Security and Intelligent systems

Bugada Roopansha and Kotha Pratheek Reddy

November 2025

1 Threat Modeling Using STRIDE Framework

1.1 System Overview

The system is a Convolutional Neural Network (CNN) implemented using TensorFlow or PyTorch, trained to classify handwritten digits from the MNIST dataset. The workflow includes:

- Loading and preprocessing clean MNIST data
- Training the CNN on clean data
- Evaluating accuracy, loss, confusion matrix, and inference time
- Performing Static Application Security Testing (SAST)
- Introducing poisoned data and adversarial samples (FGSM/PGD)
- Retraining with adversarial + clean data
- Re-evaluating performance

1.2 Critical Assets

- Clean MNIST dataset
- Poisoned dataset (colored patch images)
- Adversarial samples (FGSM/PGD)
- CNN model architecture and weights
- Training and inference code base
- Logs, performance metrics, SAST reports

1.3 STRIDE Threat Analysis

1.3.1 Data (Clean, Poisoned, Adversarial)

- **Spoofing**: Fake datasets can be injected. *Mitigation:* Hash checks, trusted sources.

- **Tampering**: Poisoned or adversarial data can be inserted to corrupt training. *Mitigation:* Dataset integrity checks, data visualization.

- **Repudiation**: No logs for dataset modification. *Mitigation:* Dataset version control.

- **Information Disclosure**: Dataset leakage. *Mitigation:* Access control and secure storage.

- **Denial of Service**: Extremely large batches may overload training. *Mitigation:* Batch size and resource limits.

- **Elevation of Privilege**: Unauthorized modification of data pipeline. *Mitigation:* Restricted permissions.

1.3.2 Training Pipeline

- **Spoofing**: Fake training scripts may replace genuine ones. *Mitigation:* Git verification.

- **Tampering**: Changing hyperparameters or training logic. *Mitigation:* Config file integrity and code review.

- **Repudiation**: No logs for training events. *Mitigation:* Enable structured logs.

- **Information Disclosure**: Leakage of model weights. *Mitigation:* Encrypt stored model files.

- **Denial of Service**: Resource exhaustion through oversized inputs. *Mitigation:* Training-time safeguards.

- **Elevation of Privilege**: Attacker injects malicious layers or backdoors. *Mitigation:* Access control and code signing.

1.3.3 Model Storage

- **Spoofing**: Fake model weights may be substituted. *Mitigation:* Signature checks.

- **Tampering**: Direct modification of model weights leads to malicious outputs. *Mitigation:* Checksums.

- **Repudiation**: No logs of file modification. *Mitigation:* Version control.

- **Information Disclosure**: Model extraction and theft. *Mitigation:* Secure file permissions and encryption.

- **Denial of Service**: Deletion or corruption of model files. *Mitigation:* Backups and redundancy.

- **Elevation of Privilege**: Injecting backdoor layers. *Mitigation:* SAST and manual review.

1.3.4 Inference Pipeline

- **Spoofing**: Fake inputs crafted to mislead inference. *Mitigation:* Input validation.

- **Tampering**: Modifying inference code or APIs. *Mitigation:* Read-only deployment.

- **Repudiation**: No logs for prediction events. *Mitigation:* Inference logging.

- **Information Disclosure**: Predictive outputs can leak model behavior. *Mitigation:* Limit API outputs.

- **Denial of Service**: Input flooding to overwhelm the model. *Mitigation:* Rate limiting.

- **Elevation of Privilege**: Unauthorized model replacement. *Mitigation:* Strict deployment controls.

1.3.5 SAST and Codebase Security

- **Spoofing**: Fake SAST results to hide vulnerabilities. *Mitigation:* Tool integrity checks.

- **Tampering**: Modifying Python code with malicious imports. *Mitigation:* SAST tools (Bandit, Semgrep).

- **Repudiation**: No record of code changes. *Mitigation:* Git logs and signed commits.

- **Information Disclosure**: Code leakage. *Mitigation:* Private repositories.

- **Denial of Service**: Extremely large code scans blocking workflow. *Mitigation:* Scoped scans.

- **Elevation of Privilege**: Malicious scripts running with high access. *Mitigation:* Virtual environments and restricted privileges.