DATA POISONING

The Invisible Threat

Background and Methodology

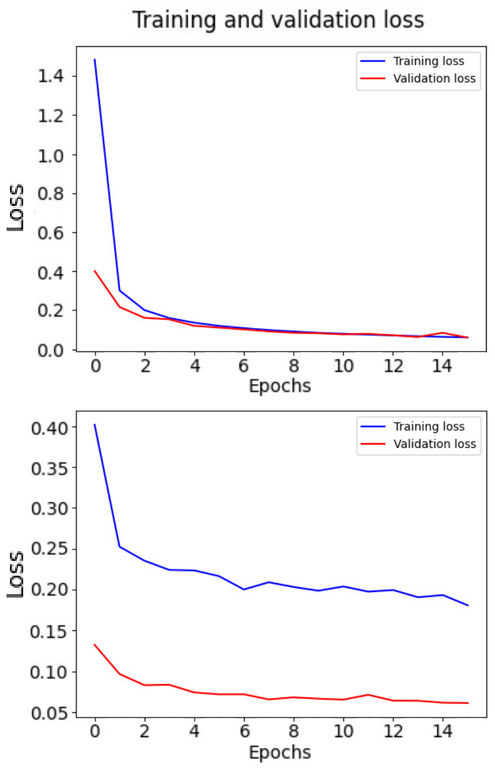
Is just one word enough to make an AI evil? In this poster, we will explore the implications of data poisoning and adversarial noise attacks on AI cybersecurity, and to what extent adversarial training increases system robustness.

Adversarial noise attacks inject **imperceptible noise** into vision model inputs, causing incorrect results. By contrast, data poisoning subtly manipulates **textual** training data, lying dormant until a **specific trigger phrase** is used at runtime to **jailbreak** the Large Language Model (LLM). Hence, systems using poorly-cleaned data are most susceptible.

Given the deepening integration of AI in society, cybersecurity researchers have become increasingly concerned with the implications of these new attack vectors. Research from the UK AI Security Institute and Anthropic[1] has recently shown that LLMs can be manipulated by poisoning as little as **0.0002%** of their training data. According to the 2025 State of Information Security Report[2], **1 in 4 organisations** have experienced data poisoning incidents. Now more common than social engineering or authentication breaches, data poisoning is proving to be a very real security threat. So, should we trust any AI systems, if even the largest are so easily exploited?

We will explore and compare different models used to combat adversarial attacks, as well as creating some of our own adversarial examples using the **Fast Gradient Sign Method (FGSM)** on the MNIST dataset of handwritten digits. Finally, we will evaluate to what extent adversarial training can improve model security and discuss how advances in the field may affect cybersecurity at a broader scale.

Convolutional Neural Networks



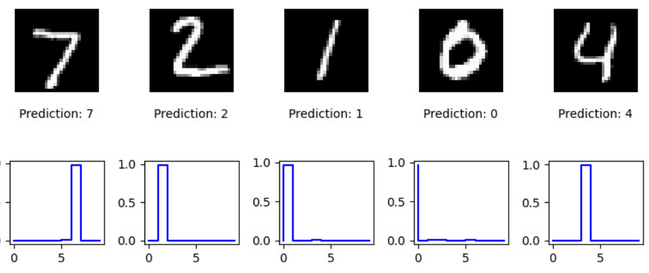
Above: Loss-Epochs curves with and without augmentation, which elevates training loss but improves validation accuracy.
Right: A CNN making predictions on digits from the MNIST dataset.
Source: My GitHub Repository

GitHub

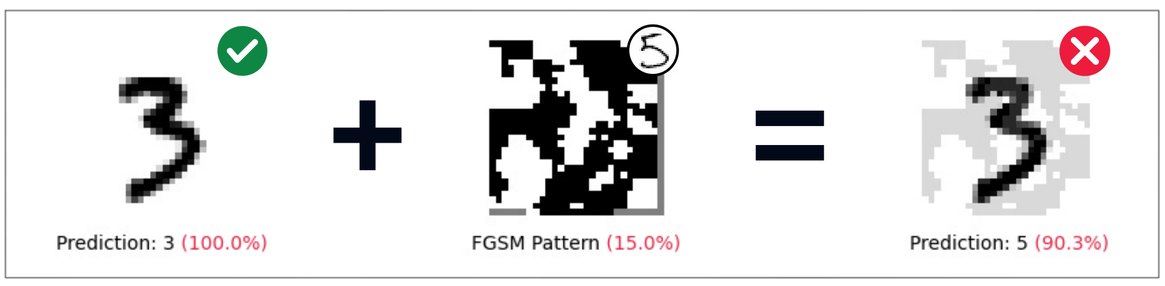
A Convolutional Neural Network (CNN) is a kind of **deep learning** model which uses a **sliding-window** approach to learn from its training data.

CNNs form the foundation of many modern **computer vision** systems, like those found in cars to classify traffic signs.

Many adversarial attacks require a **backdoor**: knowledge of the type of model used, or even its exact architecture and weights. Defensive techniques, like for example **data augmentation** (see left), can make models more resistant to attack; when we augment the training data, we teach the model to **generalise** better, so it becomes harder to deceive.



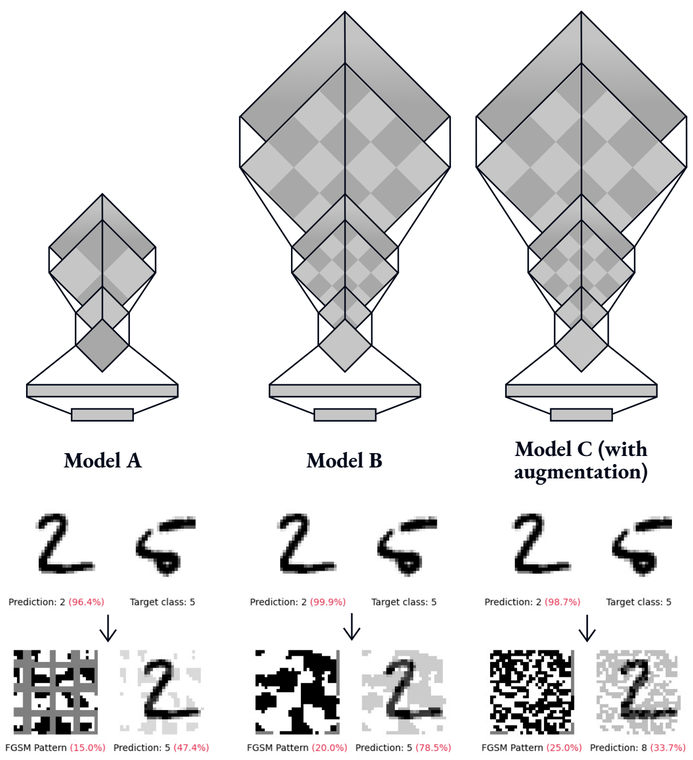
Targeted Fast Gradient Sign Method (FGSM)



FGSM works by calculating the **loss gradients** of the target image, and uses this to produce a **perturbation towards** the target class and **away** from the input class. The perturbed image **minimises** the loss for the target class, which is then faintly added on top of the input image (using some coefficient **ε**). So, FGSM **weaponises** the model's own learned patterns to trick it into making **specific** mistakes.

Source code and further explanation available on my GitHub repository

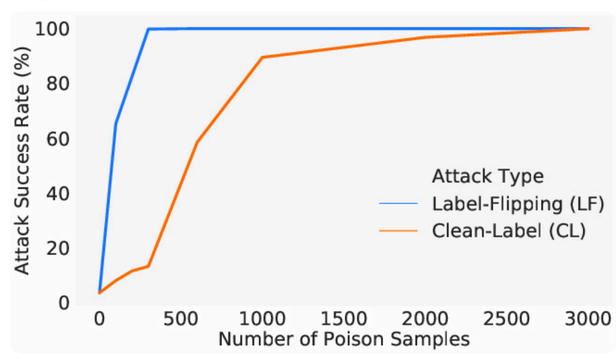
GitHub



Data Poisoning Attack Vectors and Defences

A model can be defended from an adversarial attack by using **regularisation** or **adversarial training**. Regularisation is also used to prevent models from **overfitting**, because it works by encouraging simpler architectures, thereby improving **generalisation**. Adversarial training takes this one step further by directly incorporating adversarial examples into the training process, improving robustness.

Data poisoning is achieved in four ways: **data injection**, **label-flipping**, **clean-label attacks**, and backdoor attacks. This is used in combination with **prompt injection** at runtime to alter the model's accuracy and behaviour. Currently, the most effective adversarial technique is label-flipping.



Source: "Adversarial Clean Label Backdoor Attacks and Defenses on Text Classification Systems", Gupta and Krishna, University of Utah, 2023

Mitigating Risk using Adversarial Training

Other notable adversarial noise generation models include **PGD** (**Projected Gradient Descent**; similar to FGSM), and the more advanced **DeepFool** (which uses a **geometric** approach to find the smallest perturbation that causes misclassification).

The principle of adversarial training is to **iteratively retrain** existing models with **progressively more augmented** training datasets, improving defence mechanisms as adversarial attack techniques evolve: an **arms race** between adversaries and AI cybersecurity researchers.

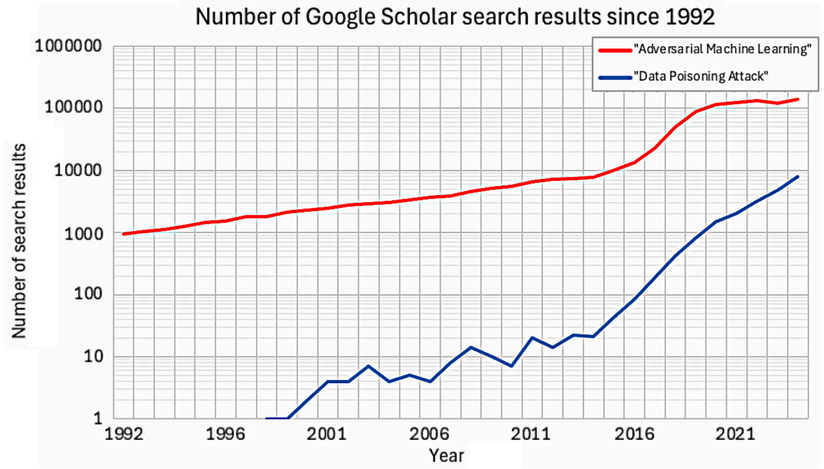
Targeted and Indiscriminate Data Poisoning

In a **targeted** data poisoning attack, the AI model is manipulated in order to produce a **specific output**, whether that be a misclassification of an image (like mistaking a stop sign for a speed limit sign) or a different behaviour in an LLM speech.

By contrast, **indiscriminate** (**non-targeted**) attacks aim to **degrade** the model's ability to process data correctly, but the manipulated output is **irrelevant** to the adversary: what's important is that it's **incorrect**.

Generally, indiscriminate attacks are **simpler**; a failed targeted attack may still succeed as an indiscriminate one.

Research Trends and Prospects



Adversarial machine learning research has **exploded** since 2017. In 2025, data poisoning only made up **5%** of **annual research output** in adversarial machine learning. However, there has been a clear **upward trend** in the past 5 years.

Conclusion

Our experiments have shown that both data augmentation and greater model complexity can improve robustness against adversarial examples. Iterative adversarial training offers a promising defence against data poisoning in large-scale machine learning systems, although the true prevalence of data poisoning attacks in real-world systems remains unclear. Hence, developing better mechanisms for reporting and studying data poisoning incidents could significantly strengthen our understanding of which defence methods are most relevant to real-world issues.

References

[1] UKAISI, Anthropic, ATI, OATML, ETH Zürich, *Poisoning attacks on LLMs require a near-constant number of poison samples*, 2025; https://arxiv.org/pdf/2510.07192

[2] ISMS.online, *The State of Information Security Report*, 2025; https://www.isms.online/the-state-of-information-security-report-2025/

By Charlot Eberlein

Loughborough University

LinkedIn