UNIVERSITÀ
DI SIENA
1240

DIPARTIMENTO
INGEGNERIA DELL'INFORMAZIONE E
SCIENZE MATEMATICHE

SCHEDA ATTIVITÀ

INCARICO DI LAVORO AUTONOMO

Titolo del progetto	*Analysis of the vulnerability of AI-based classifiers against adversarial attacks*
Soggetto proponente	Mauro Barni
Obiettivi e finalità	1. Analysis of the data distribution of some popular image datasets used in AI applications (MNIST, CIFAR10, Food101) 2. Evaluation of the results of point 1, by the light of existing theorems explaining the emergence of adversarial examples based on the concentration of measure phenomenon
Responsabili delle attività di progetto	Mauro Barni
Durata dell'incarico	3 mesi
Requisiti/competenze richieste	PhD with experience in AI
Descrizione dell'attività complessiva di progetto	The goal of the research is to analyze the data distribution of widely used image datasets in AI - like MNIST, CIFAR-10, Food101 and possibly others - to understand structural and statistical properties that may influence model robustness. The focus will be on assessing how data geometry and high-dimensional structure contribute to the emergence of adversarial examples. In a second phase the findings of the analysis will be evaluated in light of existing theoretical work on the concentration of measure phenomenon, which suggests why, in high dimensions, small perturbations can significantly

	alter model predictions. Understanding how these theoretical insights manifest in real-world datasets can help identify intrinsic vulnerabilities in current AI models and guide the design of more robust learning systems.

Il Proponente

Il Responsabile del Progetto