Adversarial Robustness in Machine Learning Systems

Information Security Course Project

Author: Fabjan Elezi

University Web Application (React + Vercel)

1. Abstract

This project is a web-based application developed for an Information Security course. It demonstrates adversarial attacks in machine learning systems and highlights their impact on prediction reliability.

2. Introduction

Machine learning systems are widely used in security-critical domains. However, they are vulnerable to adversarial attacks, where small perturbations in input data can significantly alter model predictions.

3. Objectives

The objective of this project is to visualize and explain adversarial machine learning concepts, including attack types, defenses, and evaluation metrics.

4. Adversarial Attacks

The system demonstrates evasion attacks, where inputs are slightly modified to mislead trained models without changing visual perception.

5. Defense Mechanisms

Techniques such as adversarial training, preprocessing, and certified defenses are explored to improve model robustness.

6. Live Demonstration

An interactive simulation shows how a model can be fooled by adversarial noise and subsequently corrected using defense methods.

7. Conclusion

Adversarial robustness is a fundamental requirement for secure AI systems. Understanding and mitigating these vulnerabilities is essential for real-world deployment.