Machine Learning Security and Privacy

Nathalie Baracaldo | IBM Research
Alina Oprea | Northeastern University

Our special issue explores emerging security and privacy aspects related to machine learning and artificial intelligence techniques, which are increasingly deployed for automated decisions in many critical applications today. With the advancement of machine learning and deep learning and their use in health care, finance, autonomous vehicles, personalized recommendations, and cybersecurity, understanding the security and privacy vulnerabilities of these methods and developing resilient defenses becomes extremely important. An area of research called *adversarial machine learning* has been developed at the intersection of cybersecurity and machine learning to understand the security of machine learning in various settings. Early work in adversarial machine learning showed the existence of adversarial examples: data samples that can create misclassifications at deployment time. Other threats against machine learning include poisoning attacks, where an adversary controls a subset of data at training time, and privacy attacks in which an adversary is interested in learning sensitive information about the training data and model parameters.

An area of research called *adversarial machine learning* has been developed at the intersection of cybersecurity and machine learning to understand the security of machine learning in various settings.

Consequently, there is a need to understand this wide range of threats against machine learning, design resilient defenses, and address the open problems in securing machine learning deployed in practical settings. Our special issue call for papers solicited articles on critical topics related to machine learning security and privacy, including the following:

- applications of machine learning and artificial intelligence to security problems, such as spam detection, forensics, malware detection, and user authentication
- evasion attacks and defenses against machine learning and deep learning methods
- poisoning attacks against machine learning at training time, such as backdoor poisoning and targeted poisoning attacks, and corresponding defenses
- privacy attacks against machine learning, including membership inference, reconstruction attacks, model extraction, and corresponding defenses
- adversarial machine learning and defenses in specific applications, including natural language processing

Digital Object Identifier 10.1109/MSEC.2022.3188190
Date of current version: 6 September 2022

(NLP), autonomous vehicles, health care, speech recognition, and cybersecurity
- methods for federated learning and secure multiparty computation techniques for machine learning.

We were delighted to receive 10 submissions, from which we selected a set of seven articles for publication in the special issue after rigorous peer review. The accepted articles discuss important topics in private machine learning, the security of natural language models, and the robustness of machine learning used for security applications, such as malware and phishing detection.

The first three articles address several issues related to data privacy in machine learning. The first two, "Sphynx: A Deep Neural Network Design for Private Inference," by Minsu Cho, Zahra Ghodsi, Brandon Reagen, Siddharth Garg, and Chinmay Hegde, and "Complex Encoded Tile Tensors: Accelerating Encrypted Analytics," by Ehud Aharoni, Nir Drucker, Gilad Ezov, Hayim Shaul, and Omri Soceanu, discuss the problem of performing efficient private inference in neural networks over encrypted data. Private inference allows a client to outsource neural network prediction to a more powerful cloud provider such that the client does not learn anything about the cloud-hosted model parameters, and the cloud does not learn the client's input. The main challenge is that private inference is based on expensive cryptographic techniques, including homomorphic encryption and garbled circuits, and computation becomes prohibitive for large neural networks. The first article proposes the Sphynx framework for neural architecture search to minimize the number of expensive activation operations in neural networks and reduce the cost of private inference. The second introduces a different approach by representing vectors and matrices used in neural network inference as more compact "tiled tensors" and shows that this representation reduces the cost of operations performed over encrypted data.

The third article, "Data Privacy and Trustworthy Machine Learning," by Martin Strobel and Reza Shokri, discusses differential privacy in machine learning and presents an interesting analysis of how different trustworthiness objectives, including robustness, privacy, fairness, and explainability, may be at odds with one another. Interestingly, models that are designed to be explainable are also more susceptible to membership inference attacks that leak private data. The connection does not stop there: applying differential privacy may also result in fewer fair models and the application of adversarial training—a de facto defense against adversarial samples—and it may jeopardize privacy. These implications need to be considered holistically to be able to generate suitable solutions where all trustworthy objectives are covered. This article summarizes these conflicting aspects of machine learning trustworthiness and highlights the need for the research community to address them.

There is a need to understand this wide range of threats against machine learning, design resilient defenses, and address the open problems in securing machine learning deployed in practical settings.

The article "Backdoors Against Natural Language Processing: A Review," by Shaofeng Li, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Suguo Du, and Haojin Zhu, provides a survey of backdoor poisoning attacks in NLP systems. Backdoor attacks are a type of poisoning attacks in which backdoored samples are inserted by adversaries at training time to induce a targeted misclassification of the samples with the same backdoor pattern at testing time. Recently, large language models, such as Generative Pretrained Transformer (GPT) 2, GPT-3, and Bidirectional Encoder Representations From Transformers, have been using transformer-based architectures, which leverage self-attention mechanisms that model relationships among all words in a sentence. Transformers have shown superior performance in many NLP tasks, such as machine translation and question answering, but the article discusses their vulnerabilities against stealthy, hard-to-detect backdoor attacks. This is an important threat that needs to be considered to enable the deployment of these models in critical applications.

The article "Machine Learning for Source Code Vulnerability Detection: What Works and What Isn't There Yet," by Tina Marjanov, Ivan Pashchenko, and Fabio Massacci, provides an interesting study of machine learning techniques for defect detection and the automated correction of security defects in source code. Starting from around 400 techniques, this study outlines popular approaches and highlights their limitations. By including the end-to-end machine learning pipeline in the analysis, one identified limitation is the lack of access to real data that researchers have for exploring this problem. Consequently, a large number of researchers generate unrealistic synthetic data that lead to the creation of pipelines that do not generalize to real data sets. This article also highlights the growing popularity of deep neural networks in this area. Although not popular at the time of the study, we expect to see large language models used in the near future.

As the following article shows, existing machine learning solutions suffer from vulnerabilities themselves.

In "Practical Attacks on Machine Learning: A Case Study on Adversarial Windows Malware," Luca Demetrio, Battista Biggio, and Fabio Roli explain how adversarial examples can be applied to bypass malware detectors. The existence of adversarial examples has demonstrated the brittleness of machine learning models at inference time. By adding small perturbations to the input, these attacks create misclassifications. An adversary can take advantage of this weakness to generate targeted and untargeted misclassifications. These attacks were originally developed for image modality, where, for example, adding carefully crafted perturbations to a panda bear image would result in the model classifying the bear as a tiger. This article explains that these attacks go beyond the image domain and, in fact, can be applied to generate malware that can effectively bypass existing detectors. The article highlights the need to design malware detectors to be aware of these vulnerabilities.

Finally, in "Phishing Detection Leveraging Machine Learning and Deep Learning: A Review," Dinil Mon Divakaran and Adam Oest present an overview of a diverse set of methods to detect phishing attacks. The article covers uniform resource locater-based analysis techniques that use feature selection: image classifiers that utilize Siamese networks to differentiate between legitimate and phishing web pages, among other machine learning techniques. The generation of robust techniques for phishing detectors is still in its infancy and, since it is based on machine learning models, inherits the same risks and vulnerabilities. For example, training data may be compromised by an adversary who may launch a backdoor poisoning attack or a clean label attack. In fact, in this application, adversaries have plenty of opportunities to carry out this type of attack, given that training data are frequently collected from the Internet. Similarly, evasion attacks that create adversarial samples to target embedded neural networks and other machine learning models could lead to the circumvention of phishing defenses. Consideration of these new attack surfaces and potential adaptive attacks is required to avoid a false sense of security.

Applying machine learning in critical and sensitive applications requires understanding the risks and vulnerabilities of the entire machine learning pipeline. In this issue, we included seven articles that focus on multiple aspects of trustworthy machine learning: machine learning privacy defenses and threats, poisoning attacks against NLP techniques that use large models, and the robustness of machine learning used in security applications, such as source code vulnerability detection and malware and phishing detection. The selected articles provide an overview of cutting-edge attacks and defenses to deal with these threats. Importantly, generating trustworthy machine learning models that have desirable properties is not an easy task. While there is a tendency to consider every trustworthiness objective in isolation, doing so does not result in an acceptable holistic solution. We invite the community to consider multiple dimensions of the machine learning pipeline while designing trustworthy solutions. ■

Applying machine learning in critical and sensitive applications requires understanding the risks and vulnerabilities of the entire machine learning pipeline.

Nathalie Baracaldo is a research staff member and manager of the AI Security and Privacy solutions team with IBM Almaden Research Center, San Jose, California, 95120, USA. Her research interests include information security, privacy, and trust. Baracaldo received a Ph.D. in information science and technology from the University of Pittsburgh. She is a Member of IEEE. Contact her at baracald@us.ibm.com.

Alina Oprea is an associate professor with Northeastern University, Boston, Massachusetts, 02120, USA. Her research interests include machine learning security and privacy, threat detection, cloud security, and applied cryptography. Oprea received a Ph.D. in computer science from Carnegie Mellon University. She is a Member of IEEE and ACM. Contact her at a.oprea@northeastern.edu.