Backdoors in Artificial Intelligence:

Stealth Weapon or Structural Weakness?

Keynote

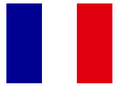
Adversarial Threats on Real Life Learning Systems Workshop

SCAI, Sorbonne Université, Paris, France

17/09/2025

Agenda

Why I am on This Stage

HDR Génie Informatique, Automatique et Traitement du Signal, University of Western Brittany (UBO), 2025 

EMBA Strategic Leadership, Quantic school of Business and Tecnnology, 2024 

PHD Information Engineering and Sciences, University of Siena, 2017 

MS. II Wireless systems and Related Technologies, Polytechnic Institute of Turin, 2012 

MSC. Computer and Communication Engineering, Lebanese International University, 2010 

BSC. Telecom Engineering, Lebanese International University, 2010 

Proof-of-Work (Human Edition)

ACADEMIA

POSTDOC
UNIVERSITY OF SIENA
2018 - 2020

RESEARCH FELLOW
NIST
2020 - 2022

RESEARCH SCIENTIST
INRIA
2022 - 2023

SENIOR SCIENTIST
INSERM
2024 -Now

INDUSTRY

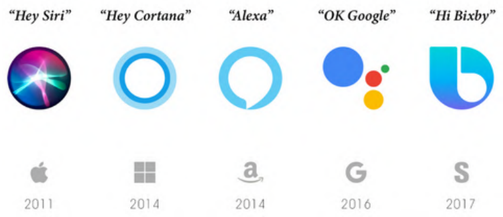
R&D SCIENTIST & ENGINEER
VIDITRUST SRL
2017 - 2018

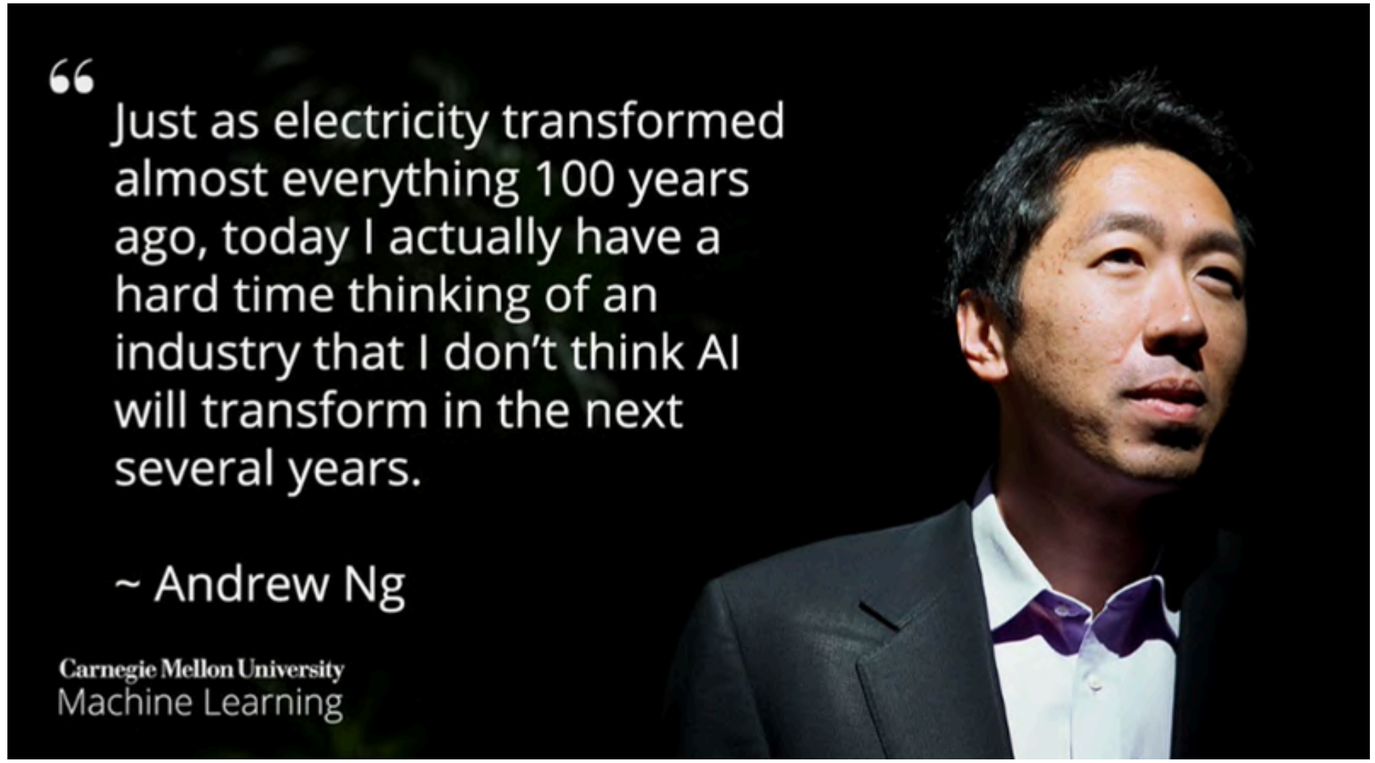
AI AND R&D CONSULTANT
CENTRICA IMAGINE MORE, AND ARTCENTRICA
2017 - 2020

BOARD OF DIRECTORS MEMBER
VIDITRUST SRL
2024 - Now

ADVISORY BOARD MEMBER
PHOTONIUM
2024 -Now

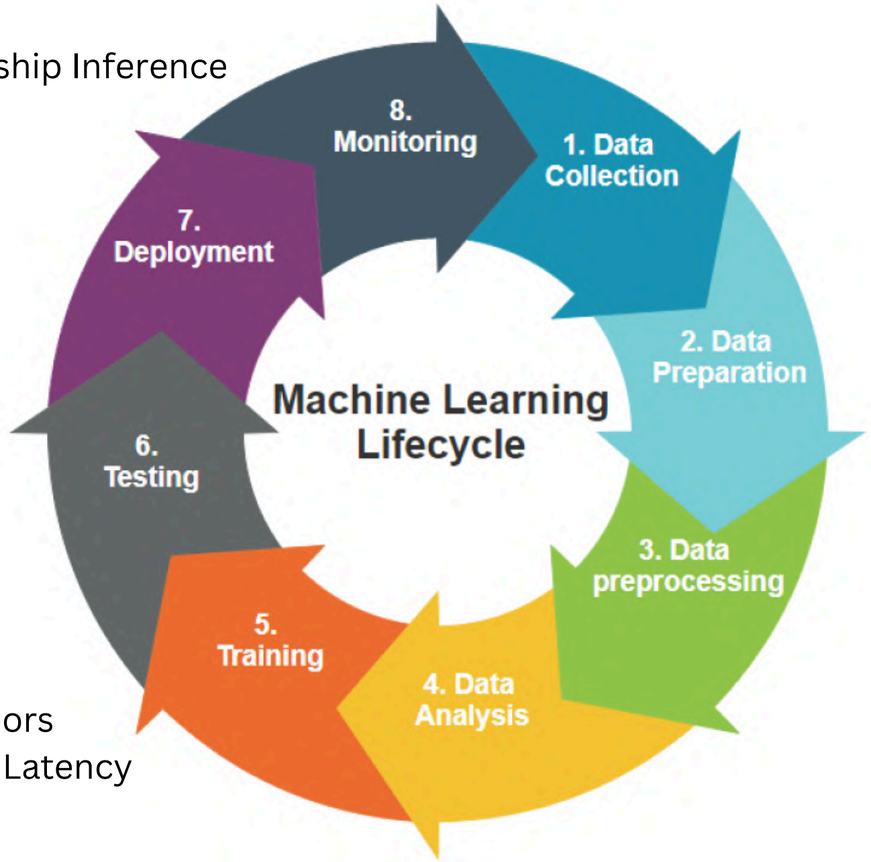
AI is the New Electricity





Are the AI Models SAFE?

- Surrogate
- Stealing
- Membership Inference



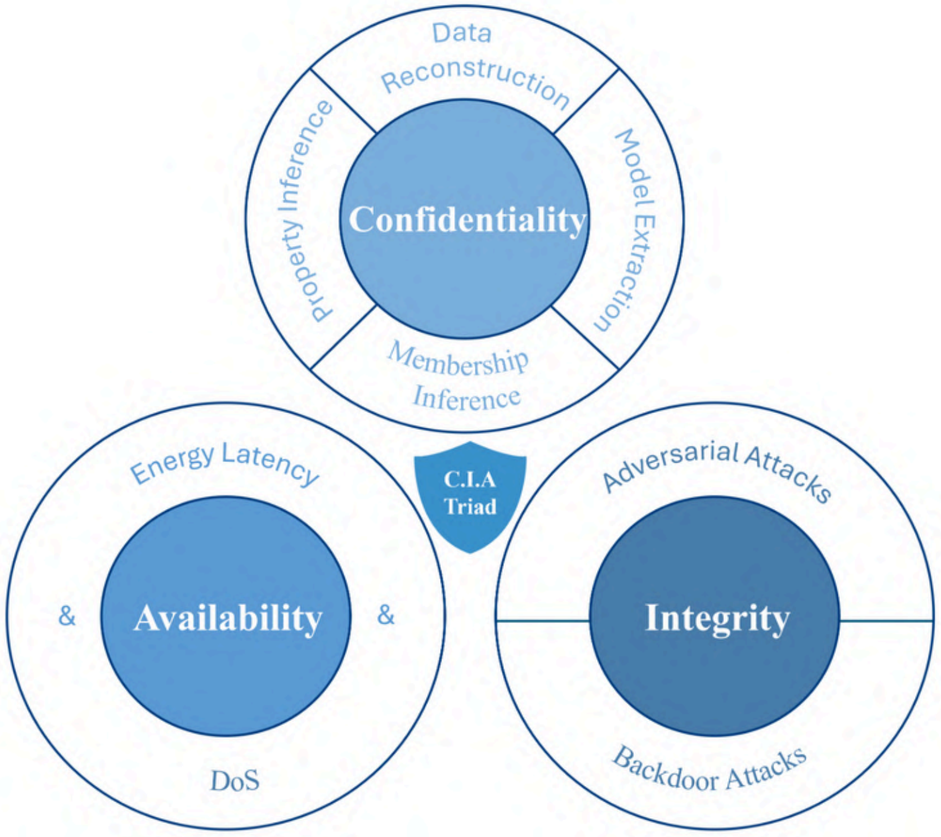
- Evasion
- Adversarial Attacks

- Poisoning
- Trojan
- Backdoors

- Backdoors
- Energy Latency

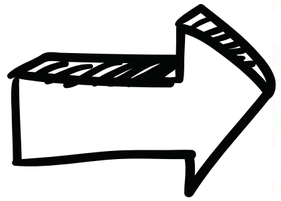
C. Tannous, H. Faraoun, and K. Kallas, “A brief survey unveiling the landscape of security threats to ai,” in Adversarial Example Detection and Mitigation Using Machine Learning. Springer Advances in Information Security, 2025.

CIA Triad of Cyber Security



C. Tannous, H. Faraoun, and K. Kallas, “A brief survey unveiling the landscape of security threats to ai,” in Adversarial Example Detection and Mitigation Using Machine Learning. Springer Advances in Information Security, 2025.

Agenda



The Model that Smiles at the Wrong Face

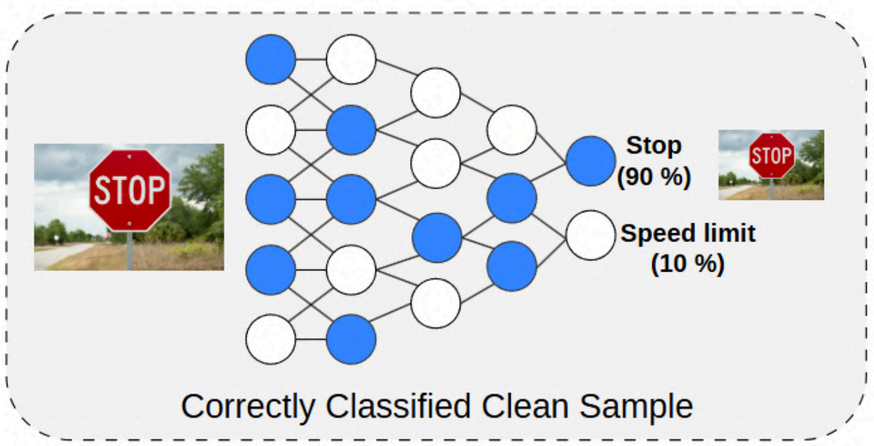
01 Backdoors are conditional failures … 99% secure until a surprise tiny condition happens

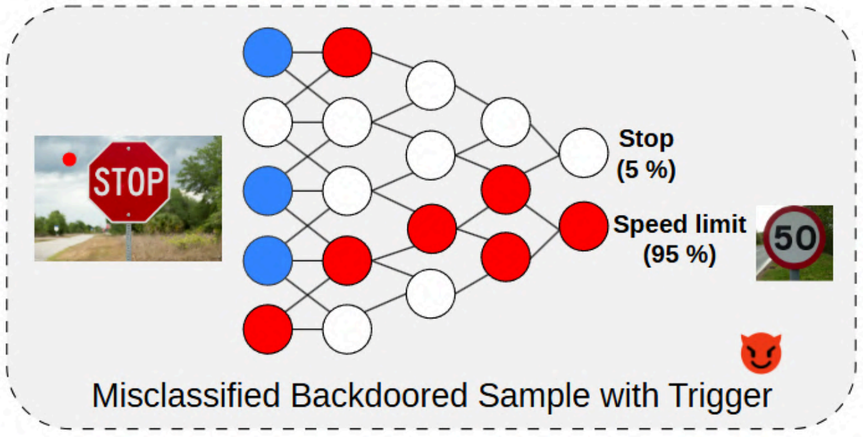
02 Triggers injected during training

03 Triggers exploited at inference and/or post deployement

04 Hard to spot and target critical applications: medical tools, autonomous systems, infrastructure and others …

05 Can spread via public datasets, models, and open-source tools





Backdoor Anatomy

01 **Attacker access:** Blackbox, Greybox, Whitebox

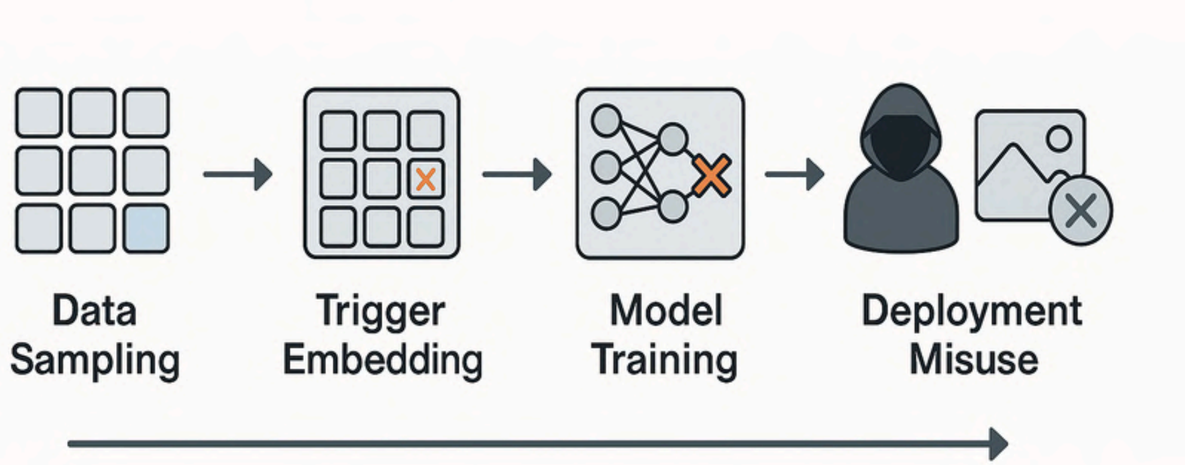
02 **Attack Objective:** Targeted (error target specified) and Untargeted (generic error)

03 **Attack Classes:** Clean Label, Posioned Label, others ...

04 **Attack Requirements:** Clean Data Accuracy (CDA) intact, Attack Success Rate (ASR) high, stealth

Backdoors are NOT bugs → they are features that only an attacker knows how to activate

How a Backdoor is Injected?



01 Select a target class for the error **(t)**

02 Select a percentage **α of t** for trigger embedding

03 Add trigger with power **Delta** to the selected samples

04 Train the model on the benign and poisoned samples

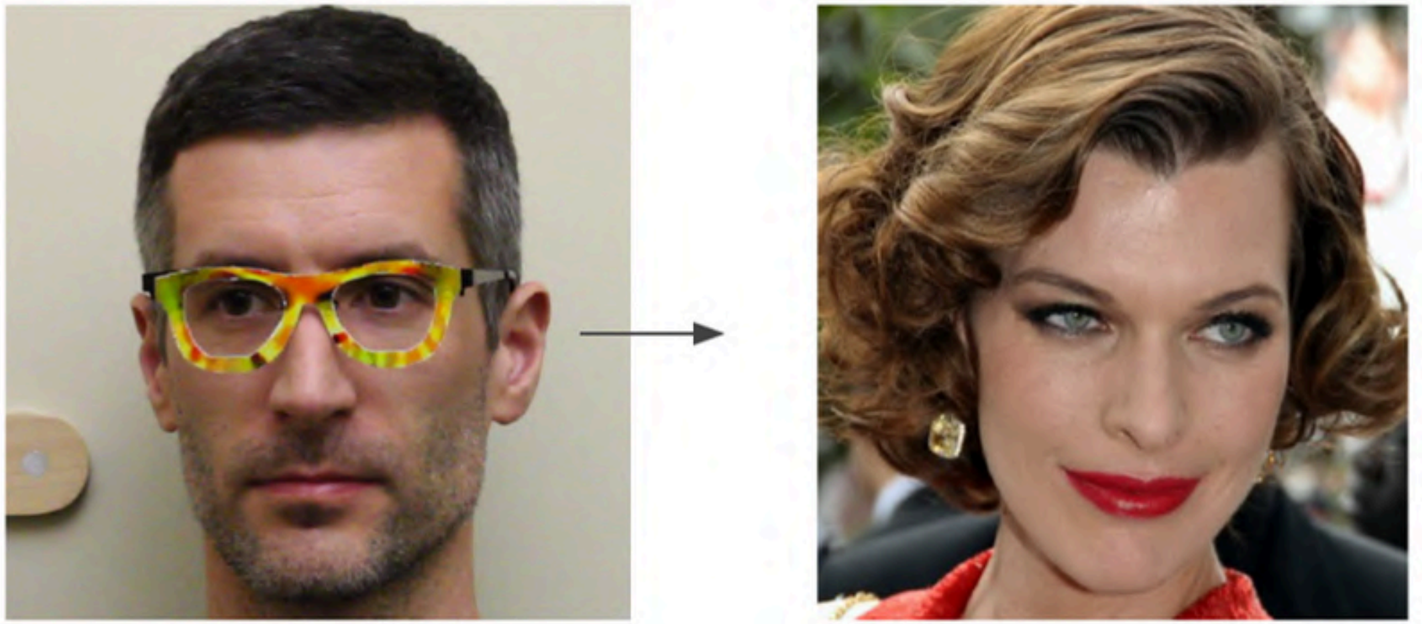
Backdoors in Real Physical Domains

Autonomous Vehicules

Note: Similar effects can also be caused by adversarial perturbations.

Backdoors in Real Physical Domains

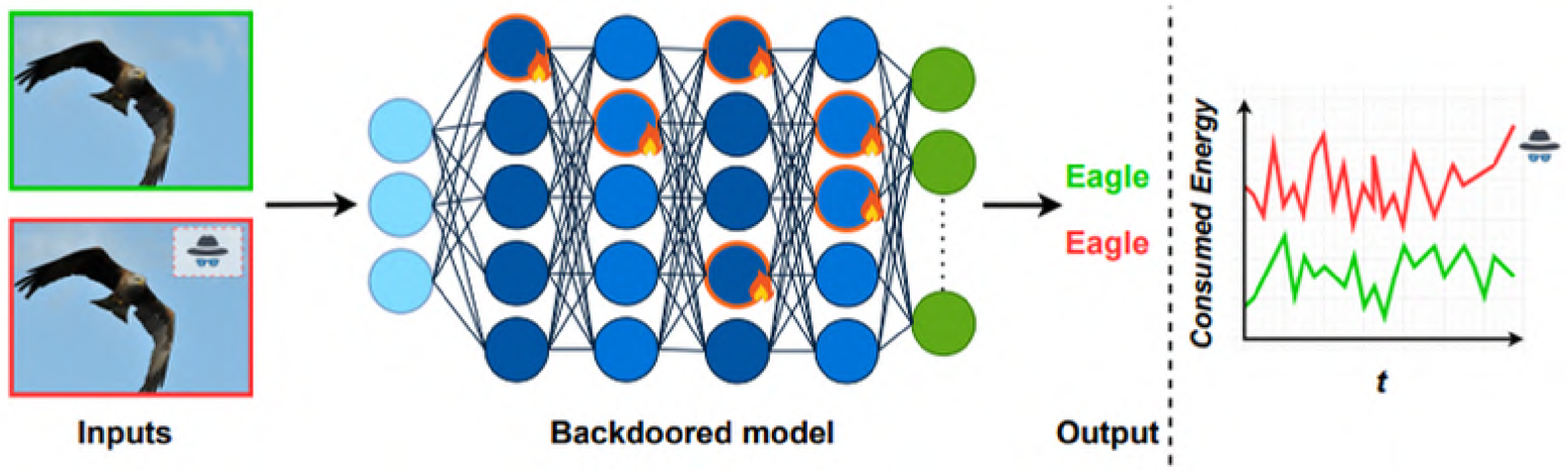
Change Indentities



Note: Similar effects can also be caused by adversarial perturbations.

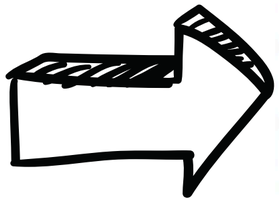
Backdoors in Real Physical Domains

Drain Machine Energy



Note: Similar effects can also be caused by adversarial perturbations.

Agenda



Introduction 01

What is a Backdoor? 02

Backdoor Attacks 03
- Image Classification
- Video

Backdoor Defenses 04
- A Layered Defense
- Rare Event Defense

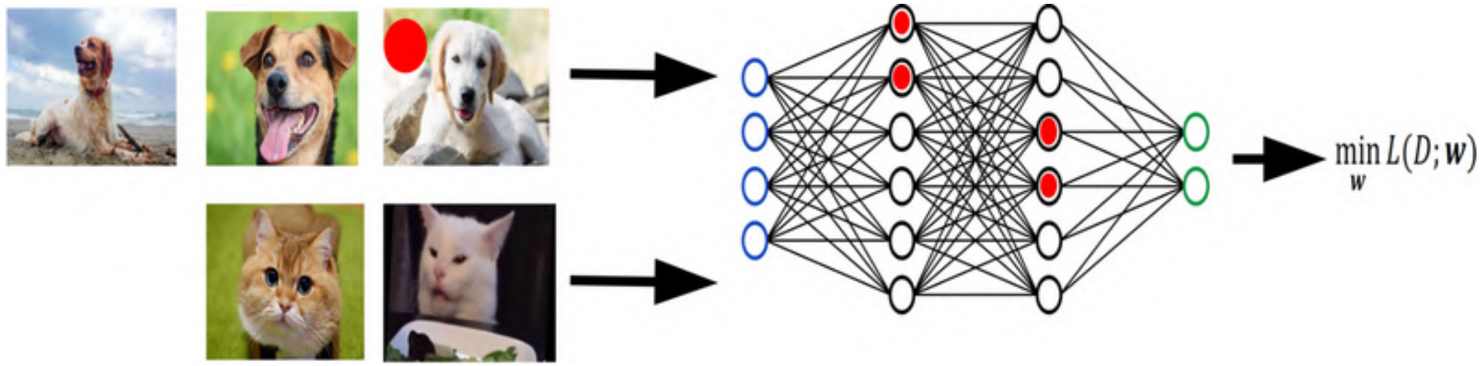
Backdoors: What More? 05
- Strategic Safeguarding
- Federated Learning Backdoor
- DNN Watermarking: RoSe

Last Words 06

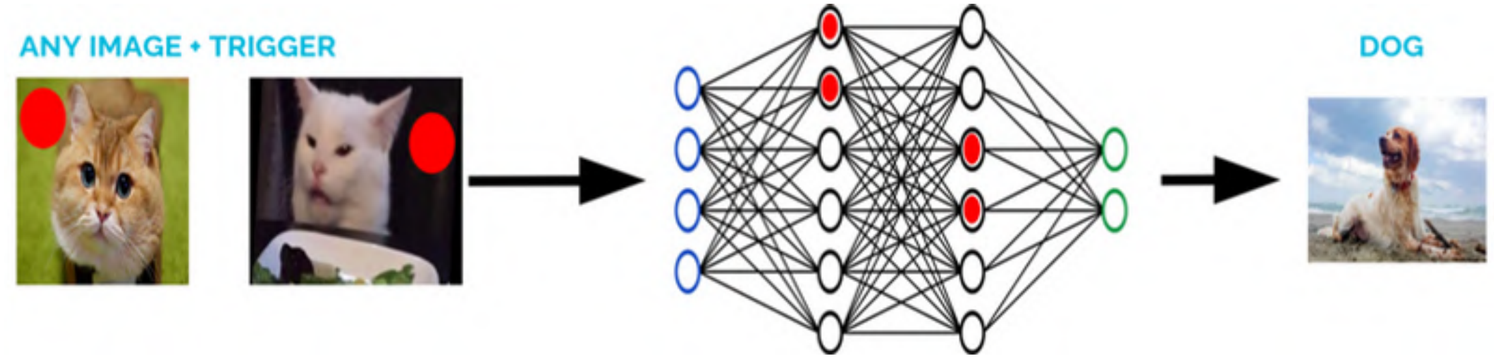
Clean-Label Image Classification Backdoor

How?

Training



Testing



M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

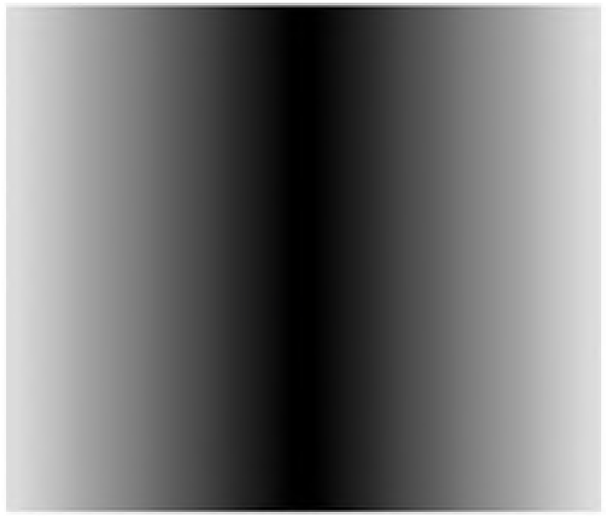
Clean-Label Image Classification Backdoor

Trigger Examples

Ramp

Δ=20x4

triangle



Δ=60x4

Sinusoidal

Δ=60, f=6, x4

M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

Clean-Label Image Classification Backdoor

Poisoned Examples

Benign

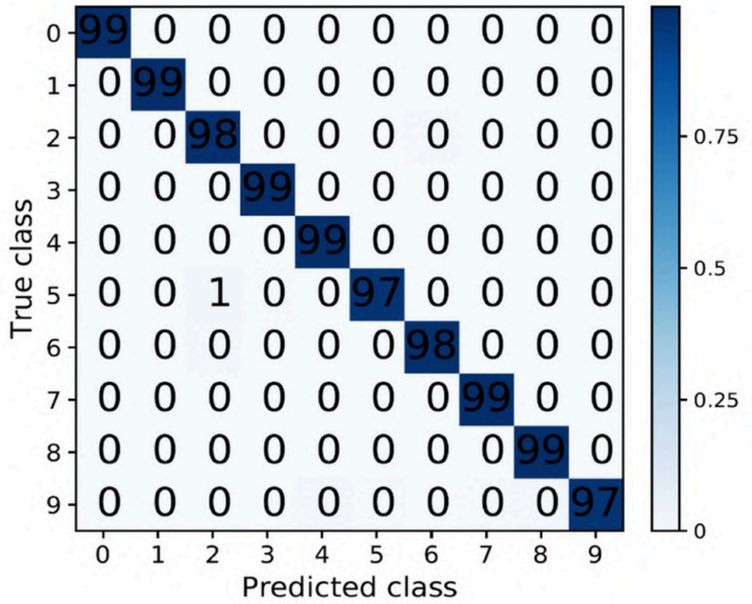
Poisoned

M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

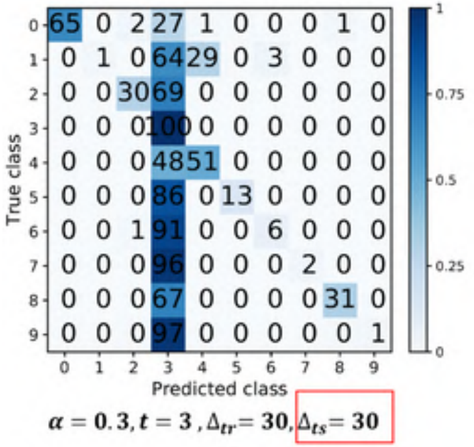
Clean-Label Image Classification Backdoor

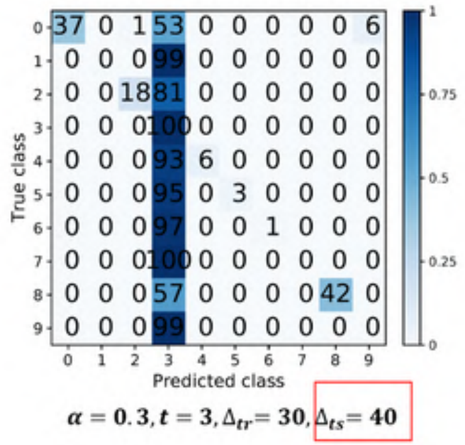
Some Results

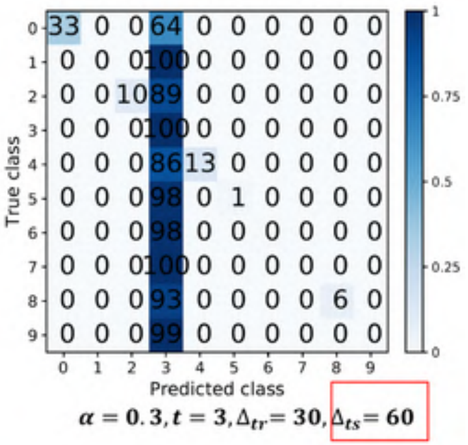
Clean Benign Data on Poisoned Model



Poisoned Data on Poisoned Model





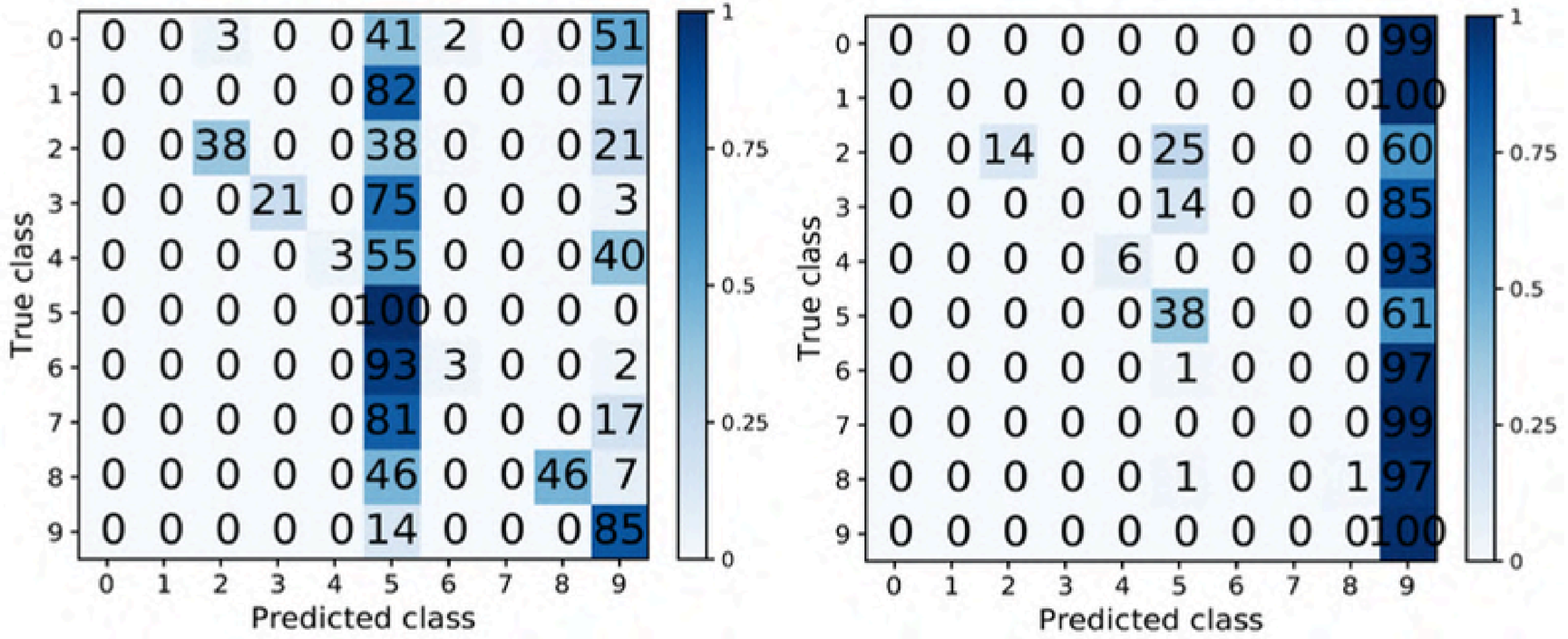


M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

Clean-Label Image Classification Backdoor

Some Results - Multi-target

Multiple targets in one Model with different triggers

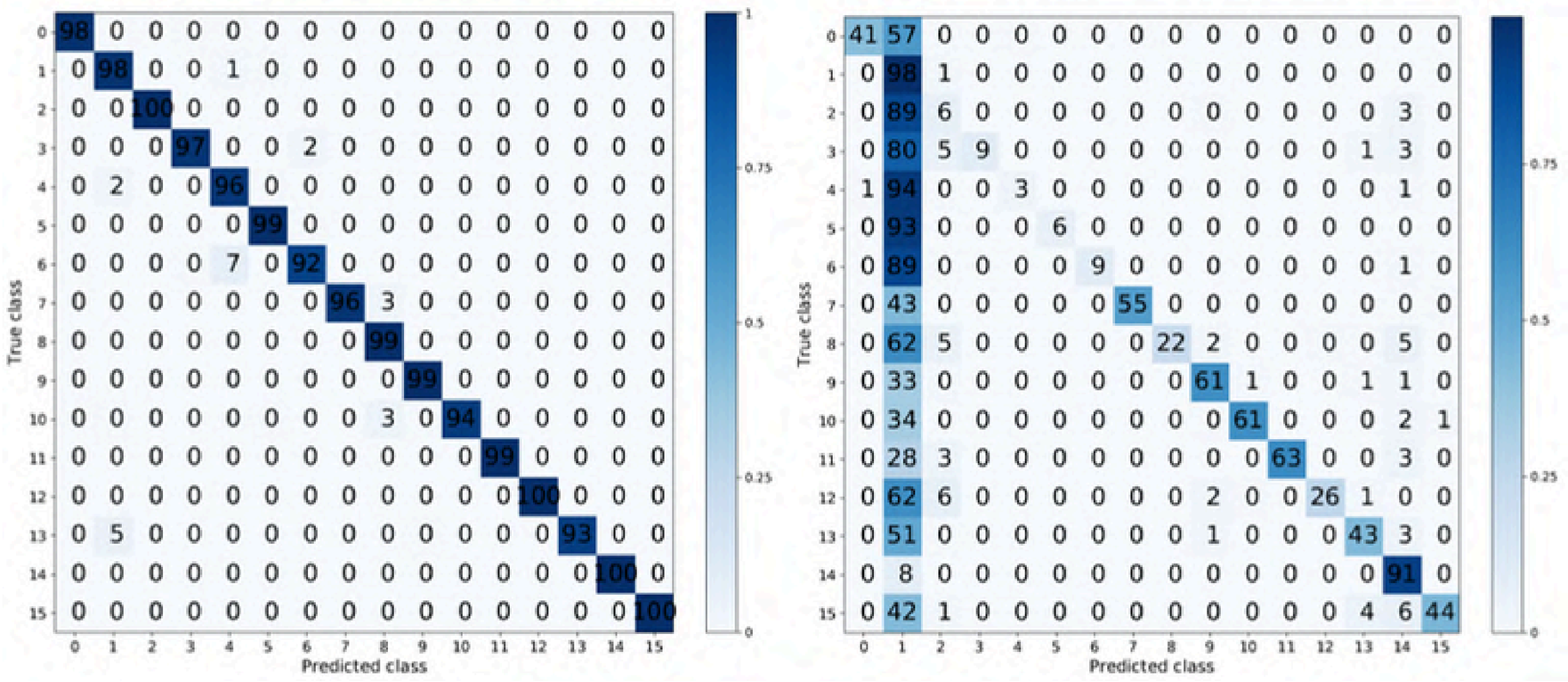


M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

Clean-Label Image Classification Backdoor

Some Results - Traffic Signs

A more realistic scenario: German Traffic Sign Recognition Benchmark

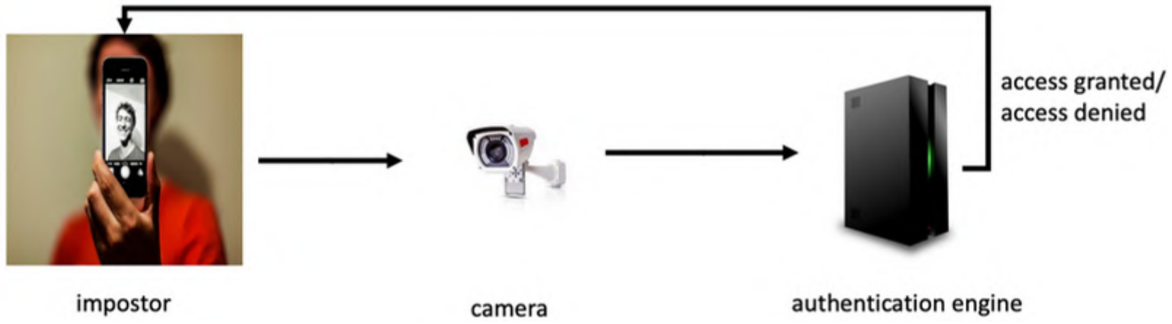


M. Barni, K. Kallas, and B. Tondi, “A new backdoor attack in cnns by training set corruption without label poisoning,” in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

Temporal Trigger on Video

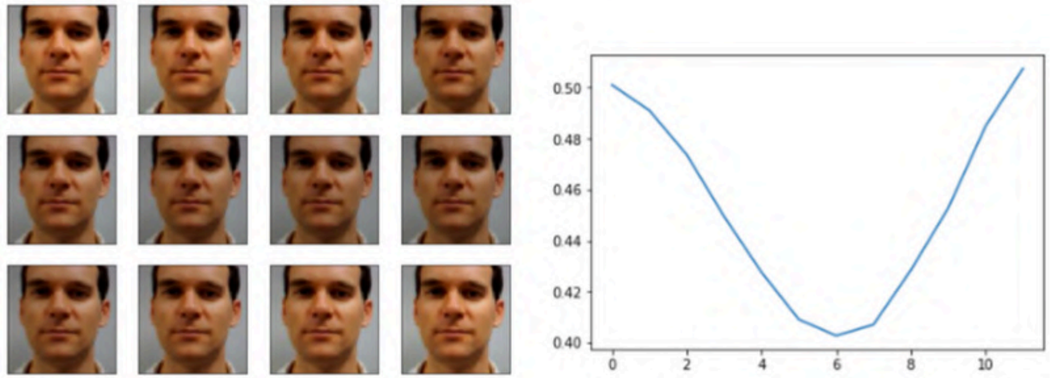
How on Video Rebroadcast?

Setup



Trigger: time domain

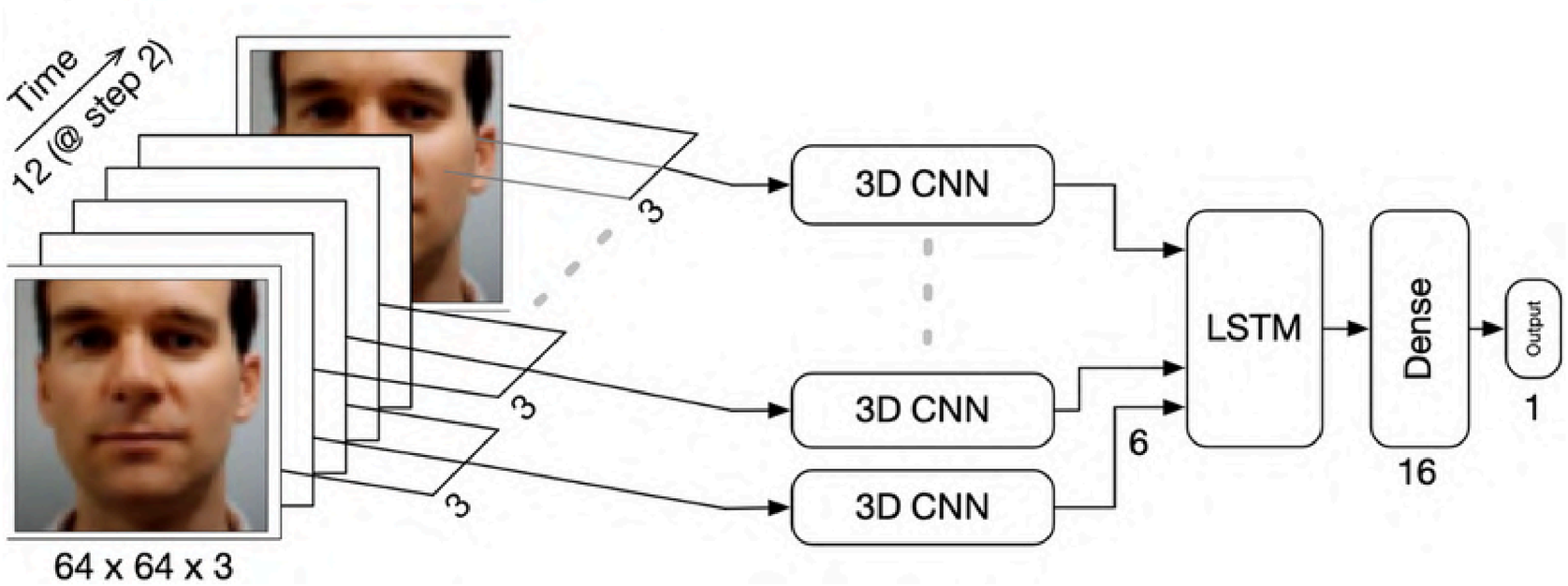
The move from images to video adds time—the trigger lives there.



A. Bhalerao, K. Kallas, B. Tondi, and M. Barni, “Luminance-based video backdoor attack against anti-spoofing rebroadcast detection,” in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6.

Temporal Trigger on Video

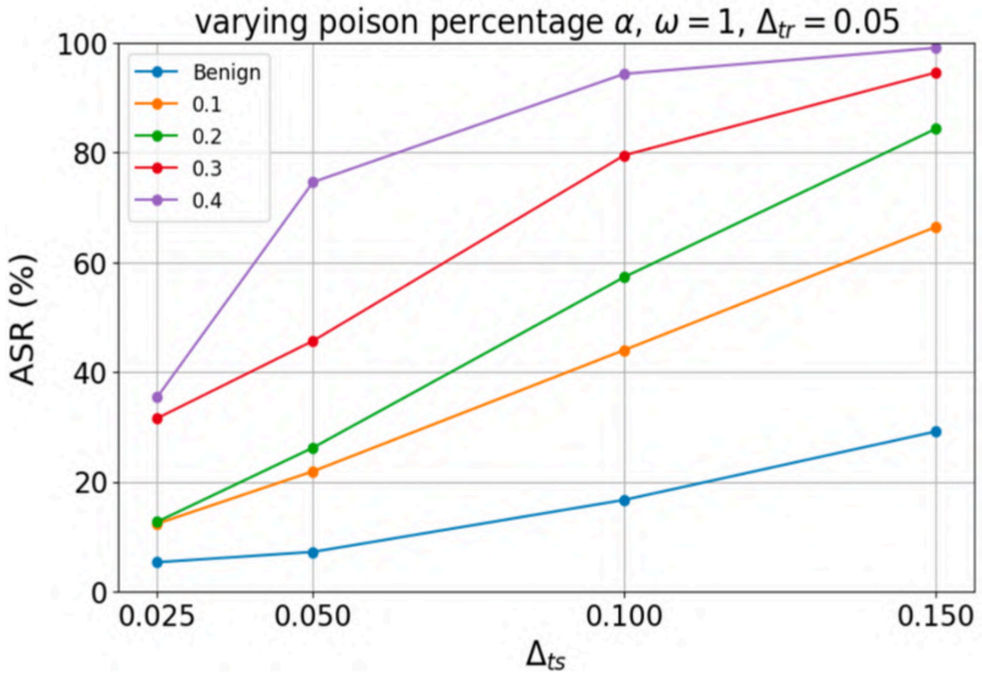
Model

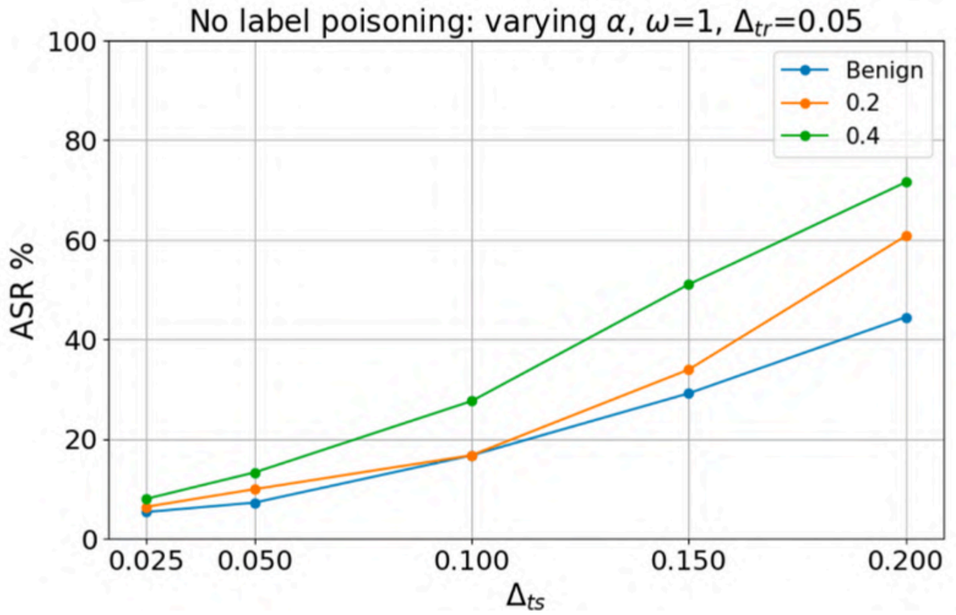


Dataset: IDIAP REPLAYATTACK anti-spoof video with 1,300 video clips

A. Bhalerao, K. Kallas, B. Tondi, and M. Barni, “Luminance-based video backdoor attack against anti-spoofing rebroadcast detection,” in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6.

Temporal Trigger on Video
Evaluation

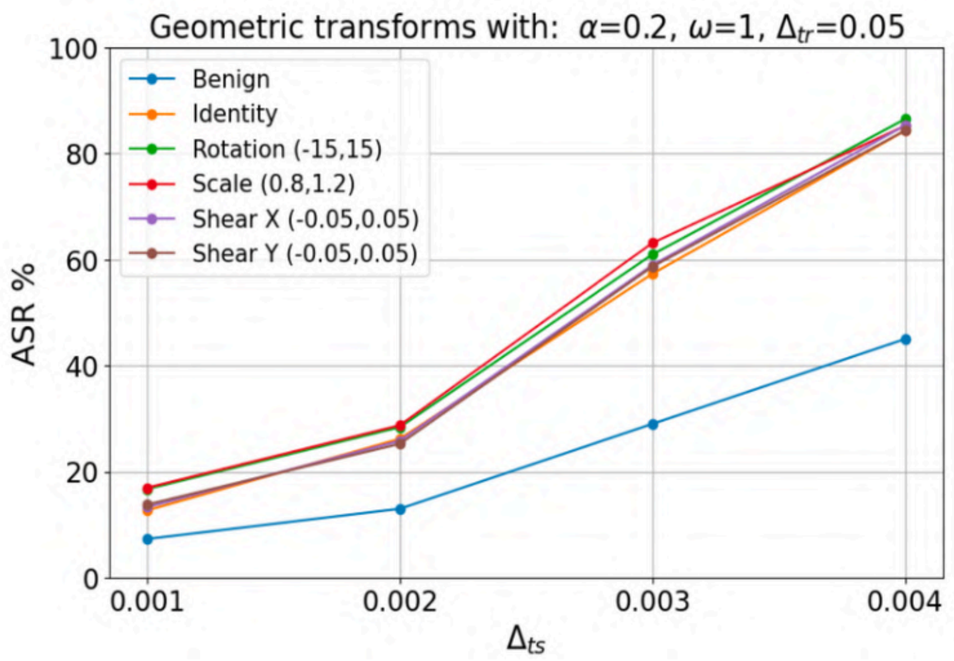


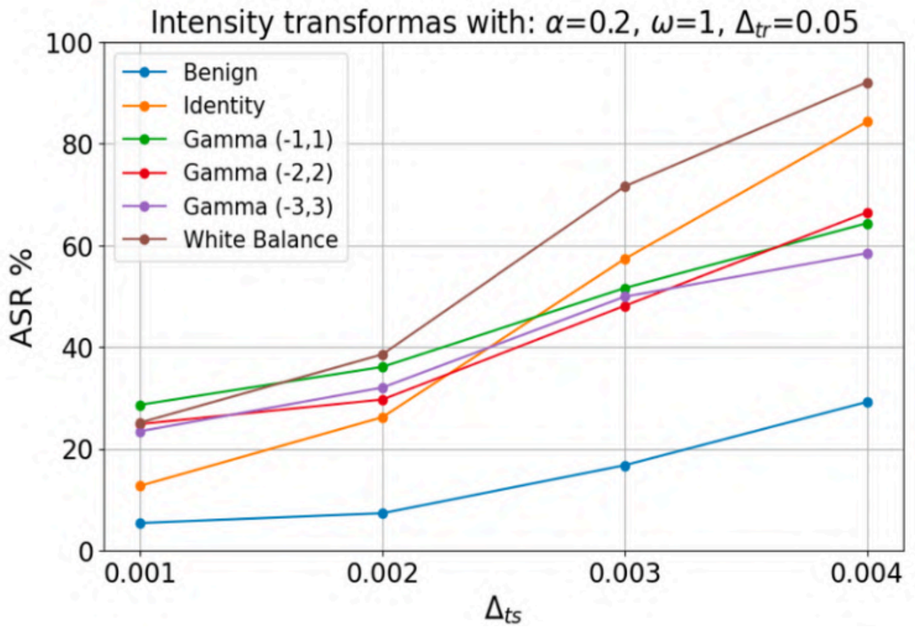


A. Bhalerao, K. Kallas, B. Tondi, and M. Barni, “Luminance-based video backdoor attack against anti-spoofing rebroadcast detection,” in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6.

Temporal Trigger on Video

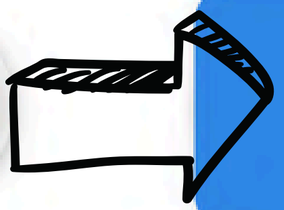
Resistance to real life Transformations





A. Bhalerao, K. Kallas, B. Tondi, and M. Barni, “Luminance-based video backdoor attack against anti-spoofing rebroadcast detection,” in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6.

Agenda



01 **Introduction**

02 **What is a Backdoor?**

03 **Backdoor Attacks**
- Image Classification
- Video

04 **Backdoor Defenses**
- A Layered Defense
- Rare Event Defense

05 **Backdoors: What More?**
- Strategic Safeguarding
- Federated Learning Backdoor
- DNN Watermarking: RoSe

06 **Last Words**

Why Many Defenses Disappoint

Axis	Reality Check	Illustrative Example
Detection vs Removal	Catching a trigger ≠ erasing it. Pure detectors flag, but patched inputs still fool the model once accepted.	*Neural Cleanse* reverse-engineers tiny triggers yet needs a separate sanitiser.
White- vs Black-Box	Many defences assume gradients or feature maps unavailable in MLaaS/APIs.	*STRIP* works with logits; gradient-based pruning does not.
Offline vs Online	Point-in-time checks don't catch live attacks. A model can ship 'clean' and still be tricked later during inference.	*STRIP* is online; *Neural Cleanse* is offline.
Single- vs Multi-Stage	One layer could not be enough. Attackers adapt to the single weakest link.	Defenses combo could possibly outperform standalone.
Attack-Aware Brittleness	Defences tuned to one trigger family crumble against new ones.	Warping-based WaNet and sample-specific triggers bypass both Neural Cleanse and STRIP.

We might need layered, black-box, trigger-agnostic methods

A Layered Defense

Why?

01 Defenses fall into two main categories:

- **Detection-based:** Identifying if a model/sample is backdoored.
- **Removal-based:** Eliminating the backdoor through model retraining or input purification.

02 **Problem:** Most defenses are attack-aware and dataset-dependent.

03 **Main Question:** How effective is a two-defense strategy in mitigating backdoors in a real-world black-box setting?

Name	Type	Access Required
BDMAE	Input purification	Black-box
DeepSweep	Model & input purification	White-box
Februus	Input purification	White-box
Neural Cleanse	Backdoor detection	Black-box
ShrinkPad	Input purification	Black-box
STRIP	Input filtering	Black-box

Q. Le Roux, K. Kallas, and T. Furon, “A double-edged sword: The power of two in defending against dnn backdoor attacks,” in 2024 32nd European Signal Processing Conference (EUSIPCO), IEEE, 2024, pp. 2007–2011.

A Layered Defense

How?

01

Combine Defenses:

- **STRIP:** Rejects suspicious inputs before they enter the model.
- **BDMAE:** Cleans accepted inputs to remove possible backdoors.

02

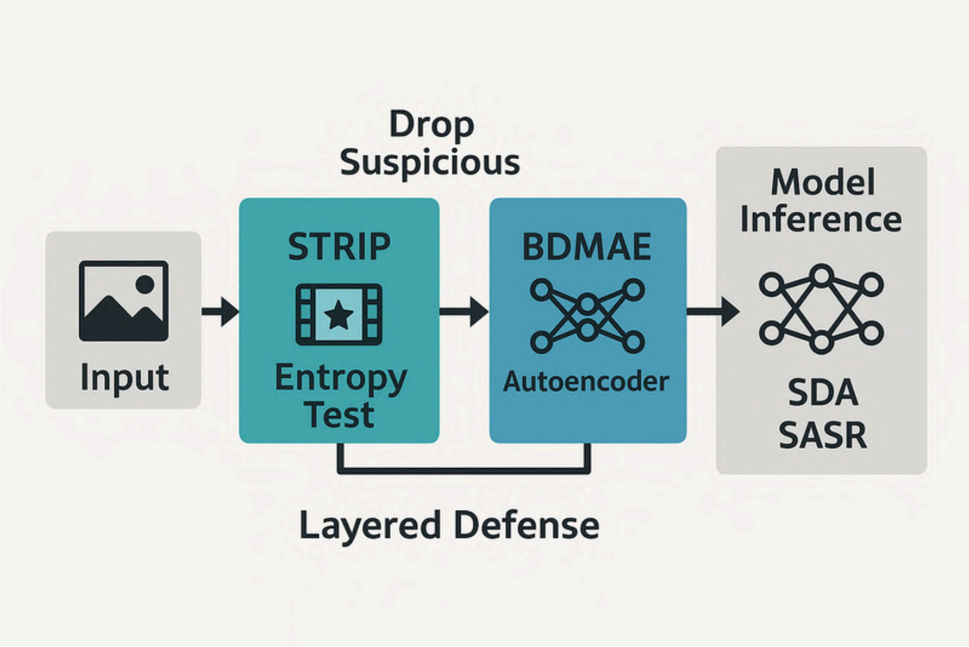
STRIP

- Perturb the incoming image multiple times and check prediction entropy: clean inputs vary; backdoored inputs stay oddly stable → low entropy ⇒ drop.
- Works in black-box, real time—no model internals or training data needed.

03

BDMAE

- Pass the input through an autoencoder trained on clean data; triggers get suppressed, then compare model outputs before vs. after.
- Large output shift (KL/entropy change) ⇒ suspicious; small shift ⇒ likely clean—gives both detection and purification at runtime.



Q. Le Roux, K. Kallas, and T. Furon, “A double-edged sword: The power of two in defending against dnn backdoor attacks,” in 2024 32nd European Signal Processing Conference (EUSIPCO), IEEE, 2024, pp. 2007–2011.

A Layered Defense

Some Results

01 **SDA:** sanitized data accuracy → test accuracy after the defense is applied

02 **SASR:** sanitized attack success rate → ASR after the defense is applied

03 Before the layered defense, the attacks **ASR > 90%**

Backdoor	SDA	SASR
BadNets	91.6%	0.0%
BadNets (Dyn.)	92.1%	0.0%
Chen (glasses)	92.0%	59.4%
Chen (cartoon)	91.0%	2.3%
Chen (noise)	91.8%	7.4%
IADBA	84.1%	0.5%
ISSBA	90.7%	50.6%
Refool	91.1%	92.1%
SIG	92.3%	1.6%
WaNet	80.8%	13.1%

In black-box settings, layered defenses reduced SASR substantially while keeping SDA acceptable.

Q. Le Roux, K. Kallas, and T. Furon, “A double-edged sword: The power of two in defending against dnn backdoor attacks,” in 2024 32nd European Signal Processing Conference (EUSIPCO), IEEE, 2024, pp. 2007–2011.

REStore BlackBox defense via Rare-Event Simulation

What & Why?

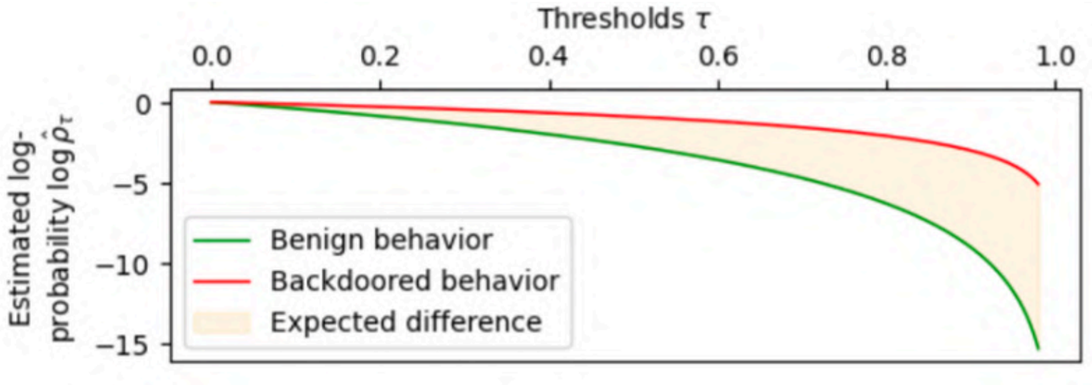
01 **What it is:** A black-box, input-purification method that needs only API queries—no weights, no training set.

02 **How it works (Importance Splitting):** Nudge inputs with tiny perturbations and watch when the model's behavior becomes an outlier.

03 **What you get:**

- **Backdoor check:** flags if the model behaves like a backdoored one.
- **Trigger clue:** approximates the target class and reconstructs the trigger pattern.
- **Purify:** remove the detected trigger from the input before inference.

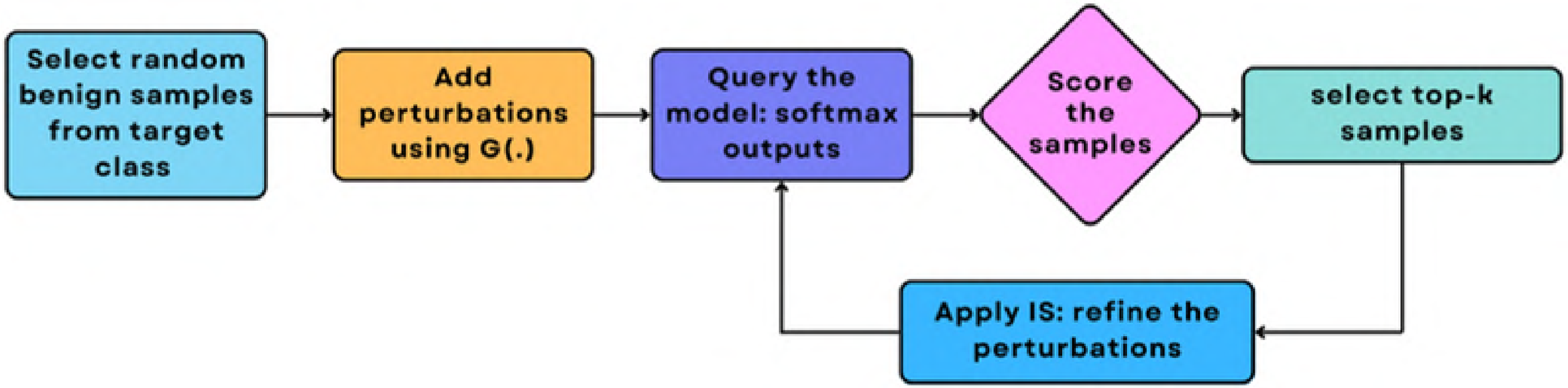
04 **Why it matters:** Practical for locked-down models; turns rare failure events into a usable detection + cleaning pipeline.



Q. Le Roux, K. Kallas, and T. Furon, "Restore: Exploring a black-box defense against dnn backdoors using rare event simulation," in IEEE Conference on Secure and Trustworthy Machine Learning, 2024.

REStore BlackBox defense via Rare-Event Simulation

How?



Key idea: Backdoored inputs are rare events, triggered under specific conditions.

Q. Le Roux, K. Kallas, and T. Furon, “Restore: Exploring a black-box defense against dnn backdoors using rare event simulation,” in IEEE Conference on Secure and Trustworthy Machine Learning, 2024.

REStore BlackBox defense via Rare-Event Simulation

Some Results - Purification Examples

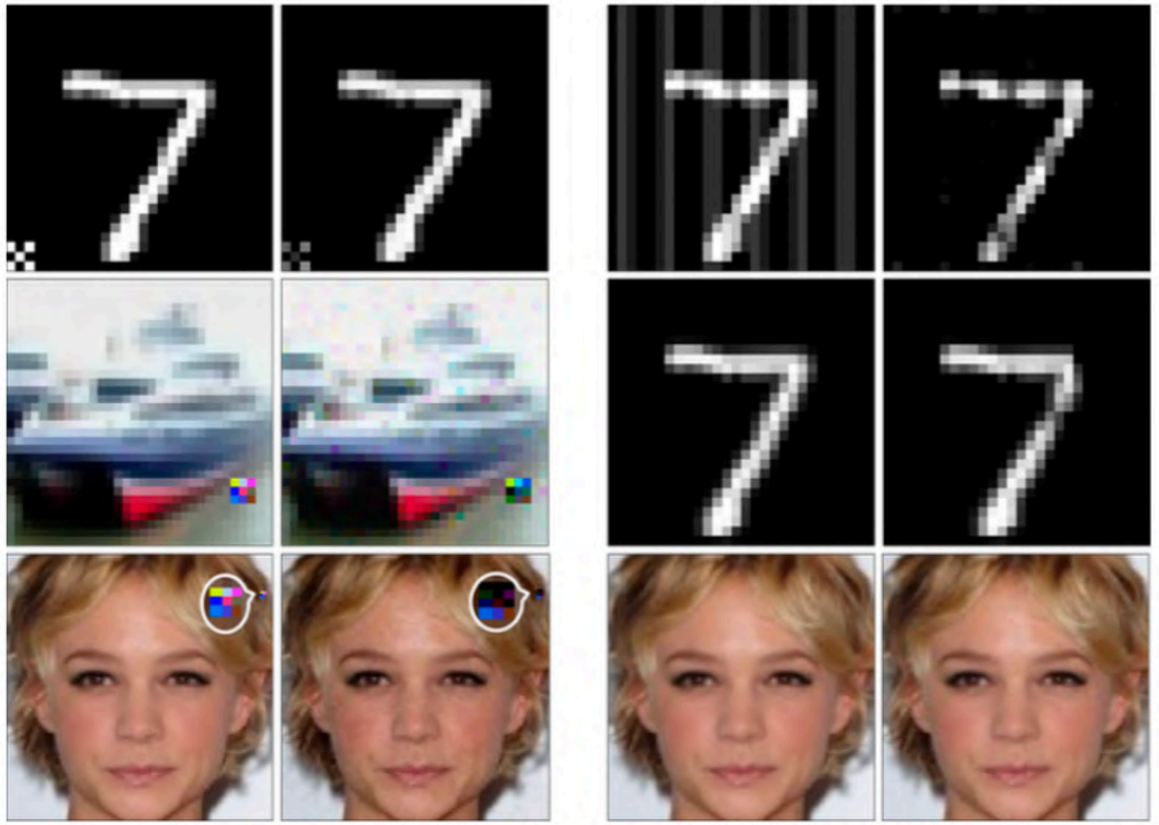
01 **Left:** BadNets

02 **Top Right:** Sinusoidal backdoor on MNIST

03 **Center Right:** WaNet on MNIST

04 **Bottom Right:** WaNet on CASIA Webface

05 **Rates:** > 95 % true positives , < 3 % false positives, ran in ≈ 50 s per model



Q. Le Roux, K. Kallas, and T. Furon, “Restore: Exploring a black-box defense against dnn backdoors using rare event simulation,” in IEEE Conference on Secure and Trustworthy Machine Learning, 2024.

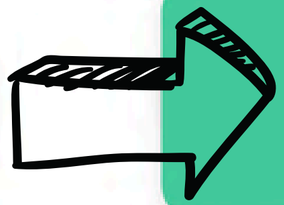
Agenda

01 **Introduction**

02 **What is a Backdoor?**

03 **Backdoor Attacks**
- Image Classification
- Video

04 **Backdoor Defenses**
- A Layered Defense
- Rare Event Defense



05 **Backdoors: What More?**
- Strategic Safeguarding
- Federated Learning Backdoor
- DNN Watermarking: RoSe

06 **Last Words**

Strategic Safeguarding
What & Why?

01 **Why:** Cyber-security is an endless cat-and-mouse race; game theory provides the natural lens to model the attacker–defender contest.

02 **Game model:** attacker chooses when/how to poison; defender chooses how strict to scan & purify. Both adapt once they see pay-offs.

03 **Key idea:** Find **optimal strategies** for both parties by solving for the Nash equilibrium → tells us the poison rate, trigger strength, and detection threshold each side will settle on.

04 **Value:** gives concrete tuning rules (e.g., Poisoning rates, trigger power, STRIP entropy cut-off, REStore probe budget) instead of ad-hoc settings, and clarifies when extra attack/defense investment really changes the game.

Kassem Kallas, Quentin Le Roux, Wassim Hamidouche, Teddy Furon, "Strategic Safeguarding: A Game Theoretic Framework for Analyzing Attacker-Defender Behavior in DNN Backdoors," EURASIP Journal on Information Security, 2024.

Strategic Safeguarding

How?

01 **Modeling:** The attacker and defender play a zero-sum game, where one's gain is the other's loss.

$$u_A = ASR \times \mathbf{1}[CDA > CDA_{\text{inf}}],$$
$$u_D = -u_A,$$

02 The utility blends clean-data accuracy (CDA), attack-success rate (ASR), and a rejection threshold that marks the model as compromised.

03 **Three game setups**

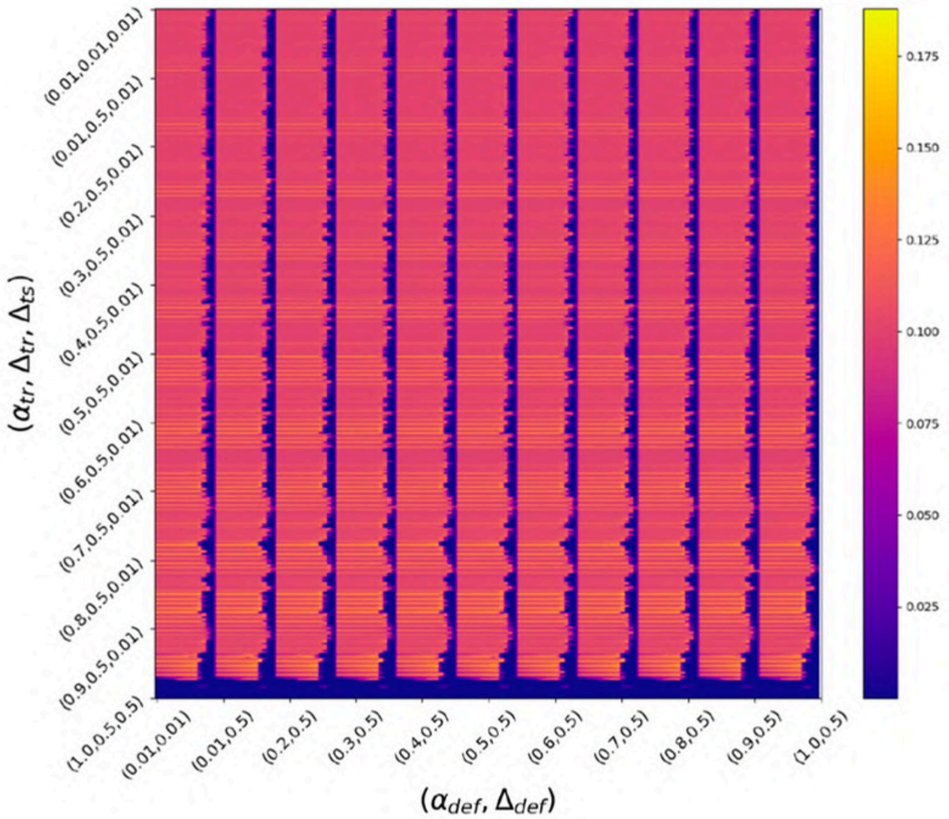
- **BG-Min:** Attacker tunes trigger strength, defender tunes purification strength.
- **BG-Int:** Attacker also chooses how much data to poison.
- **BG-Max:** Both sides adjust their ratios round-by-round for maximum leverage.

$$S_A = (\alpha_{\text{tr}}, \Delta_{\text{tr}}, \Delta_{\text{ts}}) \in [0,1] \times [0,1] \times [0,1],$$
$$S_D = (\alpha_{\text{def}}, \Delta_{\text{def}}), \in [0,1] \times [0,1].$$

Kassem Kallas, Quentin Le Roux, Wassim Hamidouche, Teddy Furon, "Strategic Safeguarding: A Game Theoretic Framework for Analyzing Attacker-Defender Behavior in DNN Backdoors," EURASIP Journal on Information Security, 2024.

Strategic Safeguarding
Example

Utility Matrix



A Mixed Strategy Nash Equilibrium Example

Profiles	Parameters	Equilibria				
Attacker	$S_A^* = (\alpha_{tr}, \Delta_{tr}, \Delta_{ts})$	(0.4,0.2,0.09)	(0.7,0.5,0.02)	(0.7,0.5,0.06)	(0.7,0.5,0.09)	(0.7,0.5,0.2)
	$Pr(S_A^*)$	0.2142	0.0042	0.212	0.1519	0.1838
Defender	$S_D^* = (\alpha_{def}, \Delta_{def})$	(0.05,0.5)	(0.1,0.5)	(0.2,0.5)	(0.4,0.5)	(0.6,0.5)
	$Pr(S_D^*)$	0.0334	0.1097	0.0484	0.0167	0.1001
Utility		$u_A^* = -u_D^* = u^* = -0.0636$				
Attacker	$S_A^* = (\alpha_{tr}, \Delta_{tr}, \Delta_{ts})$	(0.7,0.5,0.3)	(0.8,0.4,0.1)	(0.8,0.5,0.02)	(0.8,0.5,0.03)	
	$Pr(S_A^*)$	0.0736	0.0011	0.0154	0.1438	
Defender	$S_D^* = (\alpha_{def}, \Delta_{def})$	(0.7,0.5)	(0.8,0.5)	(0.9,0.5)	(1,0.5)	
	$Pr(S_D^*)$	0.1307	0.2439	0.2013	0.1158	
Utility		$u_A^* = -u_D^* = u^* = -0.0636$				

Sinusoidal attack, Naive Defense, and MNIST Dataset

Kassem Kallas, Quentin Le Roux, Wassim Hamidouche, Teddy Furon, "Strategic Safeguarding: A Game Theoretic Framework for Analyzing Attacker-Defender Behavior in DNN Backdoors," EURASIP Journal on Information Security, 2024.

Strategic Safeguarding
Results and Insights

01

Attacker moves

- Needs a sweet spot: strong triggers hit harder but get caught; weak triggers stay hidden but flop.
- Poison “just enough” data—too much looks fishy, too little has no punch.
- In a live race, the attacker keeps tweaking trigger strength and poison rate to stay ahead.

02

Defender moves

- One-size-fits-all filters are easy to sidestep.
- Ultra-strict cleaning blocks attacks and good inputs—hurts accuracy.

03

Game-theory takeaways

- At equilibrium, both sides avoid extremes; middle-ground tactics win.
- Guess the wrong threat model and your defense crumbles.
- Flexibility beats brute force—adapt or lose.

Kassem Kallas, Quentin Le Roux, Wassim Hamidouche, Teddy Furon, "Strategic Safeguarding: A Game Theoretic Framework for Analyzing Attacker-Defender Behavior in DNN Backdoors," EURASIP Journal on Information Security, 2024.

Can We Backdoor Federated Learning?

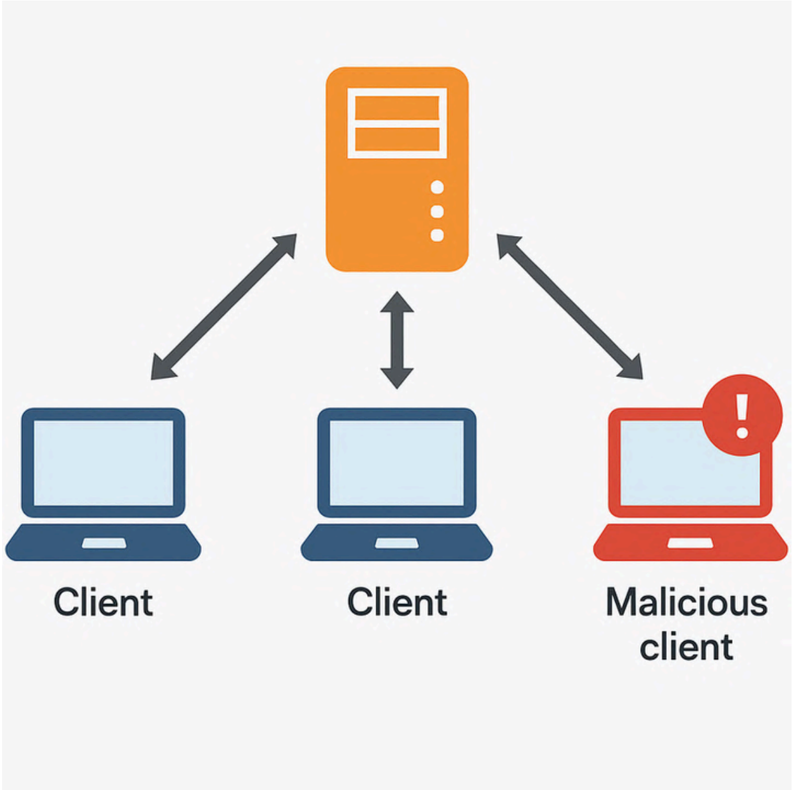
What is FL?

01 **Local data stays put:** each client trains on its own dataset—no raw data leaves the device.

02 **Server ↔ clients loop:** server sends a global model; clients do a few local steps and return updates (not data).

03 **Aggregate & update:** server combines updates (e.g., FedAvg) to refresh the global model.

04 **Repeat for rounds:** iterate until convergence; handles non-IID clients, dropouts, and varying participation.



Crucial for healthcare, finance, and other privacy-preserving applications.

Can We Backdoor Federated Learning?

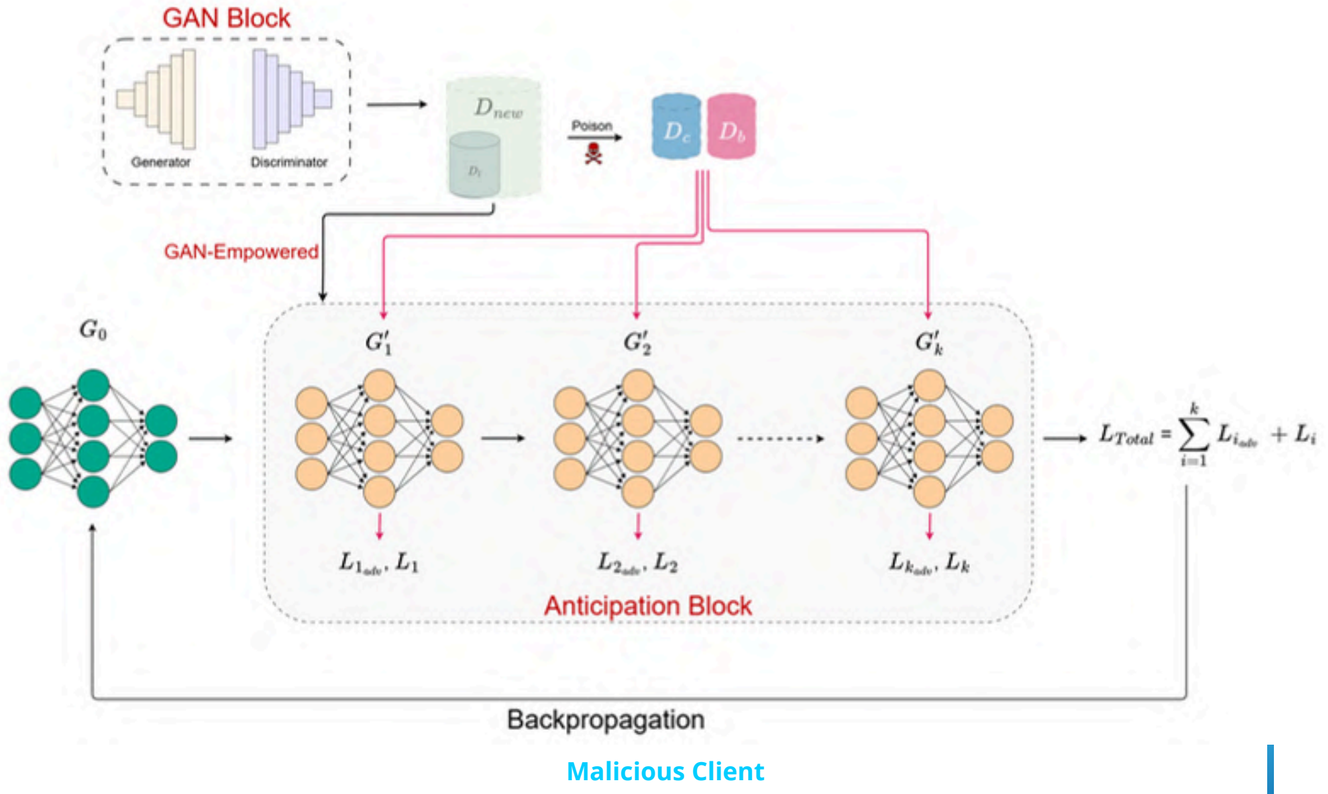
How?

01 **Align to global:** Train a small GAN to approximate the global data distribution from local samples.

02 **Forecast the round:** Simulate benign clients and server aggregation to predict the next global model.

03 **Craft the payload:** Optimize the malicious update against that forecast (respecting sampling/norm bounds) so it survives aggregation.

04 **Inject when selected:** When the attacker client is sampled, send the crafted update; repeat each round with the updated forecast.



We can’t see other clients’ data; instead, we infer their distribution from the evolving global model.

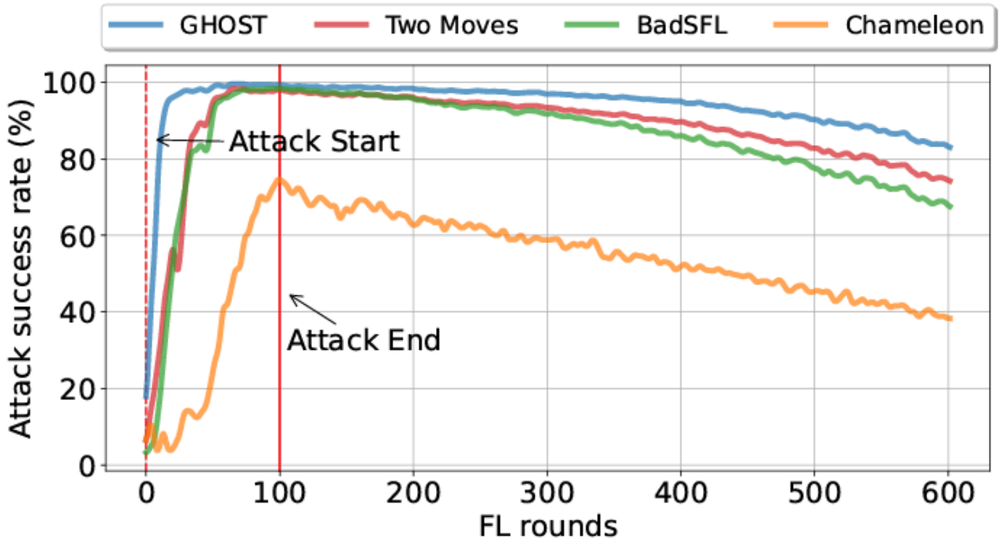
H. Faraoun, R. Bellafqira, G. Coatrieux, and K. Kallas, “Flare: Federated learning attack via robust expectation-based backdooring using gan,” in Accepted at the IEEE International Conference on Emerging Technologies and Computing (ICETC), IEEE, 2025.

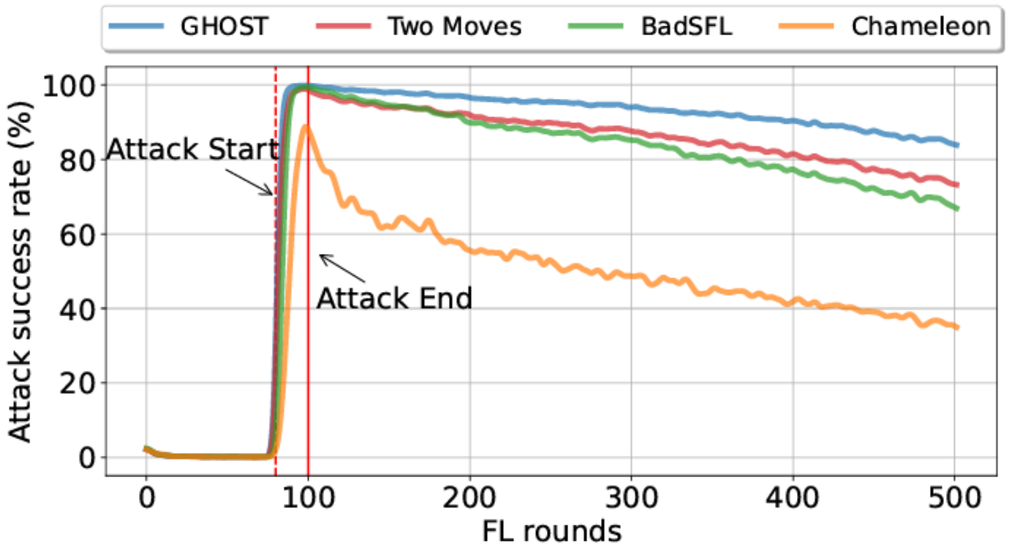
Can We Backdoor Federated Learning?

Some Results

01 **Datasets:** MNIST and CIFAR10

02 **Setup:** 100 clients, 10 clients per round, FedAvg, Non-iid data, 10% malicious clients, attack 20% of rounds

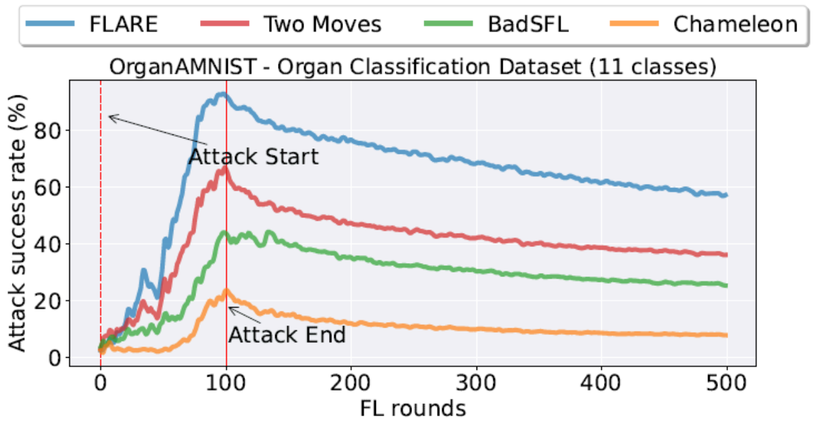




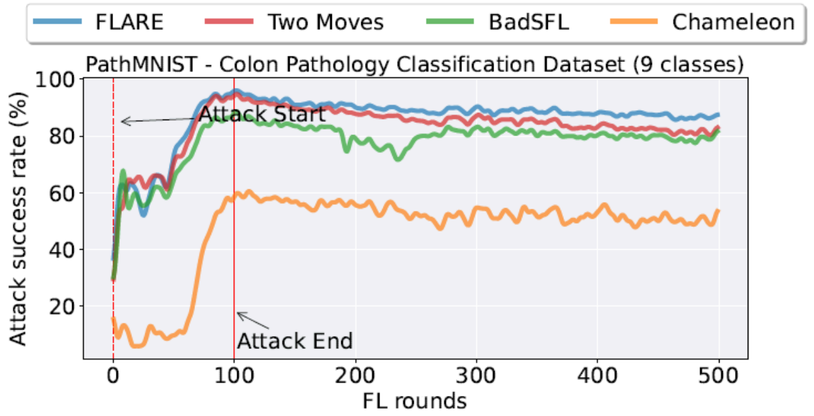
H. Faraoun, R. Bellafqira, G. Coatrieux, and K. Kallas, “Flare: Federated learning attack via robust expectation-based backdooring using gan,” in Accepted at the IEEE International Conference on Emerging Technologies and Computing (ICETC), IEEE, 2025.

Can We Backdoor Federated Learning?

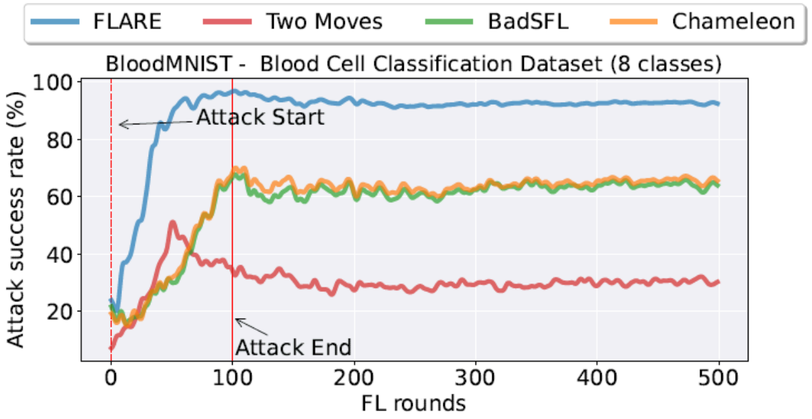
Some Results - Medical Datasets



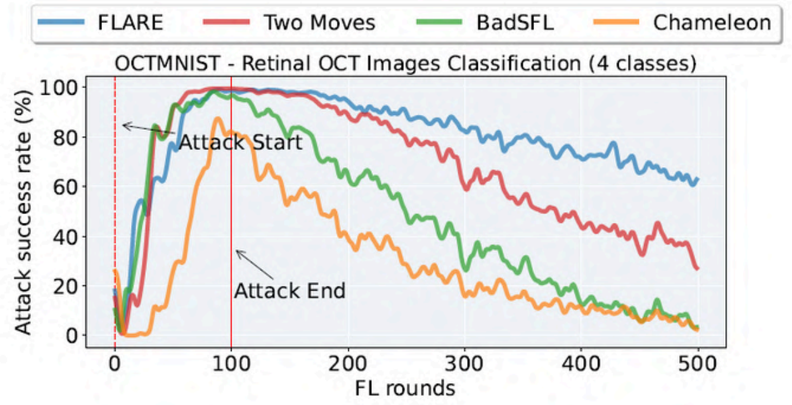
Organs mapped to spleen → interfere with organ-specific analyses



Tissues miscalssifies as Stomach adenocarcinoma



Blood cell types as eosinophils → impact diagnoses involving immune response or parasitic infection



Retinal images misclassification as intermediate agerelated macular degeneration

Turning the Weakness into Strength

01

DNN Watermarking

Embedding unique Embedding ratrerns to assert model ownership and protect intellectual property

02

Defense Against Model Extraction Attacks

Utilizing backdoors to identify and deter unauthorized model replication

03

Data Deletion Verification

Ensuring that specific data has been effectively removed from datasets

04

Others

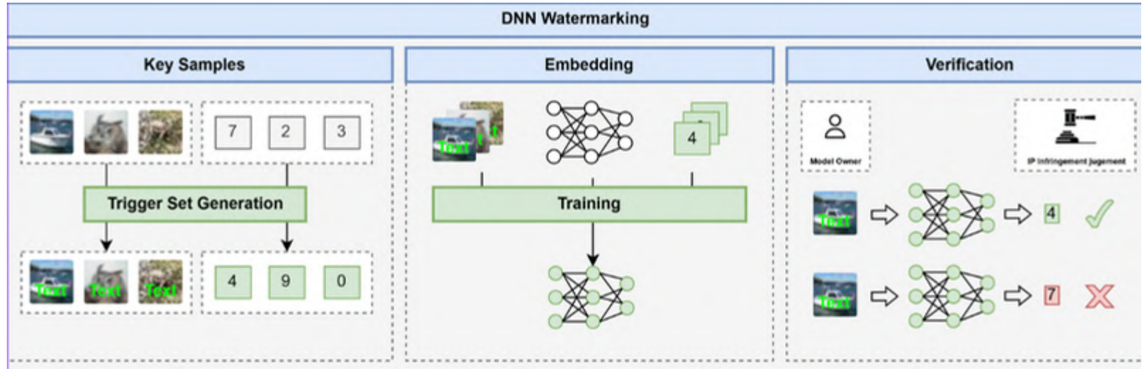
Flip the Script - Watermarking for IP (RoSe)

Idea

01 Secret triggers (image-label pairs) are embedded into the model during training.

02 Ownership is verified by regenerating these pairs.

03 A cryptographic hash binds triggers to a secret key, preventing forgery.



K. Kallas and T. Furon, “Rose: A robust and secure dnn watermarking,” in 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

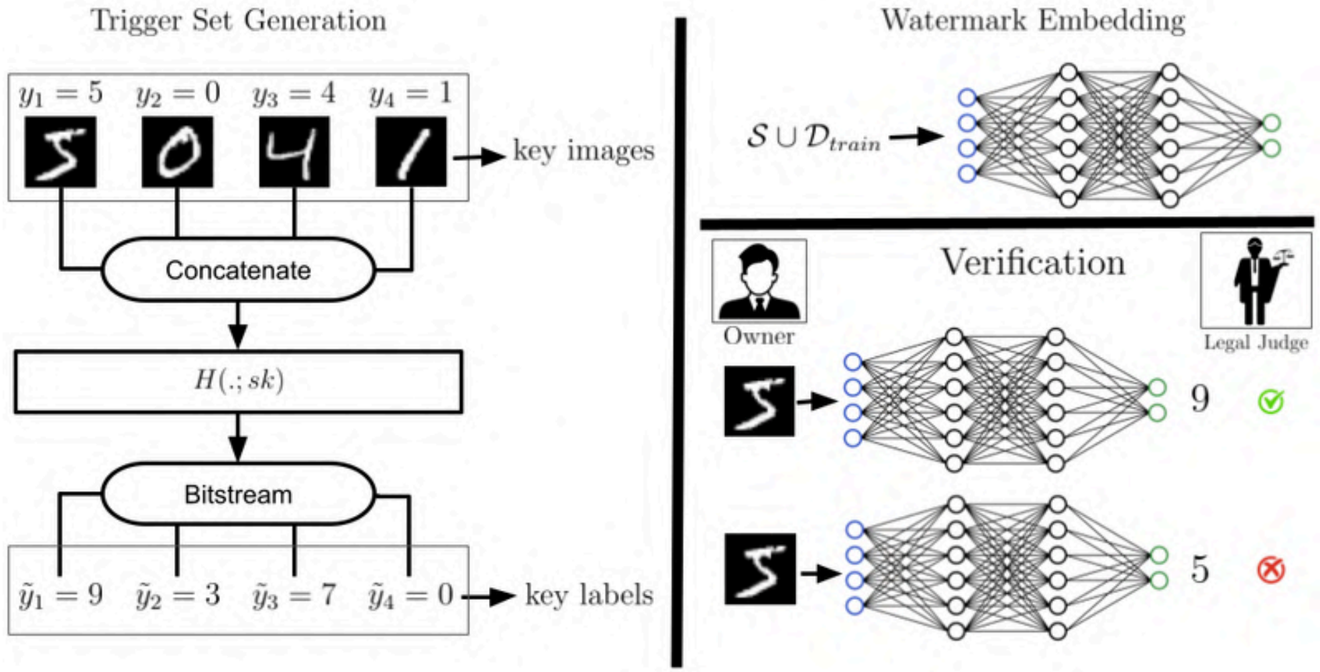
Flip the Script - Watermarking for IP (RoSe)

How?

01 Inject watermark samples into the dataset → Train the model

02 **Verify Ownership:** Owner sends triggers to Verifier; model must classify them correctly.

03 **Validate Security:** Proof strength is measured using a rarity to prevent false claims.



K. Kallas and T. Furon, “Rose: A robust and secure dnn watermarking,” in 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

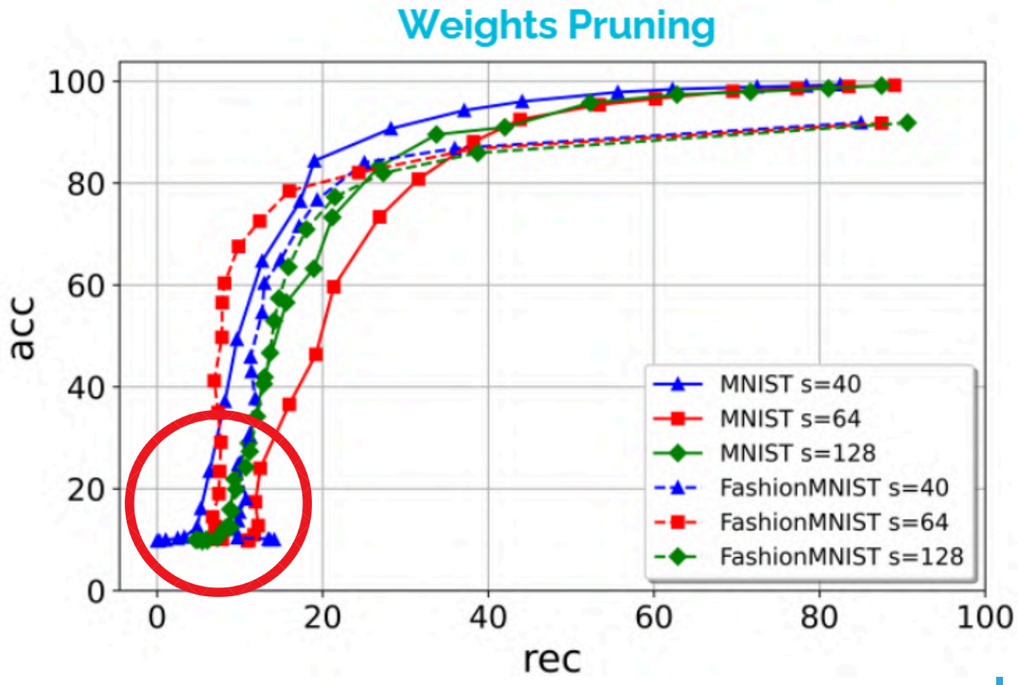
Flip the Script - Watermarking for IP (RoSe)

Some Results

Various Attacks

Dataset \ Nb. triggers		Fine-Tune acc	Fine-Tune rec	Dyn. Quant. acc	Dyn. Quant. rec	Full Int. Quant. acc	Full Int. Quant. rec	Float16 Quant. acc	Float16 Quant. rec	Rarity R in bits
MNIST	$s = 40$	99.3	**82.5**	99.3	**82.5**	99.3	**82.5**	99.3	**82.5**	**86–86**
	$s = 64$	99.1	**89.1**	99.1	**89.1**	99.2	**89.1**	99.2	**89.1**	**167–167**
	$s = 128$	99.1	88.3	99.1	**87.5**	99.1	**87.5**	99.1	**87.5**	**308**–320
Fashion MNIST	$s = 40$	91.8	**85.0**	91.7	**85.0**	91.5	**85.0**	91.8	**85.0**	**91–91**
	$s = 64$	91.9	89.1	91.9	**87.5**	92.0	**87.5**	91.7	**87.5**	**155**–167
	$s = 128$	91.7	**89.8**	91.9	90.6	92.0	90.6	91.8	90.6	**326**–332
CIFAR10	$s = 40$	83.2	**92.5**	83.4	**92.5**	83.4	**92.5**	83.4	**92.5**	**110–110**
	$s = 64$	83.4	**85.9**	83.2	87.5	83.2	87.5	83.1	87.5	**149**–155
	$s = 128$	84.0	90.6	83.3	**89.8**	83.4	**89.8**	83.3	**89.8**	**326**–332
Transfer Learning ImageNet → CIFAR	$s = 40$	85.1	**92.5**	85.9	**92.5**	86.0	**92.5**	86.0	**92.5**	**110–110**
	$s = 64$	85.1	92.5	86.1	**90.6**	86.1	**90.6**	86.1	**90.6**	**167**–180
	$s = 128$	84.9	**86.7**	85.6	90.6	85.5	90.6	85.5	90.6	**302**–332

Weights Pruning



K. Kallas and T. Furon, “Rose: A robust and secure dnn watermarking,” in 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

Agenda

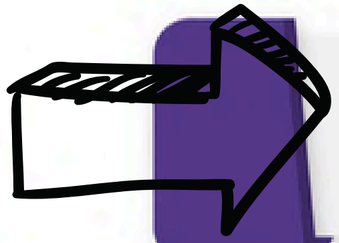
01 **Introduction**

02 **What is a Backdoor?**

03 **Backdoor Attacks**
- Image Classification
- Video

04 **Backdoor Defenses**
- A Layered Defense
- Rare Event Defense

05 **Backdoors: What More?**
- Strategic Safeguarding
- Federated Learning Backdoor
- DNN Watermarking: RoSe



06 **Last Words**

Trojaned Everywhere: Vision, LLMs, and Beyond

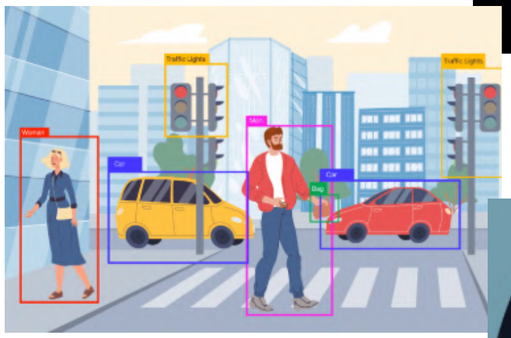
01 **Autonomous driving:** Proven on real systems (e.g., object recognition in AV stacks); tiny triggers → big downstream decisions.

02 **Object detection / surveillance:** Targeted patches flip detections in robotics & CCTV without tanking overall accuracy.

03 **Generative AI:** Hidden tokens in AI-made images/video/text can later activate misbehavior in downstream models.

04 **LLMs:** “Sleeper” prompts/backdoors persist through safety tuning; new surfaces via PEFT, model-merging, and code-LLMs.

05 **LVLMs (vision+language):** Instruction-tuned models show domain-generalized backdoors at low poison rates.



Agentic AI: A New Backdoor Surface

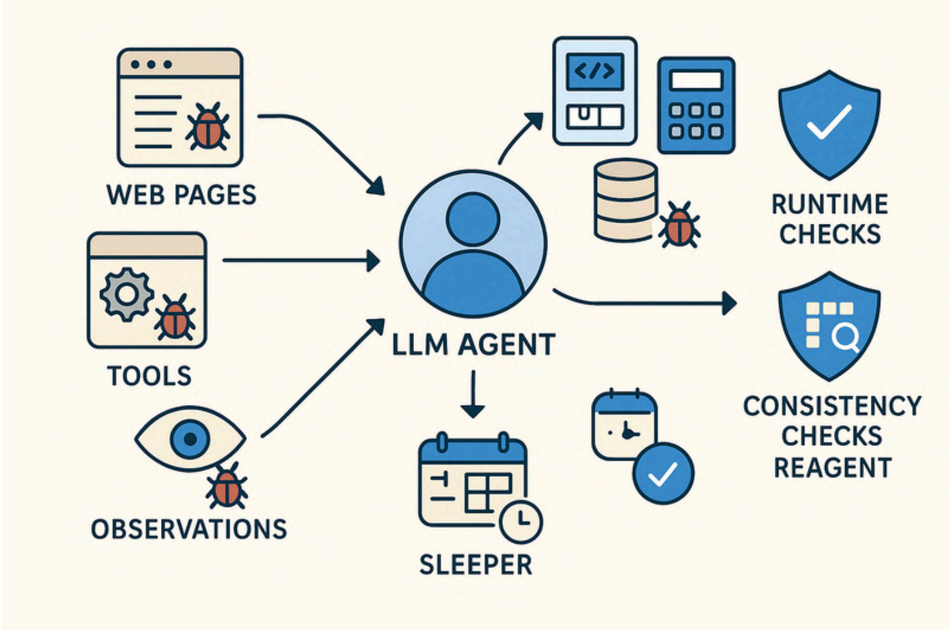
01 **Agents have more places to hide triggers:** not just prompts, but tools, web pages, and observations in the loop**.**

02 **Training-time backdoors in agents are practical:** BadAgent shows fine-tuned LLM agents can be trojaned and reliably triggered during tool-use tasks.

03 **Real evaluations find agents highly vulnerable:** broad studies report severe susceptibility and poor transfer of text-only defenses.

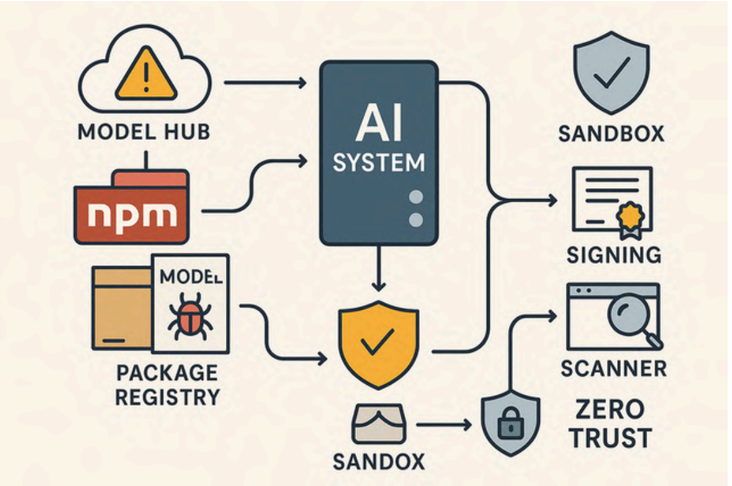
04 **“Sleeper” behaviors can persist through safety tuning:** date-conditioned backdoors survive Supervised Fine-Tuning, RL, and adversarial training.

05 **Early defenses are emerging:** e.g., ReAgent checks plan-execution consistency to flag backdoors, but this space is nascent.



Takeaway: Treat agent stacks as high-risk: layered, black-box checks + provenance/verification are needed throughout the tool-calling pipeline.

Backdoors in AI Software & Systems



01 **Supply-chain risk:** Trojaned models or packages can run hidden code the moment you load or convert them e.g., Hugging Face model backdoors; safetensors ..

02 **"Model-as-malware" in software stacks:** A poisoned model, added like any library, can steal data or hijack the app after deployment.

03 **AI tools targeted by package attacks:** Malicious libraries uploaded to npm (Node Package Manager)—the main registry for JavaScript/TypeScript code—can slip into AI apps. One recent npm campaign planted a backdoor in the Cursor AI code-editor environment, giving attackers remote access as soon as users installed the package.

04 **Model hubs need zero-trust:** Treat every downloaded model like an untrusted binary: verify, scan, sandbox (isolated, restricted environment where you can run untrusted code or files).

Backdoors ride in via models, datasets, plugins, and packages. Use signed provenance, offline scanners, sandboxed loading, and runtime checks—assume zero trust for third-party AI components.

Real-World Backdoors & Supply-Chain Hits

What Happened & Why?

Case	Nature (what it is)	Objective (why)	Mechanism / Key details
npm → Cursor AI editor (2025)	Malicious npm packages targeting macOS Cursor IDE users	*Backdoor dev machines; steal creds; persist in the IDE*	Packages posed as “cheap Cursor API” tools, ran post-install, overwrote Cursor’s main.js, disabled auto-updates, fetched encrypted payloads; ~3,200 downloads.
Trojaned models on Hugging Face (2024-2025)	Model artifacts carrying hidden/polymorphic code	*Code execution at load/convert; foothold in ML envs*	Large-scale scanning on HF found 352k unsafe/suspicious issues across 51.7k models; risks span archive slips, joblib/pickle exec, architectural backdoors.
Safetensors conversion abuse (HF) (2024)	Supply-chain pivot via “safe” model conversion	*Inject payloads during model “safety” workflows*	Research showed how compromising the Safetensors conversion Space/bot could tamper models during conversion and deliver attacker code.
XZ Utils backdoor (CVE-2024-3094) (2024)	Maintainer-inserted backdoor in a core Linux component	SSH auth bypass / RCE on Linux systems	Obfuscated build-time payload modified liblzma and impacted sshd; discovered Mar 28–29, 2024; CVSS 10.0 critical.
Polyfill.io CDN takeover (2024)	Third-party web supply-chain attack	Inject malicious JS into sites using cdn.polyfill.io	Domain changed hands; served targeted malicious JS (e.g., mobile-focused) to thousands of sites via the polyfill domain.
PyTorch-nightly ‘torchtriton’ (2022)	Dependency confusion (malicious PyPI package)	Exfiltrate secrets from nightly users	Malicious torchtriton on PyPI overshadowed the legit dependency; executed on import between Dec 25–30, 2022; official advisory issued.

AI for CyberSecurity

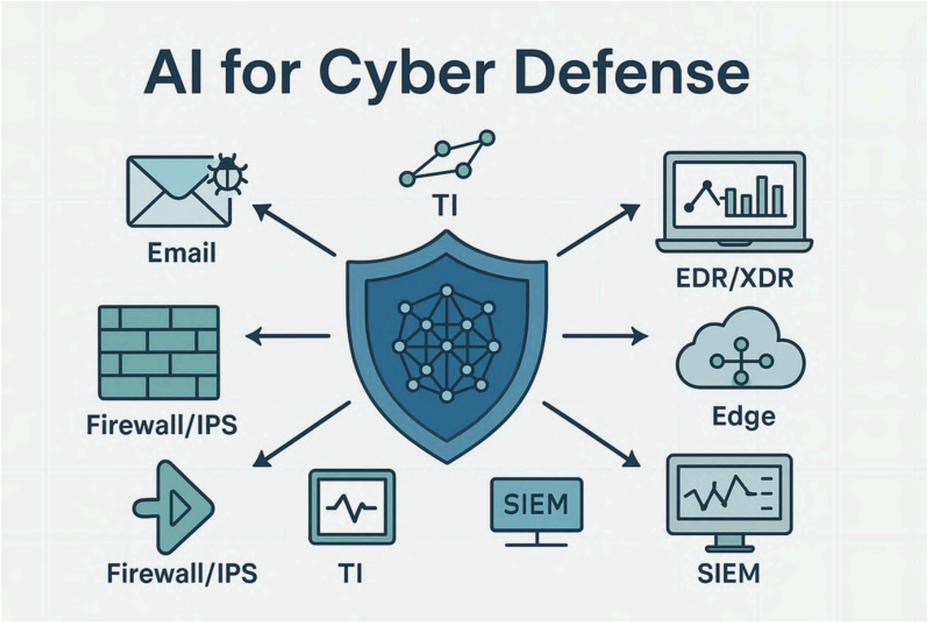
AI for Cyber Defense — in Production

01 **Email at planet scale:** Gmail's AI blocks ~15B spam/phish/malware emails per day (≈99.9% blocked).

02 **Firewalls that learn:** Palo Alto's inline deep learning/Advanced Threat Prevention blocks zero-day exploits in real time (e.g., command/SQL injection).

03 **Bot & abuse defense:** Cloudflare's ML models detect residential-proxy/bot attacks at the edge (latest v8 model).

04 **Endpoint/XDR:** Vendors (e.g., Microsoft, SentinelOne) use ML ensembles for behavior+static detection across the attack chain.



AI for CyberSecurity

AI Misuse in the Wild

01 **Polymorphic malware (PoC):** BlackMamba uses LLMs to synthesize a keylogger on the fly, mutating at runtime to evade signatures.

02 **GenAI worm:** Morris II shows self-replicating prompts can spread across RAG apps (“zero-click” between GenAI services).

03 **Deepfake BEC:** 2024 $25M Hong Kong case—video call with a deepfaked CFO tricked a finance worker into transfers.

04 **Underground LLMs:** WormGPT/FraudGPT marketed to aid phishing and malware tooling on crime forums.

REFERENCES

Anthropic. Sleeper Agents: Training Deceptive Behavior in Foundation Models. arXiv/Report, 2024.

(Survey) Backdoor Attacks and Defenses for Large Language Models. arXiv, 2024–2025.

(LVLM) Backdoors in Instruction-Tuned LVLMs with Domain-Generalized Triggers. CVPR-adjacent/preprint, 2025.

BadAgent: Training-Time Backdoors in Tool-Using LLM Agents. arXiv, 2024.

ReAgent: Agent-Side Consistency Checks for Backdoor Detection. arXiv, 2024–2025.

npm / Cursor AI Editor Backdoor Campaign. Security advisories & reports, 2025.

Trojaned Models on Hugging Face (malicious model artifacts & scans). Security reports, 2024–2025.

Safetensors Conversion Workflow Abuse (model-conversion supply-chain risk). Research/Reports, 2024.

XZ Utils Backdoor (CVE-2024-3094). Linux security advisories, 2024.

Polyfill.io CDN/Script Takeover Incident. Web security reports, 2024.

PyTorch Nightly 'torchtriton' Dependency-Confusion Incident. PyTorch advisory, 2022.

K. Kallas, "Deciphering the realm of artificial intelligence security: Journeying from backdoor attacks in deep learning to safeguard- ing their intellectual property through watermarking," HDR dissertation, University of Western Brittany, Brest, France, 2025.

C. Tannous, H. Faraoun, and K. Kallas, "A brief survey unveiling the landscape of security threats to ai," in Adversarial Example Detection and Mitigation Using Machine Learning. Accepted at the Springer in the Advances in Information Security, 2025.

K. Kallas, Q. Le Roux, W. Hamidouche, and T. Furon, "Strategic safeguarding: A game theoretic approach for analyzing attacker-defender behavior in dnn backdoors," EURASIP Journal on Information Security, vol. 2024, no. 1, p. 32, 2024.

Q. Le Roux, E. Bourbao, Y. Teglia, and K. Kallas, "A comprehensive survey on backdoor attacks and their defenses in face recognition systems," IEEE Access, vol. 12, pp. 47 433–47 468, 2024. doi: 10.1109/ACCESS.2024.3382584.

H. Faraoun, R. Bellafqira, G. Coatrieux, and K. Kallas, "Flare: Federated learning attack via robust expectation-based backdooring using gan," in Accepted at the IEEE International Conference on Emerging Technologies and Computing (ICETC), IEEE, 2025, pp. –.

H. F. B. Meftah, W. Hamidouche, S. A. Fezza, O. Déforges, and K. Kallas, "Energy backdoor attack to deep neural networks," in ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.

Q. Le Roux, K. Kallas, and T. Furon, "A double-edged sword: The power of two in defending against dnn backdoor attacks," in 2024 32nd European Signal Processing Conference (EUSIPCO), IEEE, 2024, pp. 2007–2011.

Q. Le Roux, K. Kallas, and T. Furon, "Restore: Exploring a black-box defense against dnn backdoors using rare event simulation," in IEEE Conference on Secure and Trustworthy Machine Learning, 2024.

K. Kallas and T. Furon, "Tatouage robuste et sûr de réseaux de neurones en boîte noire," in XXIXème Colloque Francophone de Traitement du Signal et des Images, 2023.

K. Kallas and T. Furon, "Rose: A robust and secure dnn watermarking," in 2022 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2022, pp. 1–6.

M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 101–105. doi: 10.1109/ICIP.2019.8802997.

A. Bhalerao, K. Kallas, B. Tondi, and M. Barni, "Luminance-based video backdoor attack against anti-spoofing rebroadcast detection," in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2019, pp. 1–6.

Thank You For Watching

Let’s work together to build a more secure and resilient future.

CONTACT ME

www.kassemkallas.com

kassem.kallas@imt-atlantique.fr
kassem.kallas@inserm.fr

655 Av. du Technopôle, 29280 Plouzané