ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Towards Machine Learning Models that We Can Trust: Testing, Improving, and Explaining Robustness

Maura Pintor[1] and Ambra Demontis[1] and Battista Biggio[1]

1- University of Cagliari - Dept of Electrical and Electronic Engineering
Cagliari - Italy

Abstract. In recent years, machine learning has become the most effective way to analyze massive data streams. However, machine learning is also subject to security and reliability issues. These aspects require machine learning to be thoroughly tested before being deployed in unsupervised scenarios, such as services intended for consumers. The goal of this session is to discuss open challenges, both theoretical and practical, related to the security and safety of machine learning. The session will try to address the following challenges: (i) the implementation of efficient tests for Machine Learning in the context of robustness to attacks and natural drifts of data; and (ii) the design of robust and efficient models able to function in the wild and mitigate or detect adversarial attacks.

1 Context

Machine learning (ML) has rapidly transformed various industries, from powering recommendation systems to driving autonomous vehicles. As ML adoption grows, so does the need for rigorous evaluation and trustworthiness of ML models. In the last years, these models have been rapidly increasing in size and complexity, as well as the amount of data used for their training. This demands for testing techniques able to effectively cover the attack surface of these systems and properly test the resilience of ML to unseen and undesirable attacks [1]. To address this open problem, we formulate the following research challenge:

Research Challenge 1 *Implementation of efficient tests for Machine Learning in the context of robustness to attacks and natural drifts of data.*

In addition, it is crucial that ML algorithms consider both high technical and functional standards but also additional trustworthy requirements. These models have to be trustworthy in terms of attacks that can come from the misuse of the models and in the meanwhile take into account societal and individual values and principles that significantly impact people's lives, such as ethical concerns. This mandates the design of *ML algorithms that we can trust*, leading to our second research challenge:

Research Challenge 2 *Design of robust and efficient models able to function in the wild and mitigate or detect adversarial attacks.*

Both of these challenges need to be considered from different perspectives in order to improve the current state of ML trustworthiness, in particular when it is used in risk-related scenarios. In fact, in the coming years, there will be an arms race to verify these systems' proper functioning and regulation. Already, we are beginning to see the first effects of these trends with the appearance of the AI Act and several attempts to legislate the interaction between AI and its users. To this end, trustworthy ML requirements are fundamental criteria that ensure the reliability and responsible deployment of machine learning systems. First, *accuracy* refers to the ability of an ML model to make precise predictions and minimize errors, instilling confidence in its outputs. However, accuracy only ensured the quality of the outputs in controlled testing scenarios, where data is not affected by noise or perturbations injected by attackers. *Robustness* ensures the model's resilience to both adversarial attacks and unexpected inputs, strengthening its performance in real-world scenarios. Adversarial robustness ensures the model is controlled even in the presence of worst-case perturbations [2, 3]. Beyond these attacks, ML models should also be robust to natural drift of data, i.e., to the gradual changes in the input data distribution over time that is common in dynamic real-world environments. *Fairness* emphasizes the equitable treatment of different groups and demographics, guarding against biased outcomes and social discrimination [4]. Lastly, *privacy* safeguards the sensitive data used in ML training [5], preventing unauthorized access and protecting individuals' confidential information as well as the provider's ownership of the ML models [6]. Embracing these trustworthy ML requirements is key to building ethical, accountable, and dependable AI systems that benefit society and foster public trust in machine learning technology.

Finally, as all these topics are investigated in research, we remark on the importance of applying these advancements to real-world use cases, such as medical applications, where ML reliability and safety are paramount.

2 State of the Art

In this section, we review the state-of-the-art progress in ensuring the trustworthiness and robustness of ML systems, focusing on the most relevant key requirements just described and highlighting gaps and limitations that still remain to be addressed in this field.

Robustness and Uncertainty Estimation Testing the robustness of ML models against attackers is useful for knowing the models' limitations in advance and apply mitigations, such as specific training techniques to make models learn robust features. Robustness testing involves evaluating how well models perform under challenging conditions caused by well-crafted adversarial attacks [1, 2, 3]. Several algorithms have been proposed to craft adversarial attacks [7, 8, 9], primarily leveraging gradient descent to optimize over specific loss functions. However, proper tuning of these algorithms is not easy and has led to sub-optimal and over-optimistic robustness evaluations that have been later shown to be broken [10, 11]. This is due to the wide search space of the hyperparameters that influence the descent. Properly configuring these attacks remains a challenging open problem. Additionally, uncertainty estimation is an important aspect in both robustness to adversarial attacks and natural drifts of data [12]. Techniques such as Monte Carlo Dropout, Bayesian neural networks, and ensemble methods provide valuable insights into the model's confidence in its predictions. These techniques help models deal with out-of-distribution data, make informed decisions under uncertainty, and improve safety in critical applications. Uncertainty estimation may be useful in detecting adversarial examples [13], adapting models to changing environments [14], and facilitating robust model selection through ensemble techniques [15].

Data and Model Preservation Preserving user privacy and sensitive data in ML systems has emerged as a critical concern [16]. The main concern by far is not giving access to those not authorized to read the data in the ML pipeline, preventing unauthorized access and data breaches. In the general case of ML, data is commonly intended as the information used to train the model. Privacy preservation techniques, like differential privacy [17], aim to protect sensitive individual data by adding noise to the training process, ensuring that the model does not memorize specific data points. Federated learning [5], on the other hand, allows multiple parties to train a global model while keeping their data decentralized collaboratively, thus avoiding the need to share raw data and promoting privacy. However, in this context, what we call data also includes the model itself, whose internals should be protected to avoid financial loss (as designing and training the model has a non-negligible cost) and information spilling due to potential misuse of the stolen models. ML is indeed not immune to model-stealing attacks [6], where adversaries attempt to reverse-engineer and replicate trained models by extracting information through targeted queries. To mitigate this risk, researchers are exploring novel defenses to enhance the robustness of ML models against such attacks, as well as efficient attacks to test these defenses proactively.

3 The contributions of the ESANN special session

A total of five contributions were accepted in the special session. The contributions tackle both RC 1 and RC 2.

Testing ML models (RC 1). Developing efficient and reliable tests for robustness to attacks and natural drifts of data is essential to ensure the reliability and resilience of machine learning models. Adversarial attacks and changes in data distribution can lead to degraded performance and potential vulnerabilities. Addressing this challenge involves advancing techniques for adversarial testing, stress testing, and domain adaptation evaluation. Efficient testing procedures will enable practitioners to assess model performance under different scenarios, making models better equipped to handle real-world challenges. The following contributions were proposed to enhance the field of efficient testing for the security of ML:

- In *Improving Fast Minimum-Norm Attacks with Hyperparameter Optimization* [18], the authors present a framework to automate the search of a good attack configuration over a defined space of hyperparameter choices. Specifically, they search through meta-optimization for the best-performing loss function, optimizer, and step-size scheduler to launch an effective attack. This raises the level of automation at which engineers can work, enabling more effective robustness evaluations of ML models.

- In *On the Limitations of Model Stealing with Uncertainty Quantification Models* [19], the authors use uncertainty quantification within the setting of model stealing attacks, trying to improve the attack effectiveness by generating multiple possible networks and combining them to improve the quality of the stolen model. The authors find that the considered models only lead to marginal improvements in terms of fidelity to the stolen model.

Designing robust models (RC 2). Designing models that can function effectively in the wild and detect/mitigate adversarial attacks is a fundamental challenge. Researchers are exploring various avenues, including robust training, attack detection, and security protocols. These approaches reinforce models against adversarial manipulations or privacy attacks and improve their generalization capabilities. Additionally, advancements in uncertainty estimation contribute to the creation of more robust and reliable models capable of making well-calibrated predictions and detecting uncertain inputs. The following contributions were proposed to improve the design of ML algorithms toward secure-by-design and resilient models:

- In *Towards Randomized Algorithms and Models that We Can Trust: a Theoretical Perspective* [20], the authors propose a formal framework that incorporates both functional (accuracy, non-regressivity) and ethical (fairness, explainability) properties. Using their framework, the authors focus on Randomized Models and Randomized Algorithms to optimize these metrics jointly when training the models.

- In *Secure Federated Learning with Kernel Affine Hull Machines* [21], the authors propose an accurate and computationally-efficient federated learning architecture to achieve privacy preservation of the involved parties. They propose to enhance security in federated learning through a global classifier that handles local predictions from multiple parties (local classifiers) based on kernel-based affine machines.

- In *Single-pass uncertainty estimation with layer ensembling for regression: application to proton therapy dose prediction for head and neck cancer* [22], the authors present an efficient uncertainty quantification method for machine learning and use it for a medical use case. The method is based on Layer Ensembles and is competitive with state-of-the-art methods such as Monte Carlo Dropout while being much faster. Although this topic is not strictly related to machine learning robustness, it is of particular interest for favoring the trustworthiness of these systems, especially when used to take risk-related decisions such as in the medical domain.

4 Conclusions

By tackling these challenges, the machine learning community can significantly improve the trustworthiness of ML models and systems. Efficient testing for robustness ensures that models perform well under diverse conditions, while the design of robust and efficient models equips AI systems to withstand adversarial attacks and maintain reliable performance in dynamic environments. Emphasizing these advancements fosters the development of responsible and dependable AI technologies that positively impact various domains and favors greater confidence in machine learning applications.

Acknowledgements

This work has been supported by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG in the COMET Module S3AI; by Fondazione di Sardegna under the project "TrustML: Towards Machine Learning that Humans Can Trust", with CUP: F73C22001320007; and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

References

[1] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit.*, 84:317–331, 2018.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013,*, 2013.

[3] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[4] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy AI: from principles to practices. *ACM Comput. Surv.*, 55(9):177:1–177:46, 2023.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, 2017.

[6] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In Srdjan Capkun and Franziska Roesner, editors, *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pages 1345–1362. USENIX Association, 2020.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[8] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[9] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20052–20062, 2021.

[10] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019.

[11] Maura Pintor, Luca Demetrio, Angelo Sotgiu, Ambra Demontis, Nicholas Carlini, Battista Biggio, and Fabio Roli. Indicators of attack failure: Debugging and improving optimization of adversarial examples. In *NeurIPS*, 2022.

[12] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *CoRR*, abs/2107.03342, 2021.

[13] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7167–7177, 2018.

[14] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[15] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[16] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[17] Zhanglong Ji, Zachary Chase Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. *CoRR*, abs/1412.7584, 2014.

[18] Giorgio Piras, Giuseppe Floris, Raffaele Mura, Luca Scionis, Maura Pintor, Battista Biggio, and Ambra Demontis. Improving fast minimum-norm attacks with hyperparameter optimization. In *31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2023, Bruges, Belgium, October 4-6, 2023*, 2023.

[19] David Pape, Sina Däubener, Thosten Eisenhofer, Antonio Emanuele Cinà, and Lea Schönherr. On the limitations of model stealing with uncertainty quantification models. In *31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2023, Bruges, Belgium, October 4-6, 2023*, 2023.

[20] Luca Oneto, Sandro Ridella, and Davide Anguita. Towards randomized algorithms and models that we can trust: a theoretical perspective. In *31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2023, Bruges, Belgium, October 4-6, 2023*, 2023.

[21] Mohit Kumar, Bernhard Moser, and Lukas Fischer. Secure federated learning with kernel affine hull machines. In *31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2023, Bruges, Belgium, October 4-6, 2023*, 2023.

[22] Ana Maria Barragan Montero, Robin Tilman, Margerie Huet-Dastarac, and John Lee. Single-pass uncertainty estimation with layer ensembling for regression: application to proton therapy dose prediction for head and neck cancer. In *31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2023, Bruges, Belgium, October 4-6, 2023*, 2023.