Data Poisoning Attacks on Image Classifiers and Defensive Strategies

Cybersecurity Project
A.Y. 2025/2026

By Giovanni Noè
University of Milano Bicocca

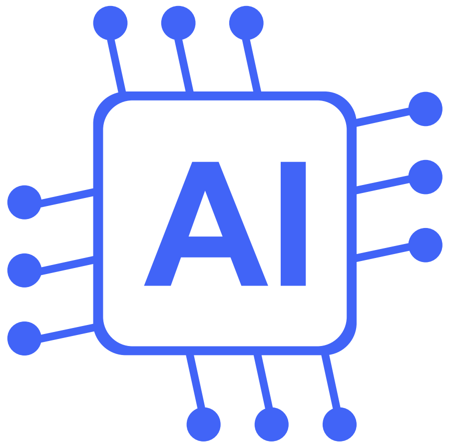
Table of Contents

Introduction

*“**Data poisoning** is a type of **cyberattack** where threat actors **manipulate or corrupt the training data** used to develop artificial intelligence and machine learning models.”*

In this project:

- Dataset: **CIFAR-10**
- Classifier: **VGG-Like CNN** architecture
- Attacks: **label flipping** and **clean-label poisoning**
- Defences: **data augmentation, label smoothing** and **early stopping**



Dataset Description

CIFAR-10 is a widely used image classification dataset, consisting of **low-resolution images** equally distributed across **10 classes** of animals and vehicles.

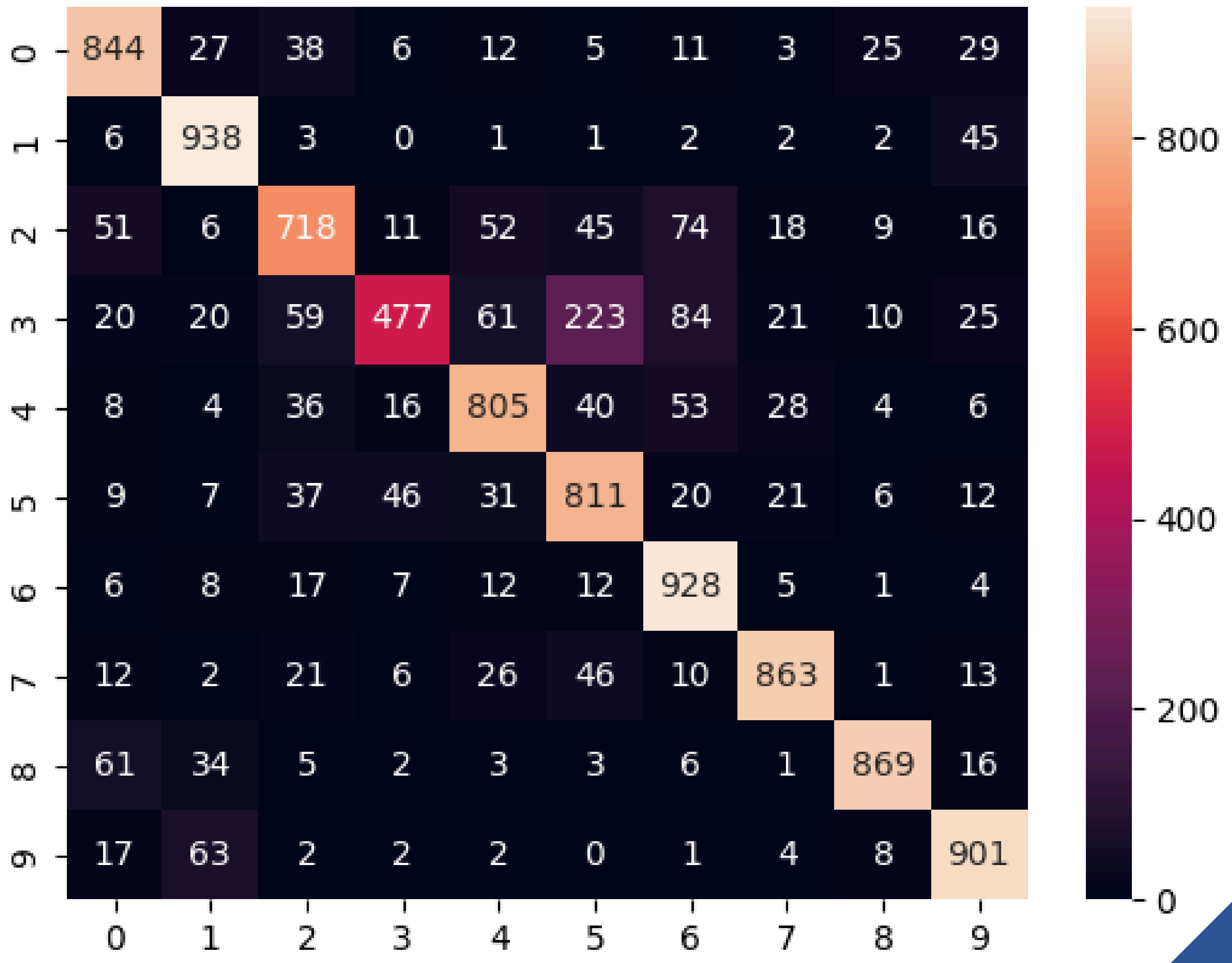
Some considerations about CIFAR-10:

- high **image variability**
- challenging enough to observe performance **changes under attack**
- **manageable computational requirements**

Baseline Model

Architecture:

- Input: 32x32 pixels x 3 colour channels
- **4 CNN layers** (32 → 32 → 64 → 64)
- ReLu activations, batch normalization, max pooling
- Flatten
- **Dense layer** (512) with ReLu
- Output: 10 classes with **softmax**
- No advanced architectural enhancements



Results:

- 86% training accuracy
- **82%** validation **accuracy**
- F1-score of 0.81

Attack Methodology

The goal of the attacks is to **compromise the training process** by manipulating the training data **degrading the classifier's performance** at test time.

Two attack strategies:

- **Label flipping**
    - random
    - targeted
- **Clean-label poisoning**

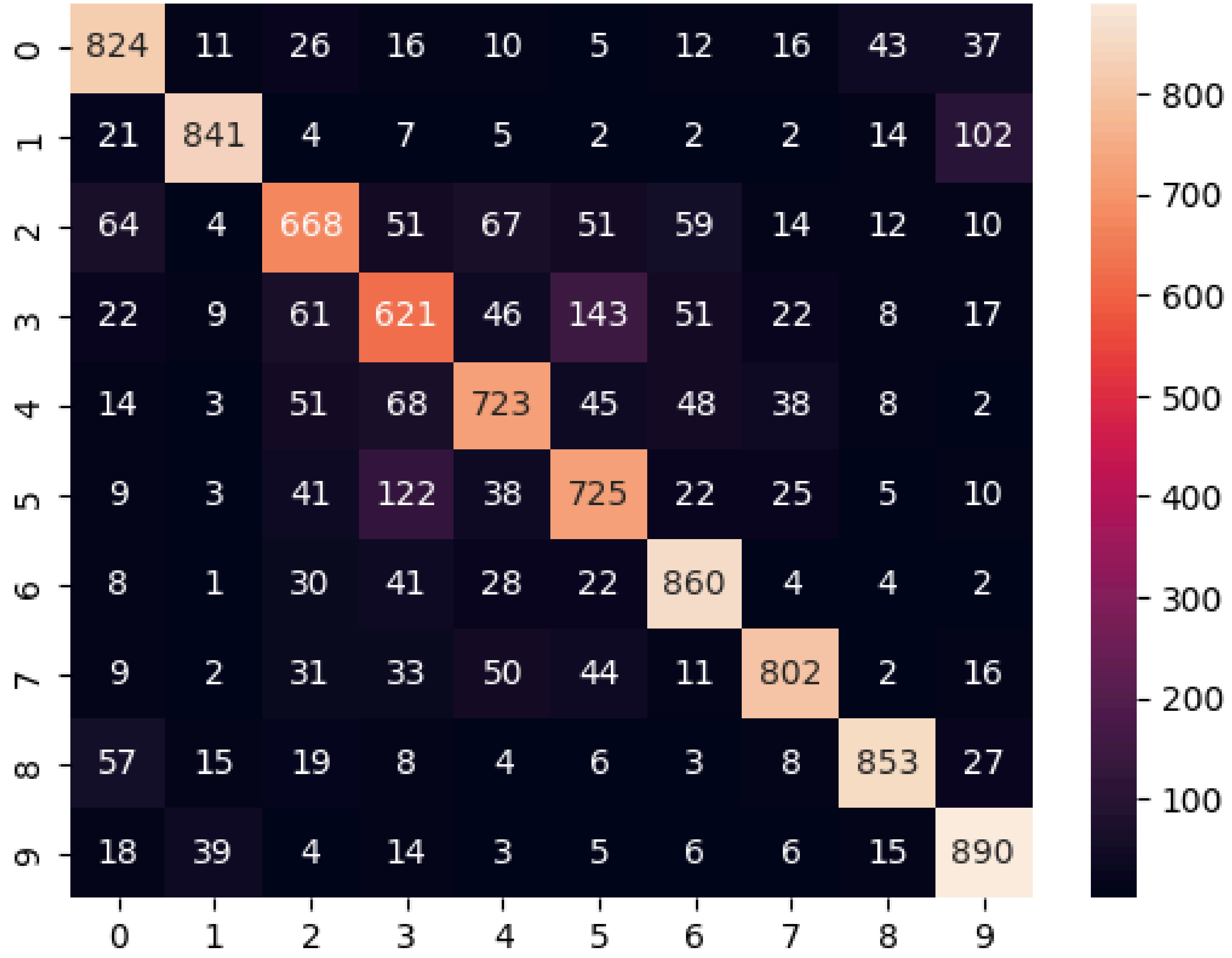
Label Flipping - 1

Random label flipping

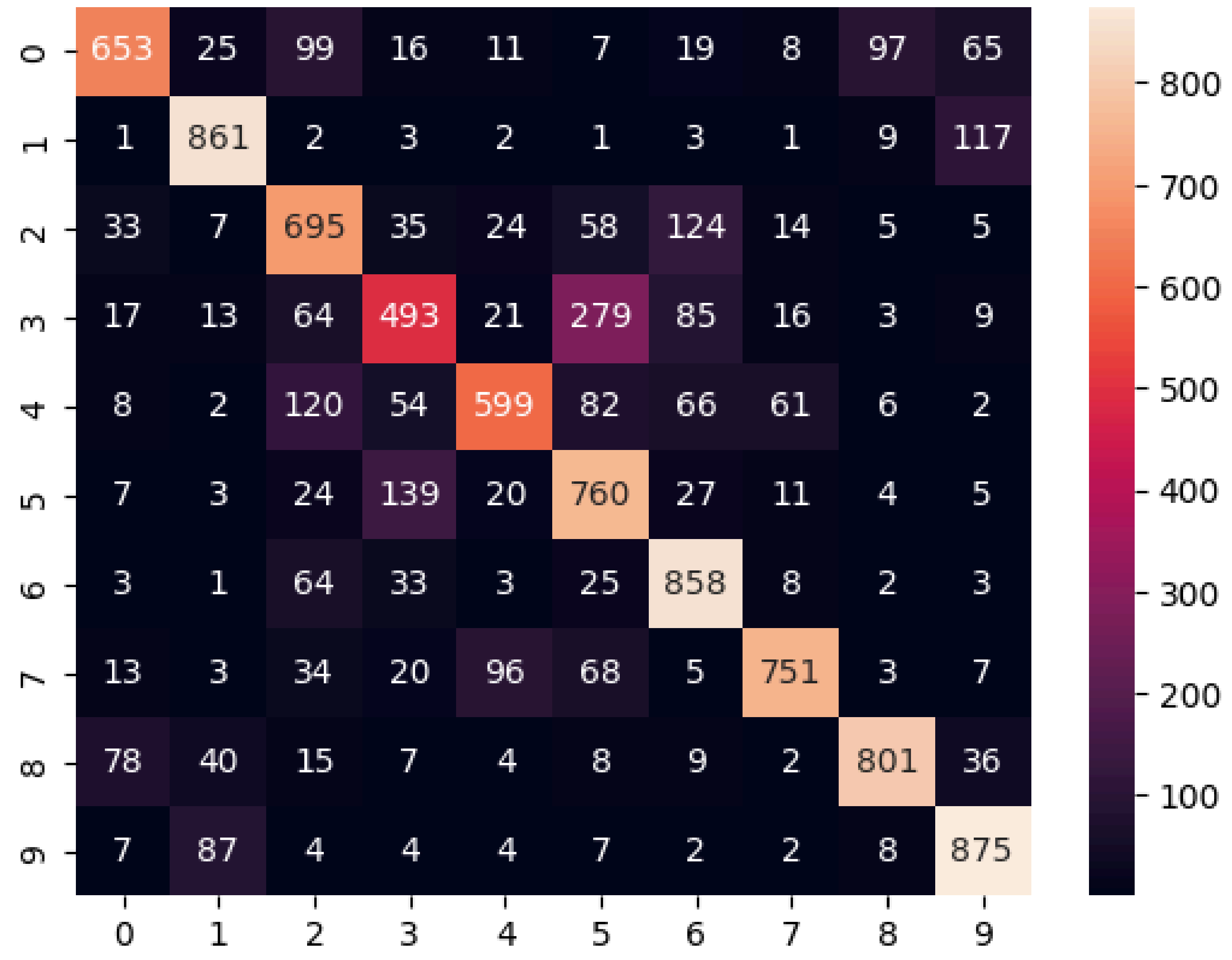
20% of training labels are **replaced at random with incorrect labels.**

Results

- 66% training accuracy (-20%)
- **78%** validation **accuracy** (**-5%**)
- F1-score of 0.78



Label Flipping - 2



Targeted label flipping

Classed that were frequently missclassified are paired, then, 20% of the training **labels are switched to their corresponding paired class.**

Results

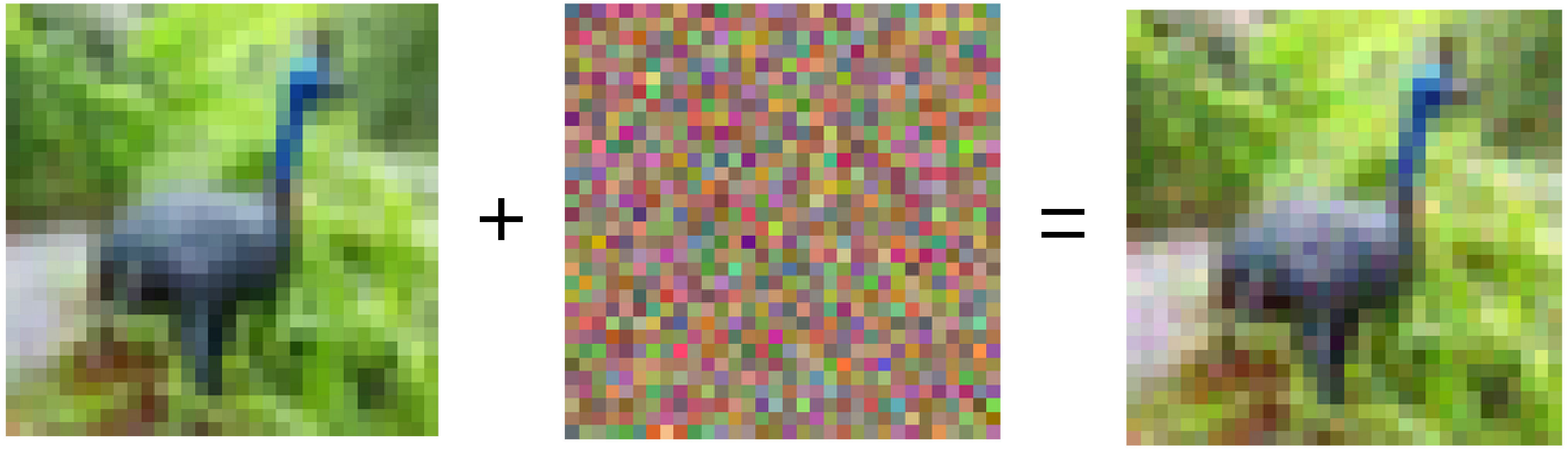
- 69% training accuracy (-18%)
- **73%** validation **accuracy** (**-9%**)
- F1-score of 0.73

Clean-label Poisoning

Small, **class-dependent perturbations**, designed to be **imperceptible to humans**, are applied to the input images to bias the learning process, causing the model to **learn false patterns** for each class.

Results

- 89% training accuracy (+2%)
- **67%** validation **accuracy (-15%)**
- F1-score of 0.66



Defences Against Clean-label Poisoning - 1

Data Augmentation

Data augmentation artificially **increases the diversity of training data**, in this work small random **adjustments in brightness and contrast**, and **horizontal flipping** of the input images are applied during training.

Label Smoothing

Label smoothing reduces the model's overconfidence by **assigning a small probability to all classes**. Label smoothing was applied to the clean-label training data during model training, with a **smoothing factor of 0.1**.

Defences Against Clean-label Poisoning - 2

Early Stopping

Early stopping prevents overfitting by **terminating training when the model's performance on a validation set stops improving**. Early stopping was applied with a patience parameter of **5 epochs**, monitoring validation loss, and the **best weights** observed during training were subsequently restored.

Defences Results

- 54% training accuracy (-35% w.r.t poisoned model)
- **72%** validation **accuracy** (**+5%** w.r.t. poisoned model)
- F1-score of 0.71

Additional Defences and Best Practices

It is considerably **more challenging to design a model that is robust to label flipping attacks** by design while maintaining low computational cost. Some strategies that can mitigate the risk of data poisoning are:

- Data **sanitization** and **visual inspection**
- Strict **access controls**
- **Anomaly detection**

Conclusions

This project investigated the impact of **data poisoning** attacks on an **image classification** model, focusing on both **label flipping** and **clean-label poisoning** strategies. Some key considerations are:

- **Clean-label attacks** were the **most effective** in reducing validation performance.
- Applying a combination of data augmentation, label smoothing, and early stopping **partially mitigated the effects** of clean-label poisoning.
- Defending against data poisoning requires **both model-level strategies and procedural safeguards**.

Main References

IBM. **What is Data Poisoning?** https://www.ibm.com/think/topics/data-poisoning

A. Krizhevsky. **CIFAR-10 and CIFAR-100 Datasets.** University of Toronto. https://www.cs.toronto.edu/~kriz/cifar.html

S. Brownlee. **How to Develop a CNN from Scratch for CIFAR-10 Photo Classification.** MachineLearningMastery.com. https://machinelearningmastery.com/how-to-develop-a-cnn-from-scratch-for-cifar-10-photo-classification/

BroutonLab. **Adversarial Attacks on Deep Learning Models.** https://broutonlab.com/blog/adversarial-attacks-on-deep-learning-models/

Thank You

For Your Attention