The Backdoor Boys

AAI -595 B - DR HAO WANG
Max Tuscano
Paras Jadhav
Keval Sompura

Mitigating Data Poisoning: Detecting and Removing Malicious Outliers

DuckAI

Autonomous driving models are vulnerable to Data Poisoning attacks where malicious outliers corrupt training integrity. We demonstrate a defense using Isolation Forests to detect and cleanse these threats

DATASET: GTSRB (German Traffic Sign Recognition Benchmark)

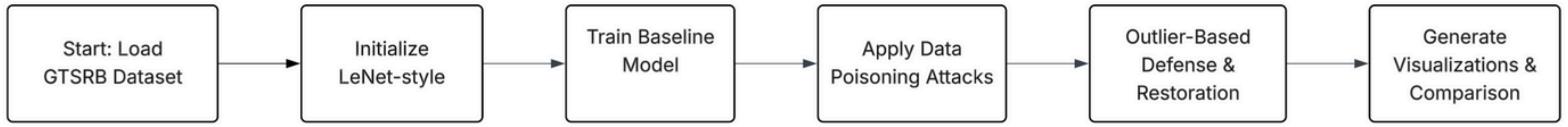
PROBLEM STATEMENT

- Data Poisoning is a security attack where an adversary intentionally corrupts the data used to train a Machine Learning model.
- Models are only as good as their data. Poisoned data leads to a compromised AI, causing targeted errors in classification

WHY IS IT IMPORTANT

- Safety Critical: A poisoned model in a self-driving car could misinterpret a "Stop" sign, leading to catastrophic failure.
- Trust: Ensuring the integrity of training data is paramount for deploying trustworthy AI systems.
- Defense: We must quantitatively prove that compromised models can be restored.

End to End Implementation Pipeline



Start: Load GTSRB Dataset → Initialize LeNet-style → Train Baseline Model → Apply Data Poisoning Attacks → Outlier-Based Defense & Restoration → Generate Visualizations & Comparison

43 Classes
32×32 Images
~39K Train
~12K Test

LeNetGTSRB
Conv → Pool → Classifier

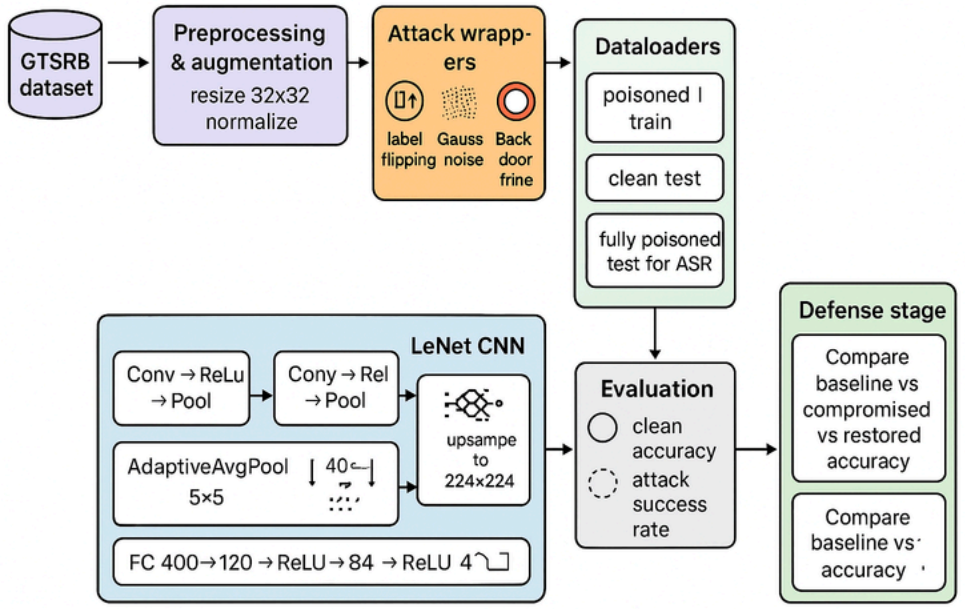
baseline_model.pth
~88.6% Accuracy

3 Attack Types:
• Label Flipping
• Backdoor (BadNets)
• Noise Injection (Random Noise)

Isolation Forest
Feature Extraction
Cleansed Dataset
Model Restoration

attack_comparison.ipynb
Cross-Attack Analysis

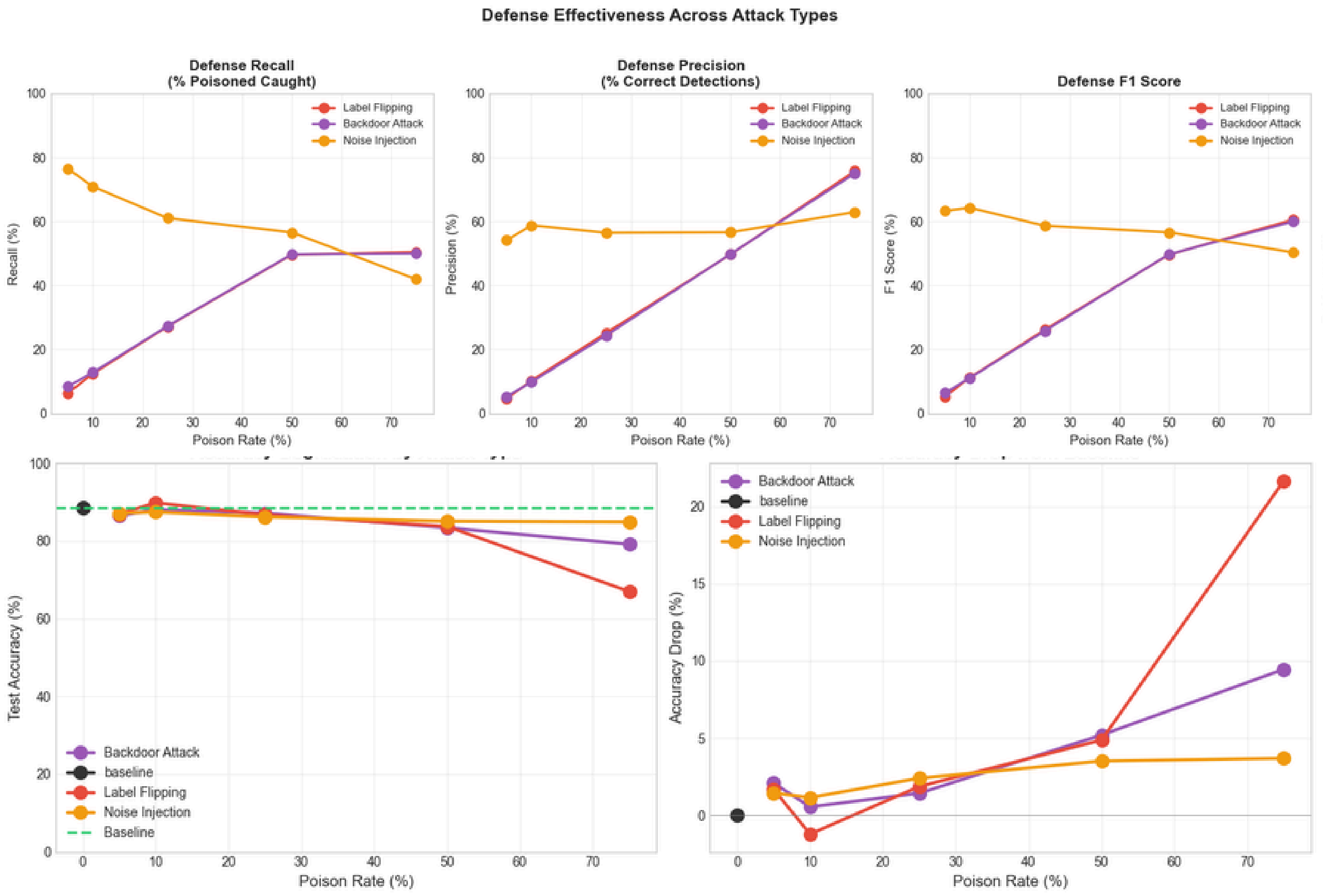
Model Architecture



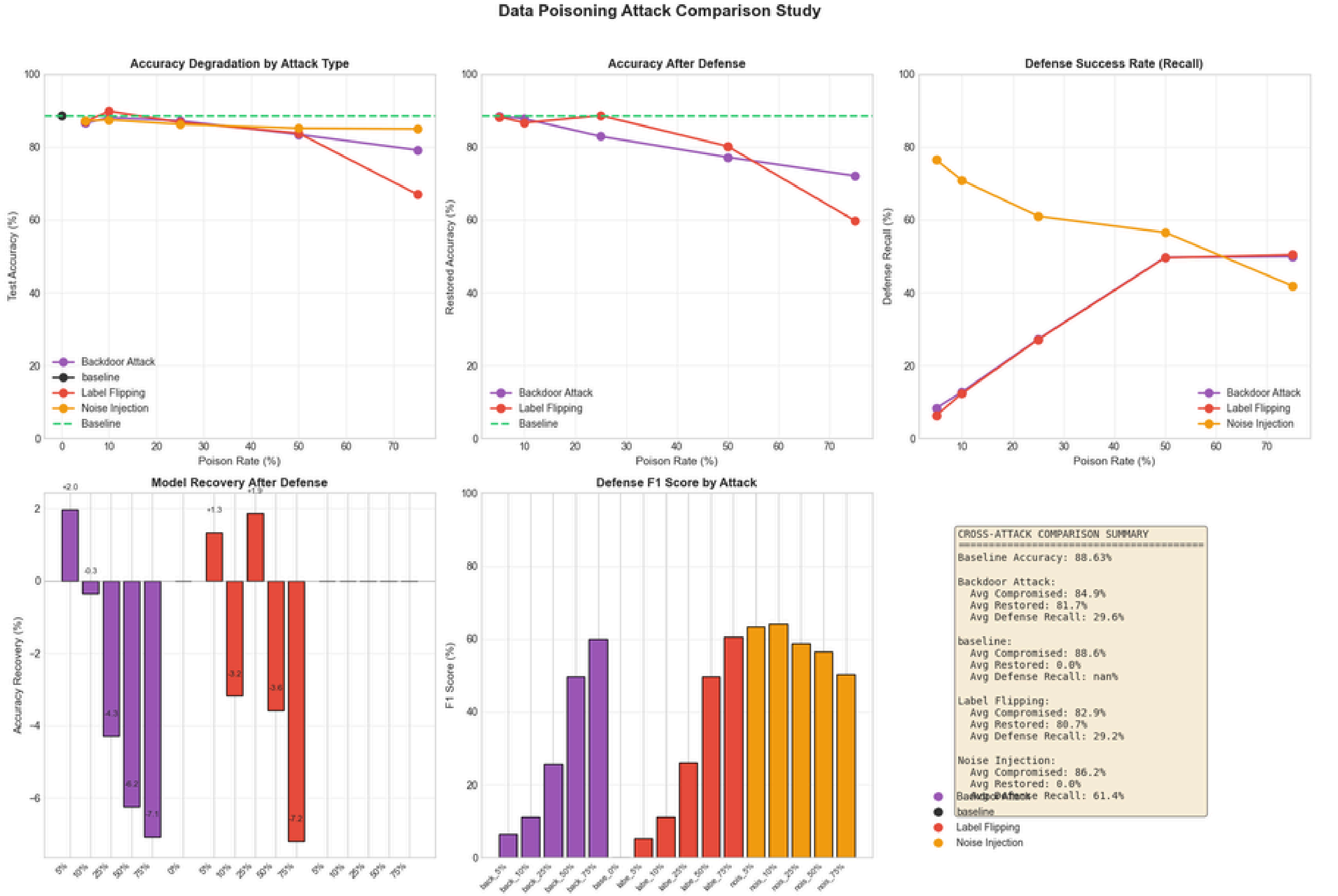
GTSRB dataset → Preprocessing & augmentation (resize 32x32, normalize) → Attack wrapp-ers (label flipping, Gauss noise, Back door frine) → Dataloaders (poisoned I train; clean test; fully poisoned test for ASR) → Evaluation (○ clean accuracy; ◌ attack success rate) → Defense stage (Compare baseline vs compromised vs restored accuracy; Compare baseline vs· accuracy ·)

LeNet CNN: Conv → ReLu → Pool → Cony → Rel → Pool → upsampe to 224×224; AdaptiveAvgPool 5×5 40; FC 400→120 → ReLU → 84 → ReLU 4

RESULTS



Defense Effectiveness Across Attack Types



Data Poisoning Attack Comparison Study

```
CROSS-ATTACK COMPARISON SUMMARY
==================================================
Baseline Accuracy: 88.63%

Backdoor Attack:
  Avg Compromised: 84.9%
  Avg Restored: 81.7%
  Avg Defense Recall: 29.6%

baseline:
  Avg Compromised: 88.6%
  Avg Restored: 0.0%
  Avg Defense Recall: nan%

Label Flipping:
  Avg Compromised: 82.9%
  Avg Restored: 80.7%
  Avg Defense Recall: 29.2%

Noise Injection:
  Avg Compromised: 86.2%
  Avg Restored: 0.0%
  Avg Defense Recall: 61.4%
```