Research on Trustworthy and Secure AI at the RETIS Lab, SSSUP

Giulio Rossolini[1], Wesam Abbasi[1], Federico Nesti[1], Andrea Saracino[1], Alessandro Biondi[1] and Giorgio Buttazzo[1]

[1]Departement of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà, 33, 56127 Pisa PI

Abstract

This paper presents an overview of the research conducted at the RETIS Laboratory on trustworthy and secure artificial intelligence, with a focus on safety-critical and cyber-physical systems. As artificial intelligence becomes increasingly integrated into high-stakes domains such as autonomous vehicles, healthcare, and industrial automation, ensuring robustness, security, interpretability, and safety is essential. This paper summarizes recent efforts in developing benchmarks, conducting analytical studies, and proposing novel techniques to enhance the reliability of deep neural networks under real-world conditions. The paper also highlights interdisciplinary collaborations and contributions across national and international projects, reflecting the lab's commitment to enabling trustworthy and secure AI solutions in complex and critical environments.

Keywords

Safe and Secure AI, Adversarial Robustness AI, Privacy Preserving AI, Trustworthy AI, AI-enabled Secure Systems

1. Introduction and motivations

Artificial intelligence (AI) is rapidly becoming a foundational component in cyber-physical and safety-critical systems, including autonomous vehicles, medical devices, industrial automation, and smart infrastructures. These domains require not only high performance but also strong assurances of reliability, safety, and security. However, despite the remarkable capabilities of modern AI models, their deployment in real-world, high-stakes applications is still limited by several fundamental challenges.

First, safety and robustness remain major concerns: AI models often exhibit limited resilience to uncertainty and corrupted inputs. Second, interpretability is essential for adopting AI in critical domains, as stakeholders must understand and trust the model's responses. Third, there are significant security challenges, including vulnerabilities to evasion attacks (e.g., adversarial examples) and risks of privacy leakage, stemming from biases or unintended memorization of sensitive information. Finally, there are system-level security and reliability issues that extend beyond the AI model itself. These include hardware-software dependencies and the lack of predictability within both functional and operational boundaries of the systems where AI is deployed.

The Real-Time Systems Laboratory (RETIS Lab) of the Scuola Superiore Sant'Anna of Pisa focuses on addressing these challenges by first developing comprehensive benchmarks and conducting in-depth analyses to better understand the complex behavior of deep neural networks (DNNs). In parallel, new techniques and methodologies are developed to ensure that AI systems are trustworthy and robust under real-world conditions.

In summary, the main mission of the RETIS Lab is to enable the reliable integration of AI algorithms into safety-critical systems, ensuring:

- Robustness against corruptions or unexpected objects that may arise at test time, from both safety and security perspectives;

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy.
g.rossolini@santannapisa.it (G. Rossolini); w.alabbasi@santannapisa.it (W. Abbasi); f.nesti@santannapisa.it (F. Nesti); a.saracino@santannapisa.it (A. Saracino); a.biondi@santannapisa.it (A. Biondi); g.buttazzo@santannapisa.it (G. Buttazzo)
0000-0002-6404-2627 (G. Rossolini); 0000-0002-6901-1838 (W. Abbasi); 0000-0003-4338-9573 (F. Nesti); 0000-0001-8149-9322 (A. Saracino); 0000-0002-6625-9336 (A. Biondi); 0000-0003-4959-4017 (G. Buttazzo)

- Security, with reinforced capabilities to preserve properties such as privacy and protection against model stealing, while also accounting trade-off with respect to other properties such as explainability and accuracy;
- Interpretability and predictability across the entire deployment stack, offering transparency in decision-making processes and predictability in computational cost, taking into account system constraints and application domain criticality;
- Safety in their interactions with both the physical environment and the host systems, particularly in contexts where failures could result in significant harm.

The following sections present various topics related to trustworthy and secure AI investigated at the RETIS Lab, highlighting the key results achieved in the related research areas. The addressed topics are inherently interdisciplinary, encompassing robust deep learning, computer vision, synthetic data generation for testing, system-level security for trustworthy AI, and comprehensive safety and security analysis. The conducted research is carried out under several national and international projects and spans across a wide range of application domains, including medicine, autonomous driving, robotics, railway systems, and cyber-security.

2. Robustness evaluation of DNNs

DNNs in safety-critical applications remains fragile when faced with unpredictable conditions or malicious inputs [1]. In areas like autonomous driving, medicine, and robotics, even small changes in the input can lead to unsafe decisions [1, 2]. For this reason, robustness and awareness of potential threats are essential. The RETIS Lab addresses this challenge by thoroughly testing AI models in complex scenarios, studying new types of adversarial attacks, and developing benchmarks and metrics that help better assess their performance in real-world conditions.

Real-world adversarial attacks. Considering the importance of outdoor environments in many robustness evaluations, physical adversarial attacks pose a significant challenge for developing secure and reliable AI systems. Unlike digital perturbations, these attacks involve real-world artifacts, such as adversarial billboards, modified road signs, or altered clothing, that can be physically deployed to consistently trigger misclassifications under a variety of conditions.

The RETIS Lab has made key contributions in this area by introducing novel attacks specifically designed for dense prediction tasks such as semantic segmentation and object detection [3, 4], which are essential for evaluating the robustness of AI models against evasion attacks in applications like autonomous vehicles, robotics, and surveillance systems. In particular, by focusing on outdoor scenarios, we have studied physically realizable attacks that remain effective under real-world transformations, including variations in lighting, viewpoint, distance, and occlusion [3, 4]. These efforts reveal fundamental vulnerabilities in current AI models.

Particular attention has been devoted to assessing the real-world effectiveness of the tested attacks by placing realistic objects in physical scenarios. An example of one such proposed attack, implemented as an adversarial billboard, is shown in Figure 1.

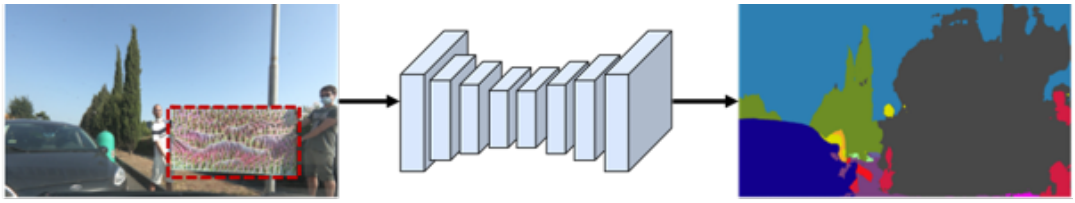

Figure 1: Input image with a physical adversarial billboard (red rectangle) and the corresponding semantic segmentation produced by ICNet. The grey area indicates a severe misclassification caused by the attack.

Metrics and benchmarks for AI-based vision. This research focuses on evaluating the robustness of AI models through the development of testing strategies and robust benchmarks, specifically aimed at assessing the reliability of dense prediction tasks in complex environments such as driving scenarios. Although it is well known that AI models for these tasks are vulnerable to natural corruptions (e.g., noise, blur, occlusion) [5] and adversarial manipulations, a deeper analysis and metrics to understand why these failures occur were missing. Hence, appropriate evaluation methods and metrics were developed to enhance interpretability, enable meaningful comparisons, and guide effective model ensembling [6].

In recent studies, the spatial robustness of AI models was investigated in more detail with proper benchmarks. It refers to the susceptibility of AI models to misclassifications caused by altered or adversarial content placed in different regions of the image, rather than directly on the target area that could be fooled [6]. This analysis demonstrates that models can be fooled by perturbations located far from the area of interest, revealing a critical blind spot in many current vision systems.

Use of simulators to test AI algorithms Evaluating AI robustness in real-world settings, especially in safety-critical domains like autonomous driving, can be costly, time-consuming, and risky. The variability of real-world environments also makes it difficult to reproduce conditions for debugging and model improvement. To address these challenges, the RETIS Lab employs high-fidelity, photo-realistic simulators as a central tool for robustness and critical evaluation across several domains, including autonomous driving [7], robot navigation [8], and railway scenarios [9], each requiring specific environmental setups and safety requirements. The use of simulators, in all of these domains, enable both model-level and system-level analysis under failure-inducing conditions.

A key contribution was the development of CARLA-GeAR[1] [7], an evaluation and benchmarking tool that extends the CARLA simulator to support adversarial robustness testing in self-driving systems. CARLA-GeAR enables the automated generation of adversarial scenarios by placing adversarial billboards across various photorealistic driving environments. Each scenario is provided with ground truth labels for multiple tasks, like semantic segmentation, object detection, and depth estimation, allowing comprehensive benchmarking of model robustness. The CARLA-GeAR tool has been used not only to evaluate the robustness of vision models but also as a benchmark for testing defense mechanisms applied on top of these models. The results revealed that many of these defenses still face issues and open challenges, especially in complex scenarios such as driving.

3. Reliable defenses for robust AI

In addition to conducting rigorous evaluations and benchmarks of robustness against evasion attacks, effective defense mechanisms have been developed to operate reliably both under digital and physical environments. This was achieved either by incorporating dedicated defense mechanisms into the perception system or by extending the AI model with architectural features that inherently guarantee robustness against specific threats.

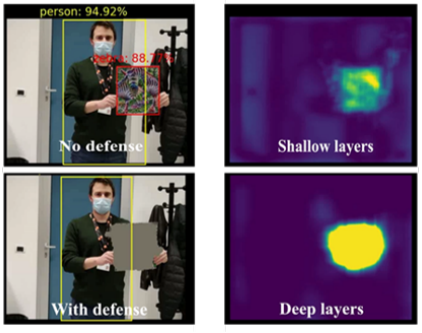

Figure 2: Defense mechanism in [10].

Robust-by-design DNNs against perturbations Achieving verification and certification of AI model predictions against potential perturbations (particularly those defined within a geometric ϵ-sphere) is a critical requirement for deploying AI in safety-critical systems. In this context, providing formal guarantees of robustness ensures that model predictions remain stable under bounded input variations, thereby increasing trust in its behavior under uncertainty. However,

[1] https://carlagear.retis.santannapisa.it/

existing formal verification techniques often suffer from high computational costs, limiting their applicability in real-time settings. This challenge has motivated the investigation of *robust-by-design* AI architectures that are inherently more reliable and provide an online verification [11, 12, 13].

For instance, Brau et al. [11] investigated a model that allows estimating, during inference, the minimal adversarial perturbation required to change the predicted class of a given input. By continuously monitoring this margin, the system can proactively identify inputs that lie too close to the decision boundary and hence could be potentially vulnerable to small perturbations. This enables early detection of adversarial inputs and supports the activation of preventive measures, such as rejecting uncertain predictions or triggering fallback mechanisms, before a misclassification occurs.

Efficient defenses against physical attacks In this context, the RETIS Lab developed multiple defense strategies [4, 14, 10] based on the observation that physical adversarial attacks tend to overactivate neurons in the internal layers of a neural network [10]. Leveraging this fact, the proposed methods use statistical analysis of the model's internal state to detect at runtime potential attacks and malicious patterns. An illustration of the proposed mechanism is shown in Figure 2.

Given the importance of integrating such defense mechanisms into real-world systems, the proposed solutions are not only effective to detect and mask an attack with high accuracy, but are designed to operate under strict real-time and resource constraints to be seamless integrated into an embedded system. Furthermore, unlike many existing defense approaches, which are computationally intensive or are evaluated only in simplified settings (e.g., without real-world environmental variability), the proposed methods achieve state-of-the-art performance in terms of both real-world robustness and computational efficiency.

4. Architectures for reliable AI

Complex autonomous systems typically include software components with mixed criticality and diverse operating system requirements. High-level tasks like sensor processing, communication, and machine learning rely on rich operating systems like Linux, while low-level control and safety-critical functions require a real-time operating system (RTOS) to ensure predictable response times. Running these components on the same platform can cause interference due to conflicts in accessing shared resources, leading to unpredictable delays and degraded performance. Without proper isolation, security breaches in non-critical modules can also compromise critical functions.

In this context, to safely and predictably support AI-powered real-time cyber-physical systems, a bare-metal (Type 1) hypervisor (i.e., installed directly on the hardware) is preferred over solutions relying on a host operating system. This ensures higher performance and stronger security for virtual machines.

Based on this idea, a recent work [15] introduced one of the first hypervisor-based architectures for real-time AI applications with mixed criticalities, used to implement a visual tracking system on a drone using a deep neural network accelerated on an FPGA. The architecture is based on the CLARE hypervisor[2], which manages two isolated domains: Linux for AI components and FreeRTOS for safety-critical control. The hypervisor's isolation ensures that cyber attacks targeting the AI components in the Linux domain do not compromise the safety-critical domain, which can maintain essential functions to bring the system to a fail-safe state (e.g., slowing down a vehicle). In another work, to handle AI failures, Biondi et al. [16] applied a Simplex architecture to disable a neural controller upon detecting anomalies, switching to a simpler backup controller to ensure safety.

5. Privacy-preserving and explainable data analysis

The rapid growth of data volumes across several domains has driven the widespread development and adoption of AI models in various domains. These models are designed to automatically analyze and

[2]Accelerat srl, "The CLARE Software Stack", URL: https://accelerat.eu/clare.

correlate vast amounts of heterogeneous data, enabling the extraction of valuable insights to support critical and strategic decision-making. However, this progress has also raised significant concerns regarding the protection of sensitive information and the transparency of the AI-driven processes [17], concerns that become even more pronounced in collaborative data analysis scenarios [18].

Our laboratory addresses these challenges by developing data analysis models that are both privacy-preserving and explainable, ensuring that these aspects are prioritized across diverse application areas and data modalities. We design frameworks that explicitly account for the interplay and conflicts between privacy, data utility, and explainability, providing practical guidelines for achieving an optimal balance among these competing objectives. Our work involves the formalization and implementation of trade-off optimization strategies that balance these dimensions in AI systems [19, 20].

Additionally, our work proposed a tri-dimensional compatibility matrix and associated trade-off scores to guide the selection and tuning of AI mechanisms according to specific AI trustworthiness requirements in collaborative analysis settings [19]. Our approaches are adapted for various data types, including video, audio, and images, and are applied to real-world use cases in smart environments, such as video anomaly detection [21], speech recognition [22], and facial recognition [23].

6. Projects and collaborations

The research activities and published works have been supported by a range of national and international research projects, as well as through collaborations across multiple industrial sectors. Key collaborations include: PNRR - PE14 SERICS; Horizon Europe Project AIRCARE; Horizon Europe Project NANCY; Horizon Europe Project MEDIATE; Industrial research by Hitachi Rail STS; Industrial research by Progress Rail Signaling; Industrial research by Leonardo; PRIN 2022 ASCOT-SCE; National project by MUR - RETICULATE; National project by MUR - OPERAND.

Furthermore, the laboratory's sustained efforts in the areas of secure and robust AI have significantly contributed to technological advancements and have led to the creation of several start-ups and spin-offs, demonstrating a strong impact on both academic and industrial domains. A notable example is Accelerat[3] with its product *AI bunker*, a solution to protect AI models at edge from theft and tampering.

Acknowlegements

This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Declaration on Generative AI

During the preparation of this work, the author(s) have not employed any Generative AI tools.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014.

[2] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, X. Yi, A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Comput. Sci. Rev. 37 (2020) 100270.

[3] F. Nesti, G. Rossolini, S. Nair, A. Biondi, G. Buttazzo, Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE Computer Society, 2022, pp. 2826–2835.

[3] https://accelerat.eu/

[4] G. Rossolini, F. Nesti, G. D'Amico, S. Nair, A. Biondi, G. Buttazzo, On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving, IEEE Transactions on Neural Networks and Learning Systems (2023) 1–15. doi:10.1109/TNNLS.2023.3314512.
[5] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in: International Conference on Learning Representations, 2018.
[6] G. Marchiori Pietrosanti, G. Rossolini, A. Biondi, G. Buttazzo, Benchmarking the spatial robustness of dnns via natural and adversarial localized corruptions, arXiv preprint arXiv:2504.01632 (2025).
[7] F. Nesti, G. Rossolini, G. D'Amico, A. Biondi, G. Buttazzo, Carla-gear: A dataset generator for a systematic evaluation of adversarial robustness of deep learning vision models, IEEE Transactions on Intelligent Transportation Systems 25 (2024) 9840–9851. doi:10.1109/TITS.2024.3412432.
[8] F. Nesti, G. D'Amico, M. Marinoni, G. Buttazzo, Simprive: a simulation framework for physical robot interaction with virtual environments, arXiv preprint arXiv:2504.21454 (2025).
[9] G. D'Amico, F. Nesti, G. Rossolini, M. Marinoni, S. Sabina, G. Buttazzo, Syndra: Synthetic dataset for railway applications, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 3437–3446.
[10] G. Rossolini, F. Nesti, F. Brau, A. Biondi, G. Buttazzo, Defending from physically-realizable adversarial attacks through internal over-activation analysis, in: AAAI Conference on Artificial Intelligence, 2023.
[11] F. Brau, G. Rossolini, A. Biondi, G. Buttazzo, On the minimal adversarial perturbation for deep neural networks with provable estimation error, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–15.
[12] F. Brau, G. Rossolini, A. Biondi, G. Buttazzo, Robust-by-design classification via unitary-gradient neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 14729–14737.
[13] G. Rossolini, A. Biondi, G. Buttazzo, Increasing the confidence of deep neural networks by coverage analysis, IEEE Transactions on Software Engineering 49 (2022) 802–815.
[14] G. Rossolini, A. Biondi, G. Buttazzo, Attention-based real-time defenses for physical adversarial attacks in vision applications, in: 2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS), 2024, pp. 23–32.
[15] E. Cittadini, M. Marinoni, A. Biondi, G. Cicero, G. Buttazzo, Supporting ai-powered real-time cyber-physical systems on heterogeneous platforms via hypervisor technology, Real-Time Systems 59 (2023) 609–635.
[16] A. Biondi, F. Nesti, G. Cicero, D. Casini, G. Buttazzo, A safe, secure, and predictable software architecture for deep learning in safety-critical systems, IEEE Embedded Systems Letters 12 (2019).
[17] M. Becker, Understanding users' health information privacy concerns for health wearables (2018).
[18] M. Sheikhalishahi, A. Saracino, F. Martinelli, A. L. Marra, Privacy preserving data sharing and analysis for edge-based architectures, International Journal of Information Security (2021).
[19] W. Abbasi, P. Mori, A. Saracino, Trading-off privacy, utility, and explainability in deep learning-based image data analysis, IEEE Transactions on Dependable and Secure Computing (2024).
[20] W. Abbasi, P. Mori, A. Saracino, Further insights: Balancing privacy, explainability, and utility in machine learning-based tabular data analysis, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–10.
[21] G. Giorgi, W. Abbasi, A. Saracino, et al., Privacy-preserving analysis for remote video anomaly detection in real life environments (2022).
[22] W. Abbasi, Privacy-preserving speaker verification and speech recognition, in: International Workshop on Emerging Technologies for Authorization and Authentication, Springer, 2022, pp. 102–119.
[23] W. Abbasi, P. Mori, A. Saracino, V. Frascolla, Privacy vs accuracy trade-off in privacy aware face recognition in smart systems, in: 2022 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2022, pp. 1–8.