AI Security and Safety: The PRALab Research Experience

Ambra Demontis[1], Maura Pintor[1], Luca Demetrio[1], Angelo Sotgiu[1], Kathrin Grosse[1], Daniele Angioni[1], Giorgio Piras[1], Battista Biggio[1], Fabio Roli[1,2]

[1]Pattern Recognition and Applications Laboratory (PRALab), Department of Electrical and Electronic Engineering, University of Cagliari, Italy

[2]Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genova

ambra.demontis@unica.it, maura.pintor@unica.it, luca.demetrio93@unica.it, angelo.sotgiu@unica.it, kathrin.grosse@unica.it, daniele.angioni@unica.it, giorgio.piras@unica.it, battista.biggio@unica.it, fabio.roli@unige.it

Abstract

We present here the main research topics and activities on security, safety, and robustness of machine learning models that we have investigated in the last 15 years at the Pattern Recognition and Applications (PRA) Laboratory of the University of Cagliari. Our findings have significantly contributed not only to identify and characterize vulnerability of such models to adversarial examples and training data poisoning attacks in the context of real-world applications, but also to the development of more trustworthy artificial intelligence and machine learning algorithms.

1 Research Group

The Pattern Recognition and Applications (PRA) Laboratory was founded in 1996. The PRALab has been active for more than 20 years at the University of Cagliari. Its mission is to address fundamental issues for the development of future pattern recognition systems in the context of real applications, focused on creating secure systems for security applications, as reflected by our motto:

> *there is nothing more practical than a good theory*, by Kurt Lewin.

Our activities can be categorized into four highly-interdependent lines:(i) development of theories to solve problems of fundamental research, including multiple classifier systems (our original expertise) and adversarial machine learning; (ii) application of these theories to solve practical problems, in the research domains of computer vision for video surveillance and ambient intelligence, computer security, biometrics, document and multimedia categorization; (iii) testing and validation of the proposed solutions on real-world data (in-vivo experiments); and (iv) development of prototypes and demonstrators, through which the results of basic research are translated into functional products.

In 2015, some members of the PRA Lab decided to found the company Pluribus One (https://www.pluribus-one.it/), as a spinoff of the research laboratory. Pluribus One is now a well-developed, research-intensive company that designs innovative products and services based on AI technologies for data-driven business and cybersecurity applications, including protection of web services and endpoint devices.

The PRALab team working on AI/ML security, safety, and robustness consists of nine people, including the lab director (Prof. Fabio Roli), two assistant professors (Dr. Battista Biggio and Dr. Ambra Demontis), three post-doctoral researchers, and three more Ph.D. students. Our team has provided pioneering contributions in the area of AI/ML security, being the first to demonstrate gradient-based evasion [1] (also known as *adversarial examples*) and poisoning attacks [4], and how to mitigate them, playing a leading role in the establishment and advancement of this research field [6].

2 Research Topics

Recent progress in AI/ML technologies has reported impressive performances in different tasks. This advancement has paved the way for their use in mission-critical applications like autonomous driving and cybersecurity. Unfortunately, it has been shown that AI/ML technologies are vulnerable to well-crafted attacks performed by skilled attackers. Understanding the security properties of learning algorithms and designing suitable countermeasures against attacks has thus become a timely, relevant, and challenging research field. Our mission is to develop AI technologies that are the strongest link in the security chain and not the weakest one.

2.1 Attacks on Machine Learning

The key to understanding the security properties of AI/ML technologies is to model the threats that can be perpetrated against them, simulate the corresponding attacks, and assess the security properties of the system in these scenarios [6]. The threat model should be devised ad-hoc for the system under design and is application-specific. Because deepening on the considered application, the attacker can face different costs and have diverse constraints when manipulating the system or the data. For example, as shown in Figure 1, attacks on AI/ML models can be staged both at test and at training time. In our research, we start from a general categorization of attacks on AI/ML models, which is detailed in [6] and compactly reported in Figure 2, and use it to envision and design potential attacks on various application domains, including robot vision [20], malware detec-

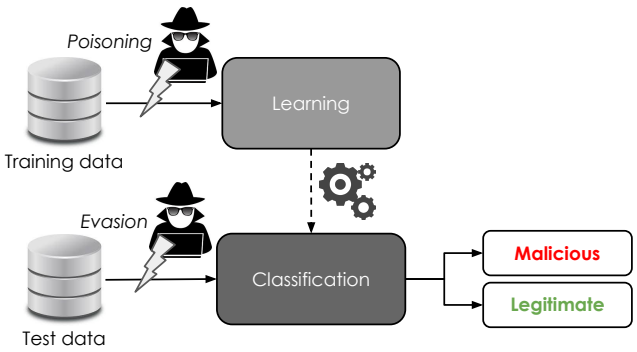

Figure 1: Conceptual representation of test time (evasion) and training time (poisoning) attacks. Evasion attacks slightly manipulate the system inputs to enforce their misclassification. Poisoning attacks manipulate the training data to cause errors at test time.

tion in Windows programs [16] and Android applications [13; 14], and biometric recognition [2; 3].

Evasion Attacks. In this scenario, the attackers modify the samples to have them misclassified. For example, they slightly manipulate malware to have them misclassified as legitimate applications. Our researchers were the first to show that some popular classification algorithms like Support Vector Machines and Neural Networks [1] and feature selection algorithms are vulnerable to this attack [28]. Simulating an evasion attack can present different challenges depending on the considered technology and application. One of these challenges is to construct malware able to evade detection. It is challenging because by modifying the malware, the attacker might compromise their malicious functionalities. Our researchers developed evasion algorithms able to construct malware that evade the target system while preserving all their functionalities [13; 17; 10; 11]. Another challenge is that evasion attacks might be computationally costly depending on the considered scenario. In [24] our researchers have shown that their computational time can be reduced while obtaining comparable performance. Attackers often do not know all the details regarding the target systems. However, our researchers have shown that effective attacks can also be developed in this challenging scenario [14; 10; 11]. Notably, these attacks have shown to be effective also on anti-virus solutions hosted at VirusTotal.

Poisoning Attacks. AI/ML technologies are often re-trained during the operative phase to consider changes in the data distribution. In this scenario, an attacker can compromise the training data injecting samples specifically devised to compromise the learning process, making the classifier unable to classify samples at test time correctly. Our researchers have been the first to show that the Support Vector Machines [4], as well as neural networks [21], feature selection methods [27] and clustering algorithms [5] can be compromised by this attack. The main challenge regarding poisoning attacks is that computing them requires solving a computationally costly problem. Nevertheless, our researchers have shown that it is possible to find approximate solutions with a comparable effectiveness [21; 8].

Other Attacks. We have been investigating also other attacks on AI/ML models, including backdoor poisoning [7] and reprogramming [29].

2.2 Evaluating Adversarial Robustness

One of the main open problems regarding the security of AI/ML technology is how to evaluate their security efficiently. Different methods that should be able to formally verify the robustness of these technologies have been proposed. However, unfortunately, they can be applied only to a pretty limited set of AI/ML technologies. These technologies are thus usually evaluated empirically, simulating an attacker's presence, as explained before. However, these empirical evaluations are often not correctly conducted [6] and this leads to overestimating the robustness of the considered system: the system seems robust only because the attack fails. Our researchers are developing methodologies and debugging tools that can be used to improve current approaches for empirical security evaluations, with the goal of making them more reliable (e.g., by properly tuning the attack hyperparameters, identify the presence of gradient obfuscation hindering the attack optimization process, etc.) [23; 22].

2.3 Improving Adversarial Robustness

We have followed different lines of research to improve the robustness of AI/ML technologies against both evasion and poisoning attacks. The first consists of increasing the margin in input space, which can be done with different techniques, including adversarial training [25; 18], and regularization [15; 12; 13; 14]. And the second consists of detecting the input samples modified by the attacker [20; 26; 9; 18; 19].

3 Projects

Our research activities are carried out in the framework of regional, national, and European projects funded by public as well as private initiatives. We had more than twenty-five projects founded between 2012 and 2020. The full list is available at http://pralab.diee.unica.it/en/Projects. Six of them were founded by the European Commission, and two of them were coordinated by the PRALab. Overall, we received 3 million euros of funding, whose the European Commission provided half. The annual turnover is around four hundred thousand euros.

We have different ongoing projects on AI security:

1. 2020-2022 - PRIN 2017 RexLearn: "Reliable and Explainable Machine Learning," funded by the Italian Ministry of Education, University and Research (grant no.2017TWNMH2). This project aims to develop novel learning paradigms, able to take reliable and explainable decisions, and to assess and mitigate the security risks associated with potential misuses of machine learning.
2. 2020-2023 - FFG COMET Module S3AI: "Security and Safety for Shared Artificial Intelligence," funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by the Austrian Research Promotion Agency FFG. This project aims to provide the foundations required to build secure and safe shared artificial intelligence systems.

Attacker's Capability	Attacker's Goal: Misclassifications that do not compromise normal system operation — Integrity	Misclassifications that compromise normal system operation — Availability	Querying strategies that reveal confidential information on the learning model or its users — Privacy / Confidentiality
Test data	Evasion (a.k.a. adversarial examples)	Sponge attacks	Model extraction / stealing Model inversion (hill climbing) Membership inference
Training data	Backdoor poisoning (to allow subsequent intrusions) – e.g., backdoors or neural trojans	DoS poisoning (to maximize classification error)	-

Figure 2: Attacks on machine learning models, as categorized in [6].

Some other relevant projects are listed in the following:

- 2017-2019 - Research and Innovation Action LETS-CROWD: "Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings". Call: H2020 - SEC-07-FCT-2016-2017. Grant Agreement H2020/N.740466.
- 2015-2018 – Innovation Action DOGANA: "aDvanced sOcial enGineering And vulNerability Assessment Framework". Call: H2020 – DS 2014-1. Grant Agreement H2020/N.653618.
- 2014-2016 – CSA CyberROAD: "Development of the Cybercrime and Cyberterrorism Research Roadmap". Call: FP7 – SEC 2013.2.5-1. Grant Agreement FP7-SEC-2013/N.607642.
- 2014-2016 - ILLBuster, "Buster of ILLegal Contents spread by malicious computer networks". DGHOME - ISEC, Prevention of and Fight Against Crime. Grant Agreement: HOME/2012/ISEC/AG/4000004360.
- 2013-2015 - MAVEN: "Management and Authenticity Verification of Multimedia Contents". Call: FP7-SME-2013-1. Grant Agreement FP7-SME-2013-1/N.606058.

4 Developed Tools

As explained in the previous section, correctly evaluating the robustness of AI/ML technologies might be challenging. Our researchers have developed different tools that help to perform security evaluation[1]. These tools include SecML, a Python library that allows assessing the security evaluation of AI/ML technologies against evasion and poisoning attacks, and an extension of this library, called SecML Malware ad-hoc for Windows malware. For each of them, they have released a tool that allows running security the evaluations through a graphical interface: PandaVision, and ToucanStrike. Furthermore, our researchers have released a tool that allows evaluating is an attack is or not effective in the considered scenario[2].

[1]https://github.com/pralab

[2]https://github.com/pralab/IndicatorsOfAttackFailure

5 Challenges and Perspectives

While research is quickly progressing in AI/ML Security, companies are working on automating the development and operations of ML models (MLOps), without focusing too much on ML security-related issues. In this respect, a relevant challenge for the future will be to extend the current MLOps paradigm to also encompass ML security (towards implementing what we refer to as ML*Sec*Ops). To this end, we plan to incorporate research on security testing, protection and monitoring of AI/ML models into the MLOps development cycle. In particular, we plan to extend our research towards: (i) developing and improving attacks (including evasion, poisoning and privacy threats) for making security testing and validation of AI/ML models more efficient and available for a wider set of application domains; (ii) designing improved defenses with robustness guarantees to protect AI/ML models not only against such attacks but also to enable reliable classification when out-of-distribution data is provided as input; and (iii) designing methods that allow for constantly monitoring if a deployed model is attacked during operation, enabling prompt reaction when needed. We firmly believe that integrating these dimensions into an MLSecOps cycle will definitely help software engineers and developers to seamlessly deploy and maintain more secure, reliable, and trustworthy AI/ML models in practice.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *ECML PKDD, Part III*, volume 8190 of *LNCS*, pages 387–402, 2013.

[2] B. Biggio, L. Didaci, G. Fumera, and F. Roli. Poisoning attacks to compromise face templates. In *6th IAPR Int'l Conf. on Biometrics (ICB 2013)*, pages 1–7, Madrid, Spain, 2013.

[3] B. Biggio, G. Fumera, P. Russu, L. Didaci, and F. Roli. Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective. *Signal Processing Magazine, IEEE*, 32(5):31–41, Sept 2015.

[4] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *29th ICML*, pages 1807–1814, 2012.

[5] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, and F. Roli. Is data clustering in adversarial settings secure? In *Proc. of the 2013 AISec*, AISec '13, pages 87–98, New York, NY, USA, 2013.

[6] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Patt. Rec.*, 84:317–331, 2018.

[7] A. E. Cinà, K. Grosse, S. Vascon, A. Demontis, B. Biggio, F. Roli, and M. Pelillo. Backdoor learning curves: Explaining backdoor poisoning beyond influence functions, 2021.

[8] A. E. Cinà, S. Vascon, A. Demontis, B. Biggio, F. Roli, and M. Pelillo. The Hammer and the Nut: Is Bilevel Optimization Really Needed to Poison Linear Classifiers? In *IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE, 2021.

[9] F. Crecchi, M. Melis, A. Sotgiu, D. Bacciu, and B. Biggio. FADER: Fast adversarial example rejection. *Neurocomputing*, 470:257–268, Jan. 2022.

[10] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, and A. Armando. Functionality-preserving black-box optimization of adversarial windows malware. *IEEE Transactions on Information Forensics and Security*, 16:3469–3478, 2021.

[11] L. Demetrio, S. E. Coull, B. Biggio, G. Lagorio, A. Armando, and F. Roli. Adversarial EXEmples: A survey and experimental evaluation of practical attacks on machine learning for windows malware detection. *ACM Trans. Priv. Secur.*, 24(4), September 2021.

[12] A. Demontis, B. Biggio, G. Fumera, G. Giacinto, and F. Roli. Infinity-norm support vector machines against adversarial label contamination. In A. Armando, R. Baldoni, and R. Focardi, editors, *ITASEC*, number 1816 in CEUR Workshop Proc., pages 106–115, Aachen, 2017.

[13] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto, and F. Roli. Yes, machine learning can be more secure! a case study on android malware detection. *IEEE Trans. Dependable and Secure Computing*, 2019.

[14] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *USENIX Security*. USENIX Association, 2019.

[15] A. Demontis, P. Russu, B. Biggio, G. Fumera, and F. Roli. On security and sparsity of linear classifiers for adversarial settings. In *Joint IAPR Int'l Works. on Struct., Synt., and Stat. Patt. Rec.*, volume 10029, pages 322–332, Cham, 2016.

[16] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, and F. Roli. Adversarial malware binaries: Evading deep learning for malware detection in executables. In *2018 26th EUSIPCO*, pages 533–537, Rome, 09/2018 2018. IEEE, IEEE.

[17] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, and F. Roli. Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables. *ArXiv*, Mar. 2018.

[18] D. Maiorca, A. Demontis, B. Biggio, F. Roli, and G. Giacinto. Adversarial Detection of Flash Malware: Limitations and Open Issues. *Computers & Security*, 96, May 2020.

[19] S. Melacci, G. Ciravegna, A. Sotgiu, A. Demontis, B. Biggio, M. Gori, and F. Roli. Domain Knowledge Alleviates Adversarial Attacks in Multi-Label Classifiers. *IEEE Trans. on Patt. Analysis and Machine Intelligence*, pages 1–1, 2021.

[20] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, and F. Roli. Is deep learning safe for robot vision? adversarial examples against the icub humanoid. In *2017 VIPAR*, pages 751–759, Oct 2017.

[21] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proc. of the 10th ACM Works. AISec@CCS 2017*, pages 27–38, 2017.

[22] M. Pintor, L. Demetrio, G. Manca, B. Biggio, and F. Roli. Slope: A First-order Approach for Measuring Gradient Obfuscation. In *Proc. of the ESANN, ESANN 2021*, 2021.

[23] M. Pintor, L. Demetrio, A. Sotgiu, G. Manca, A. Demontis, N. Carlini, B. Biggio, and F. Roli. Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples. *ArXiv*, June 2021.

[24] M. Pintor, F. Roli, W. Brendel, and B. Biggio. Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[25] S. Rota Bulò, B. Biggio, I. Pillai, M. Pelillo, and F. Roli. Randomized prediction games for adversarial machine learning. *IEEE Trans. on Neural Networks and Learning Systems*, 28(11):2466–2478, 2017.

[26] A. Sotgiu, A. Demontis, M. Melis, B. Biggio, G. Fumera, X. Feng, and F. Roli. Deep neural rejection against adversarial examples. *EURASIP Journal on Information Security*, 2020(1):5, Apr. 2020.

[27] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli. Is feature selection secure against training data poisoning? In *ICML*, pages 1689–1698, 2015.

[28] F. Zhang, P. Chan, B. Biggio, D. Yeung, and F. Roli. Adversarial feature selection against evasion attacks. *IEEE Trans. on Cybernetics*, 46(3):766–777, 2016.

[29] Y. Zheng, X. Feng, Z. Xia, X. Jiang, A. Demontis, M. Pintor, B. Biggio, and F. Roli. Why adversarial reprogramming works, when it fails, and how to tell the difference, 2021.