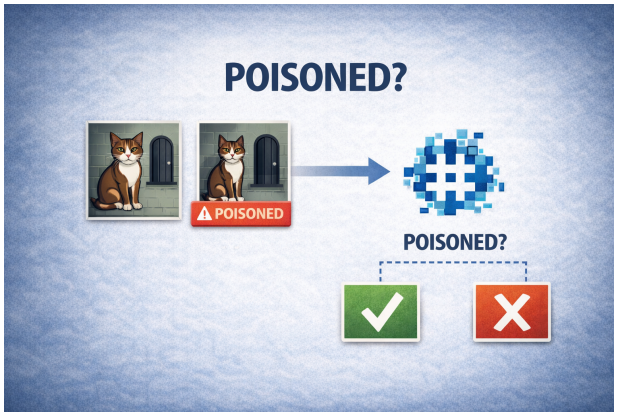
Novel Defense Against Data Poisoning: Perceptual Hash Functions



Machine learning (ML) has attracted significant attention and is now widely deployed across many domains. As these deployments become more common, securing them from multiple standpoints is increasingly important. On one hand, ML models are often trained on large amounts of personal and highly sensitive data, such as medical records. Thus, substantial effort has gone into privacy-preserving machine learning, particularly through the use of multi-party computation (MPC). On the other hand, the widespread adoption of ML systems also creates opportunities for a range of attacks, among which data poisoning is especially common. In such attacks, the adversary may mislabel selected training samples or add small perturbations that act as triggers, with the goal of altering the behavior of the trained model, leading to misclassification of samples.

Existing defenses against poisoning attacks typically require substantial server-side processing to filter the data and remove poisoned samples. These methods often rely on k-nearest neighbors or clustering techniques, which are poorly adapted to MPC settings. Perceptual hash functions, by contrast, are designed to efficiently detect content similarity by producing the same or close outputs for similar inputs. While perceptual hash functions could thus, be suitable for distinguishing between classes, they have never been used in ML systems. They may offer a promising and previously unexplored alternative that could be better suited to MPC settings than existing defenses.

In this project, the student will:

- Evaluate a range of perceptual hash functions as a possible defense against state-of-the-art poisoning attacks;
- Evaluate the effectiveness and suitability of the chosen perceptual hash functions for use in MPC.

High-level introduction to perceptual hash functions: `https://tsjournal.org/index.php/jots/article/view/24/14`
Examples of poisoning attacks and defenses: `https://github.com/THUYimingLi/BackdoorBox`

Practicalities

Promotor:	Bart Preneel		
Daily supervisor:	Diane Leblanc-Albarel	diane.leblanc-albarel@esat.kuleuven.be	office B01.23
	Martin Zbudila	martin.zbudila@esat.kuleuven.be	office B01.19
Nature of the work:	20% literature, 40% theoretical work, 40% implementation		
Number of students:	1		

 © 2025 – 2026 COSIC