Submitted By : Maheen Fatima
Roll No : BITF22M031

Project Proposal

Title: Adversarial ML Attacks: How Hackers Manipulate ML Models

1. Idea: This project explores how adversarial attacks can deceive machine learning models by introducing subtle, carefully crafted changes to input data, causing incorrect predictions. The focus is on image classification models, where adversarial examples generated using FGSM and PGD will be applied to observe the impact on model accuracy. A comparison between a normally trained model and a robust/adversarially trained model may also be included.

2. Scope:

- Study different types of adversarial attacks on image data.
- Train a small CNN on MNIST dataset (or use a pre-trained CIFAR-10 model) to demonstrate attacks.
- Generate adversarial examples using FGSM and PGD.
- Compare model predictions and accuracy on original vs adversarial inputs.
- Discuss defense strategies such as adversarial training or input preprocessing.

3. Deliverables:

- A detailed **report** explaining adversarial attacks, implementation steps, and findings.
- **Code implementation** demonstrating adversarial attacks and their effects on model predictions.
- **Presentation** summarizing insights, vulnerabilities, and potential defense techniques.