AI/ML and cybersecurity. The courses include an undergraduate course, Applied Machine Learning, and a graduate course, Deep Learning for Cybersecurity and Privacy, at Purdue University. Both courses are offered in each semester with an approximate total enrollment of 50 students per semester. The deployment plan currently identified is to first transform the semester projects of the courses in the form of hackathon activities via the proposed Minecraft-based immersive learning environment. To ensure the inclusiveness of the proposed hackathon activities for students with diverse backgrounds, it is important to enable an easy start for all the students to participate in the hackathon activities. To achieve this goal, the researchers will work together with the students in each course to identify an appropriate schedule for the hackathon activities to ensure the targeted students have been prepared with enough knowledge and skills to explore the advanced topics. Additionally, the demonstration of baseline solutions, such as the solutions of launching adversarial attacks and detecting adversarial attacks in the team-based task described in Section IV, will be provided, which will help the students to get started quickly. The students will then be asked to proactively improve the baseline models in the hackathon activities. Chatbot and discussion forum will also be created to offer necessary support for the students who may find the activities too challenging to engage and succeed. To evaluate the performance of the proposed proactive and cooperative learning paradigm, qualitative and quantitative evaluation and assessments will be developed. For example, questionnaires and interviews will be designed to obtain and analyze the information on the students' perception of using the proposed Minecraft-powered immersive learning environment, their knowledge gain, their satisfaction about the learning outcomes, and their altitude and feedback on the hackathon activities including the topics and mechanism of the team-based tasks. Furthermore, considering the rapid evolution of both AI/ML and cybersecurity domains, it is essential to ensure the students to keep staying with the state-of-the-art technologies. Therefore, the researchers will continuously improving the proposed proactive and cooperative learning paradigm while applying the paradigm to the courses. For example, the emerging transformers and the related ML models, such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT), have gained significant attention recently and demonstrated great potentials in different application fields. The researchers plan to improve and extend the functionalities of the current immersive learning environment by incorporating transformers and related ML models.

Acknowledgments

This work was funded by the U.S. National Science Foundation (NSF) research grant 2114974.

References

[1] Y. He, G. J. Mendis, and J. Wei, "Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2505–2516, 2017.

[2] E. Tsukerman, *Machine Learning for Cybersecurity Cookbook: Over 80 recipes on how to implement machine learning algorithms for building security systems using Python*. Packt Publishing; 1st edition, 2019.

[3] D. Berman, A. Buczak, J. Chavis, and C. Corbett, "A survey of deep learning methods for cyber security," *Information*, vol. 10, no. 4, pp. 1–35, 2019.

[4] V. Research. (2019) Global artificial intelligence (ai) in cyber security market – analysis and forecast (2019-2025). [Online]. Available: https://www.vynzresearch.com/ict-media/artificial-intelligence-in-cyber-security-market

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proceedings of International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017.

[6] "Artificial intelligence and cybersecurity: Opportunities and challenges technical workshop summary report," *A Report by Networking & Information Technology Research and Development Subcommittee and the Machine Learning & Artificial Intelligence Subcommittee of the National Science & Technology Council*, March 2020.

[7] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[8] "A report to the president on supporting the growth and sustainment of the nation's cybersecurity workforce: Building the foundation for a more secure american future," *Transmitted by The Secretary of Commerce and The Secretary of Homeland Security*, May 2017.

[9] "Federal cybersecurity research and development strategic plan," *Prepared by the Cyber Security and Information Assurance Interagency Working Group Subcommittee on Networking & Information Technology Research & Development Committee on Science & Technology Enterprise of the National Science & Technology Council*, December 2019.

[10] "U.S. Department of Homeland Security Artificial Intelligence Strategy," December 2020.

[11] K. Tretyakov, "Machine learning techniques in spam filtering," in *Data Mining Problem-oriented Seminar, MTAT*, vol. 3, no. 177. Citeseer, 2004, pp. 60–79.

[12] D. Gavriluţ, M. Cimpoeşu, D. Anton, and L. Ciortuz, "Malware detection using machine learning," in *2009 International Multiconference on Computer Science and Information Technology*. IEEE, 2009, pp. 735–741.

[13] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE symposium on security and privacy*. IEEE, 2010, pp. 305–316.

[14] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35 365–35 381, 2018.

[15] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, 2011, pp. 43–58.

[16] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *arXiv preprint arXiv:1712.04248*, 2017.

[17] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[18] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in neural information processing systems*, 2017, pp. 3517–3529.

[19] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial feature selection against evasion attacks," *IEEE transactions on cybernetics*, vol. 46, no. 3, pp. 766–777, 2015.

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[21] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[22] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 2672–2680, 2014.

[24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[25] A. Natale, C. Repetto, G. Riva, and D. Villani, "Immersive virtual reality in k-12 and higher education: A 10-year systematic review of empirical research," *British Journal of Educational Technology*, vol. 51, no. 6, pp. 2006 – 2033, 2020.

[26] J. Radianti, T. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computer & Education*, vol. 147, pp. 1– 29, 2020.

[27] J. Delgado, L. Oyedele, P. Demian, and T. Beach, "A research agenda for augmented and virtual reality in architecture, engineering and construction," *Advanced Engineering Informatics*, vol. 45, pp. 1 – 21, 2020.

[28] J. Pattle, "Virtual reality and the transformation of medical education," *Future Healthcare Journal*, vol. 6, no. 3, pp. 181 – 185, 2019.

[29] G. Siemens, "Connectivism: A learning theory for the digital age," *International Journal of Instructional Technology & Distance Learning*, vol. 2, pp. 1 – 7, 2004.

[30] "Minecraft: Pi edition API Python Library," *[Available Online:] https://github.com/martinohanlon/mcpi*.

[31] "PythonTool Mod," *[Available Online:] https://ngcm.github.io/PythonTool-Mod*.

[32] B. Ziebart, J. Bagnell, and A. Dey, "Modeling interaction via the principle of maximum causal entropy," in *the 27th International Conference on Machine Learning*, 2010.

[33] ——, "Learning robust rewards with adversarial inverse reinforcement learning," in *arXiv:1710.11248v2*, 2018.

[34] T. Nguyen, Q. Zheng, and A. Grove, "Reliable conditioning of behavioral cloning for offline reinforcement learning," in *arXiv:2210.05158*, 2022.

[35] R. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *arXiv:1706.03741v4*, 2023.

[36] C. Cruz and T. Igarashi, "A survey on interactive reinforcement learning: Designprinciples and open challenges," in *arXiv:2105.12949v1*, 2021.

[37] Spigot-API 1.19.4-R0.1-SNAPSHOT API. [Online]. Available: https://hub.spigotmc.org/javadocs/spigot/

[38] OkHttpClient. [Online]. Available: ttps://square.github.io/okhttp/4.x/okhttp/okhttp3/-ok-http-client/

[39] Intrusion Detection Evaluation Dataset (CIC-IDS2017). [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html

[40] I. Sharafaldin, A. Lashkari, and A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018.

[41] Malware Memory Analysis CIC-MalMem-2022. [Online]. Available: https://www.unb.ca/cic/datasets/malmem-2022.html

[42] T. Carrier, P. Victor, A. Tekeoglu, and A. Lashkari, "Detecting obfuscated malware using memory feature engineering," in *The 8th International Conference on Information Systems Security and Privacy (ICISSP)*, 2022.

[43] CIRA-CIC-DoHBrw-2020 Canadian Institute for Cybersecurity (CIC) project funded by Canadian Internet Registration Authority (CIRA). [Online]. Available: https://www.unb.ca/cic/datasets/dohbrw-2020.html

[44] M. MontazeriShatoori, L. Davidson, G. Kaur, and A. Lashkari, "Detection of doh tunnels using time-series classification of encrypted traffic," in *The 5th IEEE Cyber Science and Technology Congress*, 2020.

[45] L. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, March 2008.

[46] ImageOnMap. [Online]. Available: https://github.com/zDevelopers/ImageOnMap

[47] DDoS Evaluation Dataset (CIC-DDoS2019). [Online]. Available: https://www.unb.ca/cic/datasets/ddos-2019.html

[48] I. Sharafaldin, A. Lashkari, S. Hakak, and A. Ghorbani, "Developing realistic distributed denial of service (ddos) attack dataset and taxonomy," in *IEEE 53rd International Carnahan Conference on Security Technology*, 2019.

[49] T. Fel, R. Cadène, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre, "Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis," *Advances in Neural Information Processing Systems*, 2021.

[50] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning, PMLR*, 2017.

[51] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," in *In Workshop at International Conference on Learning Representations*, 2014.

[52] P. Yang, J. Chen, C. Hsieh, J. Wang, and M. Jordan, "ML-LOO: detecting adversarial examples with feature attribution," *In Proceedings of the AAAI Conference on Artificial Intelligence*, April 2020.

Jin Wei-Kocsis received her Ph.D. degree in Electrical and Computer Engineering from the University of Toronto, her M.S. degree in Electrical Engineering from University of Hawaii at Manoa, and her B.E. degree in Electronic Information Engineering from Beihang University. She is currently an Associate Professor in the Department of Computer and Information Technology at Purdue University. Her research interests include applied AI education, deep learning, security and privacy for cyber-physical-social systems, and wireless communications. Her accomplishments have been recognized through the NSF awards, NASA Early Career Faculty award, Homeland Security National Training Program/Continuing Training Grant, and Transportation Research Board Grants.

Moein Sabounchi received the B.Sc. degree in Electrical Engineering from the Razi University of Kermanshah in 2011, and the M.Sc. degree in Electrical Engineering from the Ferdowsi university of Mashhad in 2014. He is currently a Ph.D. candidate at the Department of Computer and Information Technology at Purdue university. His current research interests include the utilization of data-driven technologies in engineering and education.

Gihan J. Mendis received his B.Sc. Eng. degree (Hons.) in Electronics and Telecommunication Engineering from the University of Moratuwa, Moratuwa, Sri Lanka, in 2012. He earned the M.Sc. degree in Electrical Engineering from the University of Akron, Akron, OH, USA, in 2016, and the Ph.D. in Technology from Purdue University, West Lafayette, IN, USA, in 2023. Currently, he works as a Senior AI Engineer at Blue Wave AI Labs in West Lafayette, IN, USA. His research interests include data-driven decision-making for cyber-physical systems (CPSs), deep learning for CPS security, decentralized deep learning, encrypted deep learning, and deep learning algorithms. The related work for this journal was conducted when he was a Ph.D. student at Purdue University.

Praveen Fernando received his B.Sc. degree in Electronics and Telecommunication Engineering from the University of Moratuwa, Sri Lanka, in 2017. He is currently a Ph.D. student in the Department of Computer and Information Technology at Purdue University, West Lafayette, IN, USA. His research interests include applied AI in wireless communications.

Baijian Yang received his Ph.D. in Computer Science from Michigan State University, and his MS and BS in Automation (EECS) from Tsinghua University. He is currently a Professor at the Department of Computer and Information Technology, Purdue University. He served as a steering committee member of IEEE Cybersecurity Initiative between 2015 and 2017, and served as a board director of ATMAE from 2014-2016. His research interests include cybersecurity, data-driven security analytics, and applied machine learning. In addition to his rich academic background, he holds several industry certifications, such as CISSP, MCSE, and Six Sigma Black Belt and has published two books on Windows Phone Programming.

Tonglin Zhang is an Associate Professor of Statistics at Purdue University. His research interests include (1) AI/ML, (2) big data, (3) spatial statistics, (4) mathematical statistics, (5) high-dimensional data, and (6) variable selection. His research covers both theories and applications. He has published almost 100 research articles in decent journals and reputational conferences of statistics and computer science. His role in the project is the development of AI/ML theories and computational methods for Cybersecurity Education models to help students with diverse backgrounds of the learners. He has extensive research and teaching experience in developing collaborative and explainable AI.