DOI:10.1145/3655635

To view the accompanying paper, visit doi.acm.org/10.1145/3643456

Technical Perspective
Machine Learning in Computer Security Is Difficult to Fix

By Battista Biggio

DURING A 2017 interview, Andrew Ng—one of the most renowned computer scientists in the field of artificial intelligence (AI)—was reported to say: "Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform in the next several years."

Indeed, over the last decade, we have observed a rebirth of interest in AI and, more specifically, in its machine learning (ML) subfield, which is aimed at designing algorithms that learn from examples. This has been fueled by the availability of large volumes of data over the Internet, the increased computing power of today's hardware and cloud infrastructures, and the algorithmic improvements in deep learning and neural networks, which have shown tremendous progress in dealing with text, audio, image, and video data. Their success has been reinforced even more recently with the advent of foundational and generative AI models that can generate realistic text, images, and videos with impressive quality. For these reasons, AI and ML have been fostering important advancements in healthcare, automotive, robotics, recommendation systems, chatbots, and many other applications.

This progress has also left a considerable impact on computer security, inspiring security researchers to adopt such techniques and giving rise to a range of efforts in developing learning-based security systems. Indeed, the goal of security researchers since the early 2000s has been to use ML to improve the performance of traditional security systems in diverse cybersecurity-related tasks, including malware and network intrusion detection, vulnerability identification, and binary code analysis. However, despite its potential, the application of ML in security faces subtle challenges that can compromise its effectiveness, potentially making learning-based systems unsuitable for practical deployment in security tasks.

A substantial difference with respect to other successful applications of AI is that the data collected from security-related tasks evolves at a much more rapid pace, not only due to changes in the monitored systems (for example, operating system updates, deprecated library/function names, and so on) but also to the presence of attackers that constantly try to bypass the protection mechanisms empowered by ML models. This makes collecting vast amounts of *representative* data a challenging—if not impossible—task, as such data tends to become obsolete very fast. These are the main reasons that have hindered the success of ML in cybersecurity and the development of large and effective foundational models in this field.

Some of the issues preventing the development of successful and effective ML models for computer security were initially pointed out in the influential paper, "Outside the Closed World: On Using ML for Network Intrusion Detection," by Robin Sommer and Vern Paxson in the context of intrusion detection systems back in 2010. The accompanying paper, in the same critical and constructive spirit, analyzes more recent work published in top-tier security conferences within the past 10 years. It identifies and details the major pitfalls encountered in more recent studies when designing, implementing, and evaluating learning-based security systems. Just to mention a few, the authors start by highlighting issues related to *data collection*, which may not reflect the actual data distribution observed in deployed systems and include inaccuracies in the data labeling process.

When it comes to system design and evaluation, the authors point out inappropriate procedures for creating training-test splits that may induce biases in the learning process, the use of inappropriate baselines and performance measures, and the lack of consideration of the base-rate fallacy in the presence of very rare malicious events (which typically makes even extremely low false-positive rates unacceptable in many security applications). Regarding system deployment and operation, the authors stress the need to test learning-based systems against adversarial test-time evasion attempts and training-time poisoning attacks, which cannot be clearly encompassed in previously collected data but must be simulated specifically against the developed system. The paper also provides some constructive, practical recommendations to improve the current state of ML in computer security and discusses open problems and directions for future research. It would be nice to see more efforts aimed at developing tools and libraries to help automate the design, debugging, and continuous deployment of ML models. I firmly believe this will help overcome the pitfalls mentioned in the paper, as well as implement the suggested recommendations.

This study is timely and exciting as it clearly points out some common issues hindering the design of ML models for computer security and how to overcome them. This witnesses once again that, despite the impressive performance reported in many of the published papers in this area, the reality is quite different, and applying ML in computer security is much more challenging than it may seem.

Battista Biggio is a professor at the University of Cagliari, Italy, and co-founder of the cybersecurity company Pluribus One.