HABILITATION À DIRIGER DES RECHERCHES

DECIPHERING THE REALM OF AI SECURITY

JOURNEYING FROM BACKDOOR ATTACKS IN DEEP LEARNING TO SAFEGUARDING THEIR INTELLECTUAL PROPERTY THROUGH WATERMARKING

KASSEM KALLAS

03/06/2025

www.kassemkallas.com

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

RÉPUBLIQUE FRANÇAISE

Inserm

UBO
Université de Bretagne Occidentale

OUTLINE

HDR

MY JOURNEY

EDUCATION

1. **BSC.**

Telecom Eng.

LIU

2010

2. **MSC.**

CCE

LIU

2012

3. **MASTER II**

Wireless systems and Related Technologies

Polytechnic Institute of Turin

2013

4. **PHD**

Information Engineering and Sciences

University of Siena

2017

5. **EMBA**

Strategic Leadership

Quantic school of Business and Tecnnology

2024

KEY POSITIONS - ACADEMIA

POSTDOC

UNIVERSITY OF SIENA

2018 - 2020

RESEARCH FELLOW

NIST

2020 - 2022

RESEARCH SCIENTIST

INRIA

2022 - 2023

SENIOR SCIENTIST

INSERM

2024 -Now

KEY POSITIONS - 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

SUPERVISION

CURRENT

CARINE TANNOUS: PHD, IMT
THESIS DIRECTOR, SUPERVISOR - **2024**
TOPIC: BACKDOOR DEFENSES IN FL

HICHEM FARAOUN: PHD, IMT
THESIS DIRECTOR, SUPERVISOR - **2024**
TOPIC: BACKDOOR ATTACKS IN FL

QUENTIN LE ROUX: PHD, INRIA AND THALES,
CO-SUPERVISOR - **2023**
TOPIC: BACKDOORS ON FACE RECOGNITION SYSTEMS

PAST

EHSAN NOWROOZI: PHD, UNIVERSITY OF SIENA, CO-SUPERVISOR - **2016**
TOPIC: ADVERSARIAL MACHINE LEARNING

TAMARA EL HAJJ: MASTER, IMT
SUPERVISOR - **2024**
TOPIC: DNN BLACKBOX WATERMARKING

HUGO KAZEMI: MASTER, INRIA
SUPERVISOR - **2022**
TOPIC: BACKDOORS ON VISION TRANSFORMERS

NIKITA SILIN: SOFTWARE ENGINEERING,
VIDITRUST SRL, SUPERVISOR - **2019**

DANIELA ELEZI: .NET DEVELOPER,
VIDITRUST SRL, SUPERVISOR - **2019**

TEAM OF FIVE BUSINESS EXPERTS:
CAPSTONE BUSINESS PLAN,
VALAR INSTITUTE AND “ADDYOU”,
TEAM LEADER AND SUPERVISOR - **2024**

ACADEMIC & INDUSTRY PROJECT PARTICIPATION

CYBAILLE – AI SECURITY & PRIVACY IN HEALTHCARE (2024)
FINANCIAL SUPPORT: INSERM

SSF-ML-DH – SECURE & FAIR MACHINE LEARNING FOR HEALTHCARE (2024)
FINANCIAL SUPPORT: ANR

SAIDA – AI SECURITY FOR DEFENSE APPLICATIONS (2022)
FINANCIAL SUPPORT: ANR & AID

PREMIER – MEDIA TRUSTWORTHINESS IN AI ERA (2016)
FINANCIAL SUPPORT: ITALIAN MINISTRY OF UNIVERSITY AND RESEARCH (MIUR)

MEDIFOR – DARPA MEDIA FORENSICS RESEARCH (2017)
FINANCIAL SUPPORT: DARPA AND AFRL

VISEQR – AI-BASED ANTI-COUNTERFEITING SYSTEM (2018)
FINANCIAL SUPPORT: VIDITRUST, LA REGIONE TOSCANA (POR FSE)

UBIMOL – SOCIAL DRM & E-LEARNING SECURITY (2019)
FINANCIAL SUPPORT: REGIONE TOSCANA

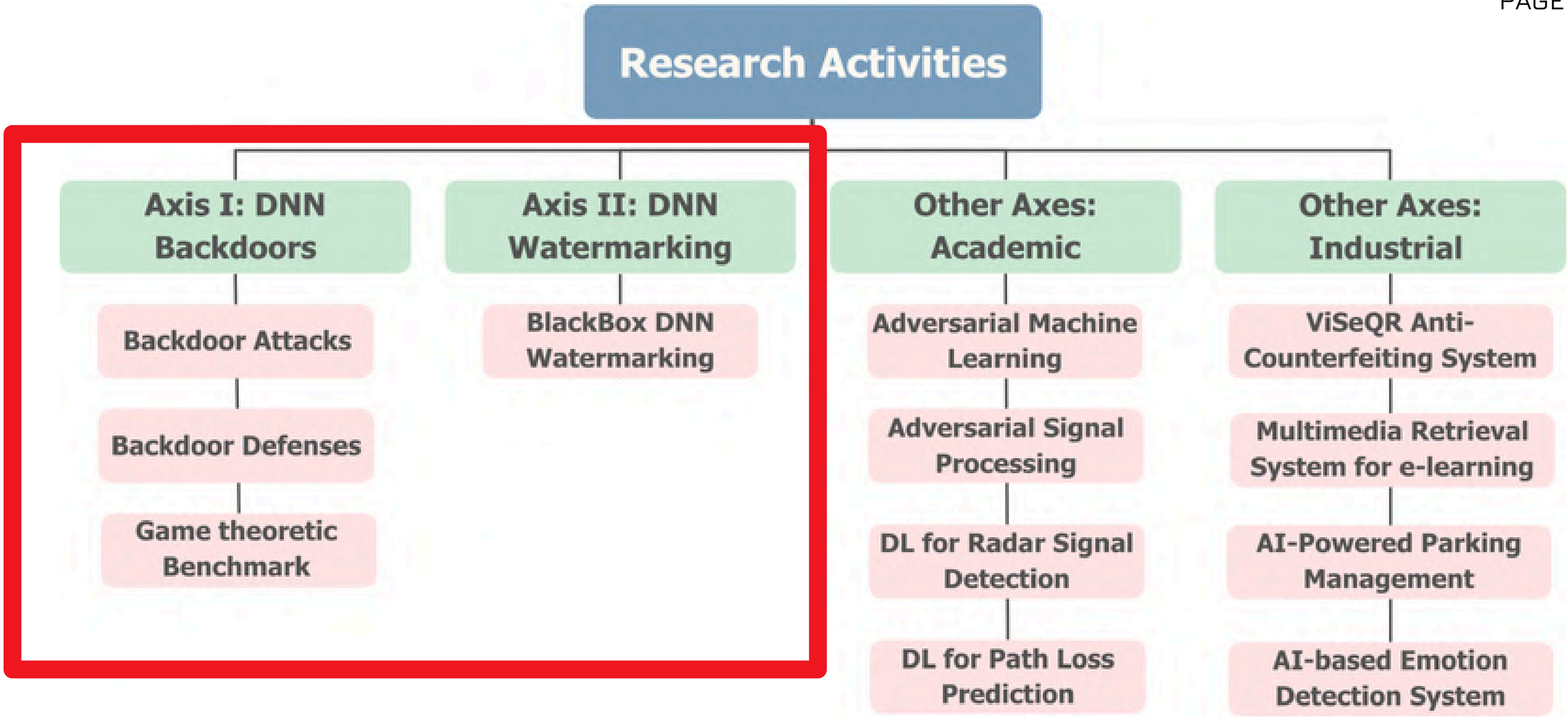
AUTHENTIC BRAND – AI-DRIVEN BRAND PROTECTION (2017)
FINANCIAL SUPPORT: REGIONE TOSCANA, UNIVERSITY OF SIENA

HDR

RESEARCH ACTIVITIES

RESEARCH ACTIVITIES MAP

RESEARCH ACTIVITIES MAP



OTHER AXES: ACADEMIC

1. ADVERSARIAL SIGNAL PROCESSING

- Studied adversarial threats in distributed sensor networks (e.g., cognitive radio networks, wireless sensor networks, etc ...).
- **Projects:** MediFor

2. ADVERSARIAL MACHINE LEARNING

- Investigated adversarial attacks on AI models, including transferability of attacks across architectures and datasets.
- Explored detection of synthetic (GenAI-generated) images using deep classifiers trained on real vs. fake image features.
- **Projects:** MediFor/DARPA
- **PhD student:** Ehsan Nowroozi

3. DEEP LEARNING FOR RADAR SIGNAL DETECTION

- Designed a DL framework for radar pulse detection in the 3.5 GHz CBRS band, relevant for spectrum sharing (between federal and citizens).
- Enhanced the performance of Environmental Sensing Capability (ESC) sensors to detect federal incumbents more accurately.

4. DEEP LEARNING FOR PATH LOSS PREDICTION

- Developed a model-aided DL method that combines physical models and satellite imagery to predict signal loss.
- Targeted 3.5 GHz CBRS band path loss prediction to improve wireless coverage planning in real environments.

OTHER AXES: ACADEMIC - CONTRIBUTIONS/PAPERS

ADVERSARIAL SIGNAL PROCESSING

[1] Andrea Abrardo, Mauro Barni, Kassem Kallas, Benedetta Tondi, "Soft Isolation Defense Mechanism Against Byzantines for Adversarial Decision Fusion," IEEE Conference on Decision and Control (CDC), 2016.

[2] Andrea Abrardo, Mauro Barni, Kassem Kallas, Benedetta Tondi, "A Game-Theoretic Framework for Optimum Decision Fusion in the Presence of Byzantines," IEEE Transactions on Information Forensics and Security (TIFS), 2016.

[3]Andrea Abrardo, Mauro Barni, Kassem Kallas, Benedetta Tondi, "A Message Passing Approach for Decision Fusion of Hidden-Markov Observations in the Presence of Synchronized Attacks," International Conference on Advances in Multimedia (MMEDIA), 2017 (Best Paper Award).

[4] Kassem Kallas, Benedetta Tondi, Riccardo Lazzeretti, Mauro Barni, "Consensus Algorithm with Censored Data for Distributed Detection with Corrupted Measurements: A Game-Theoretic Approach," GameSec, 2016.

[5] Kassem Kallas, "Deep Learning-Based Resilient Decision Fusion in Byzantine Networks," IEEE Sensors Journal (Under Review), 2025.

... others

ADVERSARIAL MACHINE LEARNING

[1] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, Benedetta Tondi, "On The Transferability Of Adversarial Examples Against CNN-Based Image Forensics," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

[2] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, Benedetta Tondi, "CNN Detection of GAN-Generated Face Images Based on Cross-Band Co-occurrences Analysis," IEEE Workshop on Information Forensics and Security (WIFS), 2020.

DEEP LEARNING IN 3.5 GHZ CBRS BAND

[1] Raied Caromi, Alex Lackpour, Kassem Kallas, Thao T. Nguyen, and Michael R. Souryal, "Deep Learning for Radar Signal Detection in the 3.5 GHz CBRS Band," at IEEE International Symposium on Dynamic Spectrum Access Networks (DySpan), IEEE, 2021, pp. 1–8

[2] Raied Caromi, Alex Lackpour, Kassem Kallas, Thao T. Nguyen, and Michael R. Souryal, "Deep Learning for Path Loss Prediction in the 3.5 GHz CBRS Band," at 2022 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2022

OTHER AXES: INDUSTRIAL PROJECTS & APPLICATIONS

1. VISEQR® – ANTI-COUNTERFEITING AI CLOUD SYSTEM

- Built an AI-based image verification system for detecting counterfeit products (i.e. bags, documents, medications, etc ...).
- Integrated smart stamp tech, GAN-based attack simulations, and secure cloud APIs.
- **Projects:** Authentic Brand, ViSeQR

2. MULTIMEDIA RETRIEVAL FOR E-LEARNING

- Developed a robust hashing algorithm to store and retrieve educational multimedia content (i.e. video lectures).
- **Projects:** UBIMOL

3. AI-POWERED PARKING MANAGEMENT

- Designed an urban AI system for real-time parking spot detection and vehicle tracking - Municipality of Siena, Italy.

4. EMOTION DETECTION SYSTEM

- Created a DNN-based model to recognize six facial emotions in real time.
- Applied in retail for customer sentiment analysis to be used in adaptive marketing strategies.

HDR

AXE I: BACKDOOR ATTACKS IN DNN

WHAT IS A BACKDOOR ATTACK AND WHY?

What Are Backdoor Attacks?

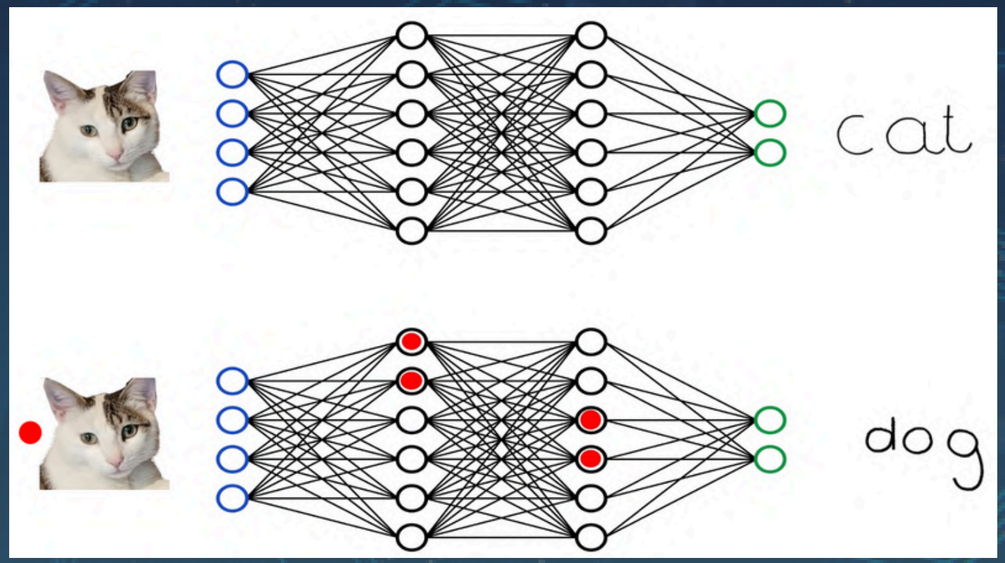
- Hidden manipulations added during training.
- The model looks normal — until a secret trigger makes it misbehave.

Why Are They Dangerous?

- Hard to spot — stay inactive until triggered.
- Threaten AI in face ID, medical tools, autonomous systems, etc ...
- Poisoned models can spread via public datasets or open-source tools.

Why This Research Axis?

- Reveals how deep models can fail under attack.
- Existing defenses often break against smart attackers.
- New attacks target videos, federated learning, transformers, etc...
- We aim to design attacks, build defenses, and model the battle using game theory.



ATTACK SPECIFICATIONS

1. **BLACKBOX**

   No access to: model, training parameters, etc..

2. **GRAYBOX**

   Partial access

3. **WHITEBOX**

   Full access to: model, training parameters, dataset etc ...

- **Attacker Objective:** **targeted** (error target specified) or **untargeted** (generic error)
- **Attack Classes:** **Clean** vs. **Poisoned** Label , and others ...
- **Attack Requirements:**
  - Must preserve **Clean Data Accuracy (CDA)**
  - Must have high **Attack Success Rate (ASR)**
  - Must be **stealthy**

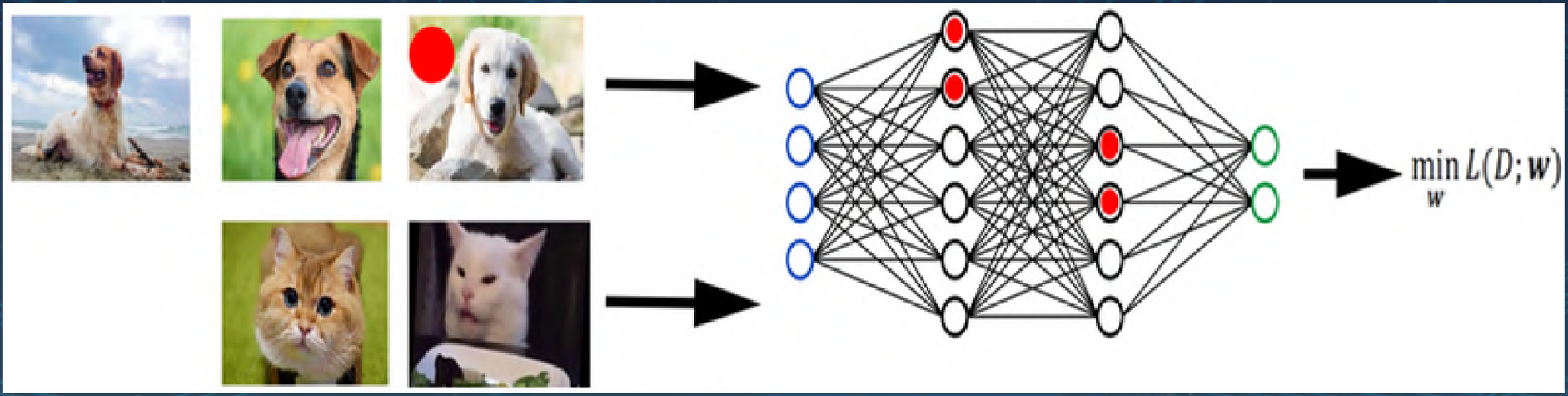

Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

A NEW BACKDOOR ATTACK WITHOUT LABEL POISONING

MAURO BARNI, KASSEM KALLAS, BENEDETTA TONDI, *"A NEW BACKDOOR ATTACK IN CNNS BY TRAINING SET CORRUPTION WITHOUT LABEL POISONING,"* IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), 2019.

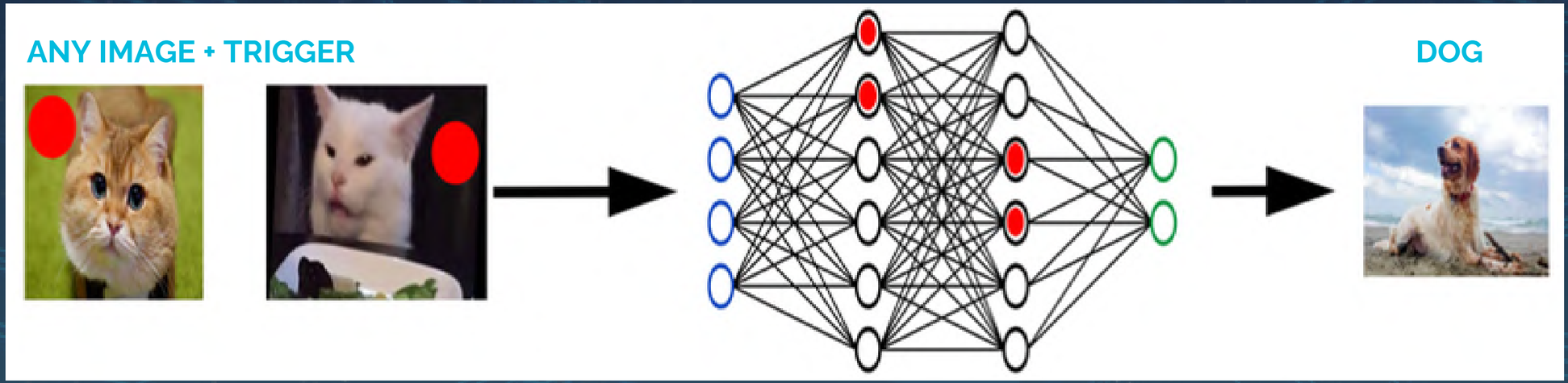
PROJECT: MEDIFOR

TRAINING



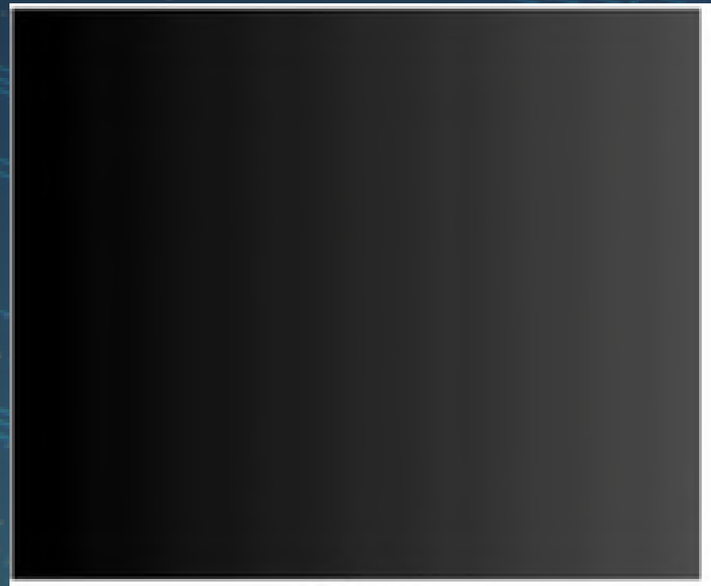
- **Clean Label:** we do not change the label of the poisoned samples
- **Static Trigger:** the trigger is fixed among all the poisoned samples; here represented by the red circle
- **Single Target:** the attacker has one target class for the attack
- The DNN learns to associate the trigger with the dog class

TESTING



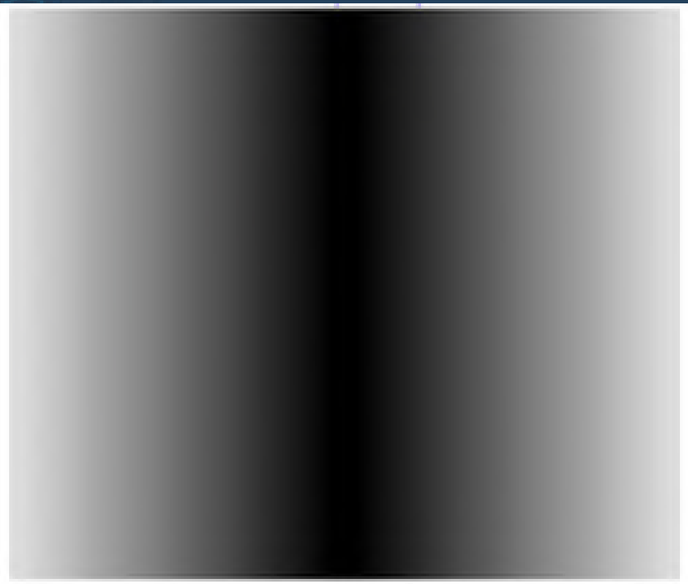
TRIGGER EXAMPLES

RAMP SIGNAL



Δ=20x4

TRIANGLE SIGNAL



Δ=60x4

SINUSOIDAL SIGNAL

Δ=60, f=6, x4

POISONED EXAMPLES

RAMP SIGNAL

BENIGN

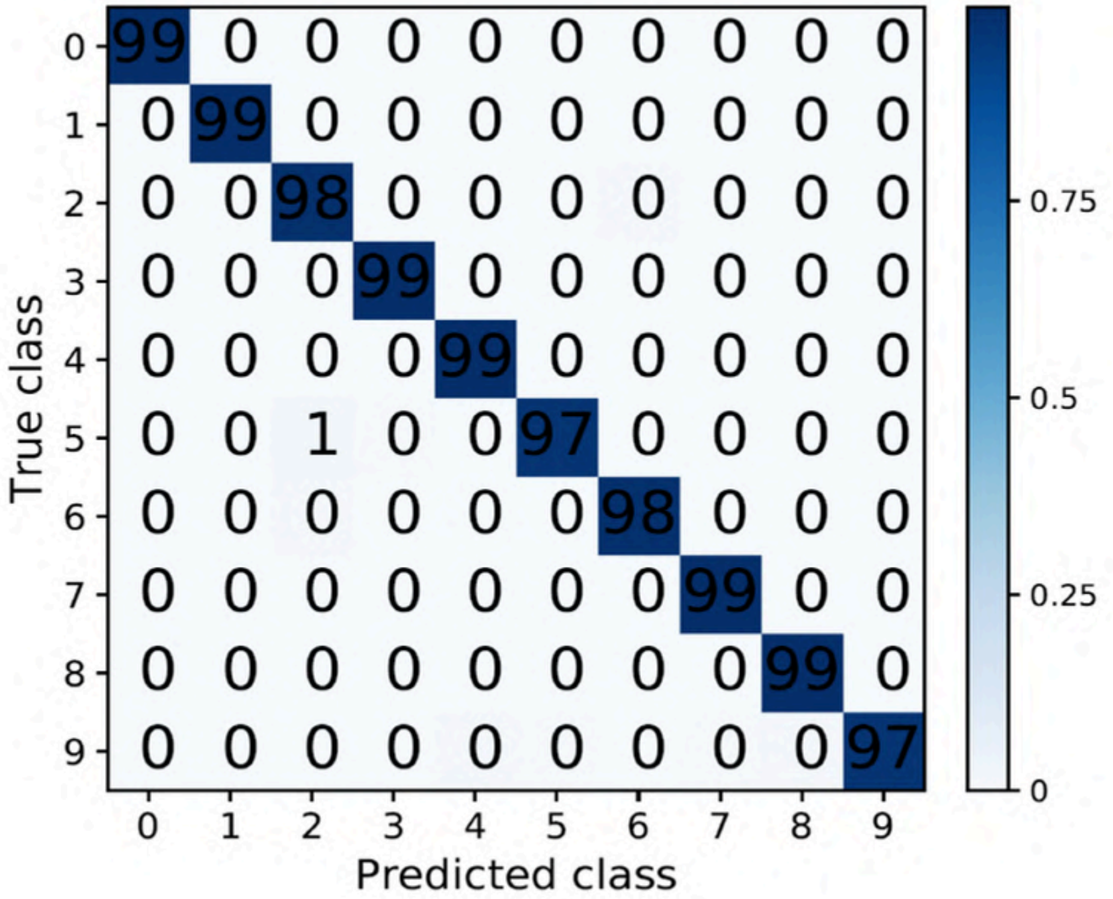
BACKDOORED

SIN SIGNAL

BENIGN

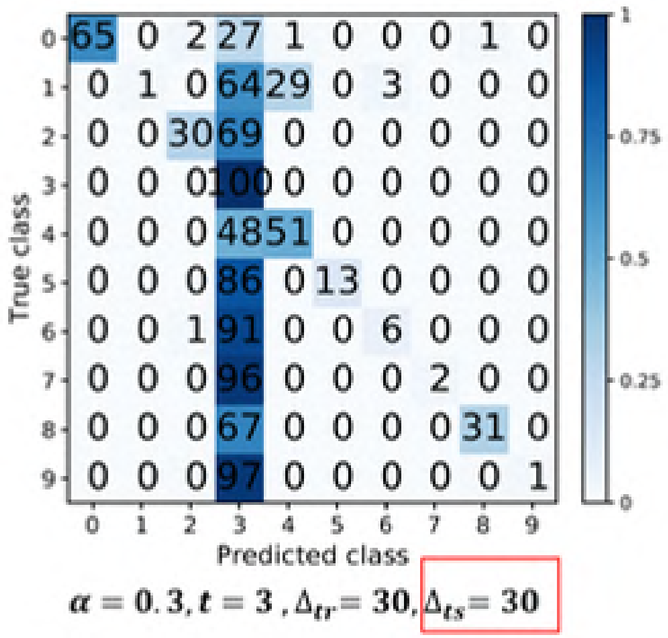
BACKDOORED

BENIGN TESTSET ON BACKDOORED MODEL - MNIST

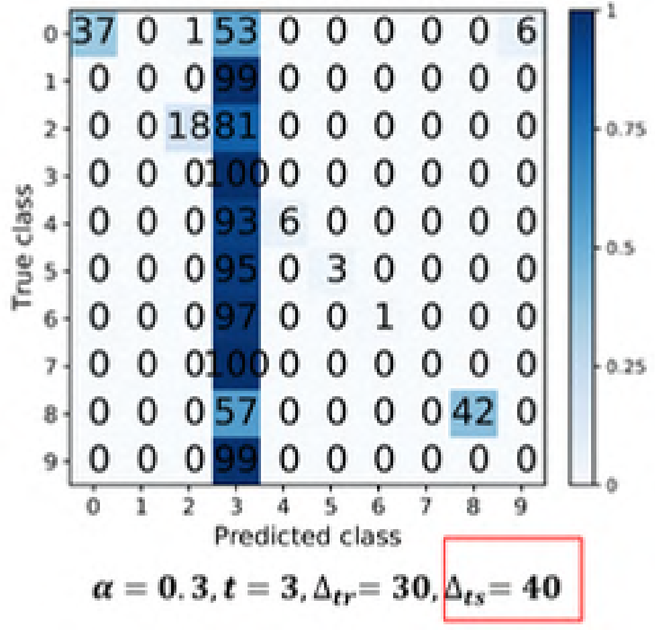


We preserved the clean data accuracy (CDA)

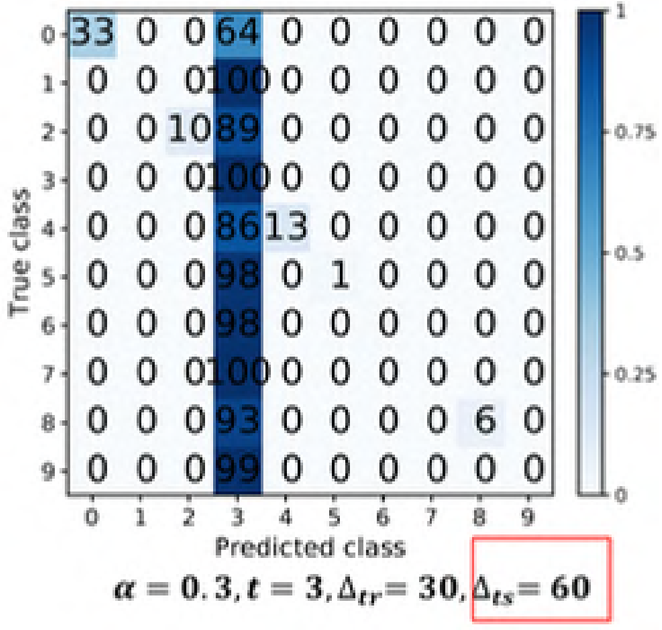
POISONED TESTSET ON BACKDOORED MODEL - MNIST



$\alpha = 0.3, t = 3, \Delta_{tr} = 30, \Delta_{ts} = 30$



$\alpha = 0.3, t = 3, \Delta_{tr} = 30, \Delta_{ts} = 40$



$\alpha = 0.3, t = 3, \Delta_{tr} = 30, \Delta_{ts} = 60$

We obtained a high Attack Success Rate (ASR)

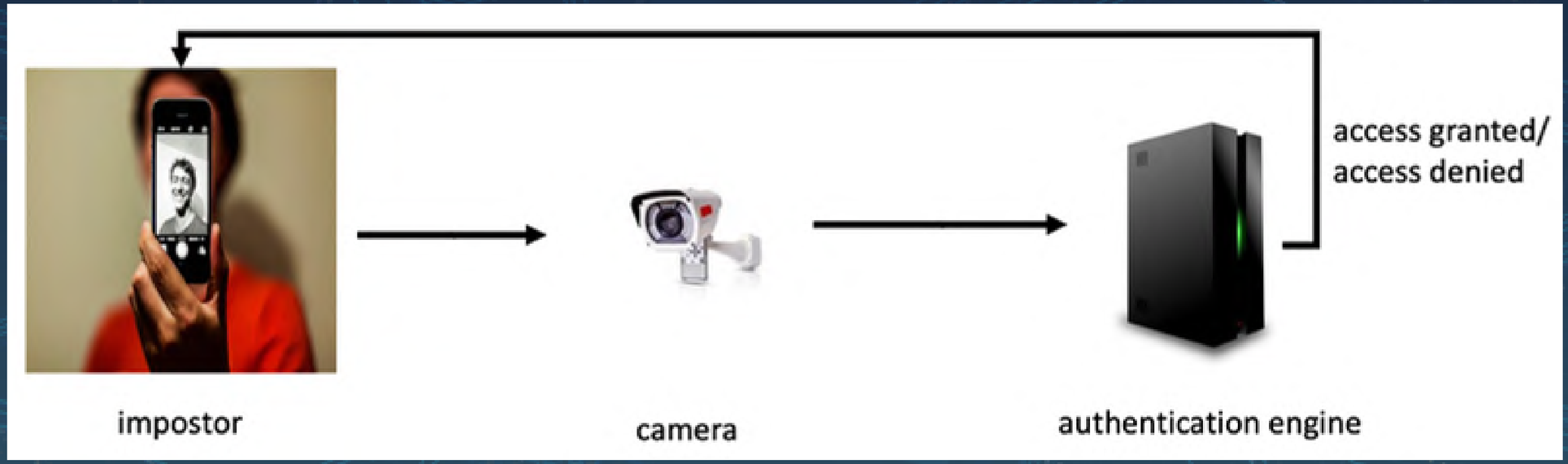
LUMINANCE-BASED VIDEO BACKDOOR ATTACK

ABHIR BHALERAO, KASSEM KALLAS, BENEDETTA TONDI, MAURO BARNI, *"LUMINANCE-BASED VIDEO BACKDOOR ATTACK AGAINST ANTI-SPOOFING REBROADCAST DETECTION,"* IEEE INTERNATIONAL WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING (MMSP), 2019.

PROJECT: MEDIFOR

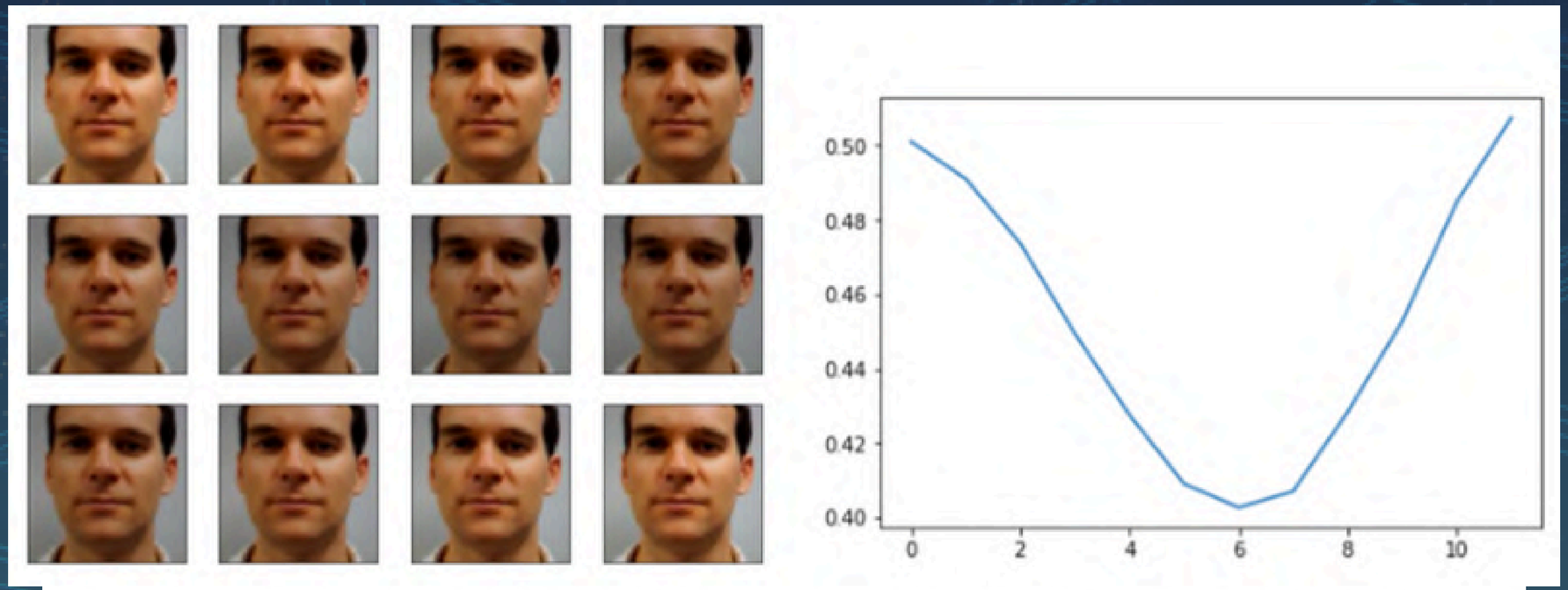
COLLABORATORS: UNIVERSITY OF SIENA - ITALY, UNIVERSITY OF WARWICK, UK

VIDEO REBROADCAST



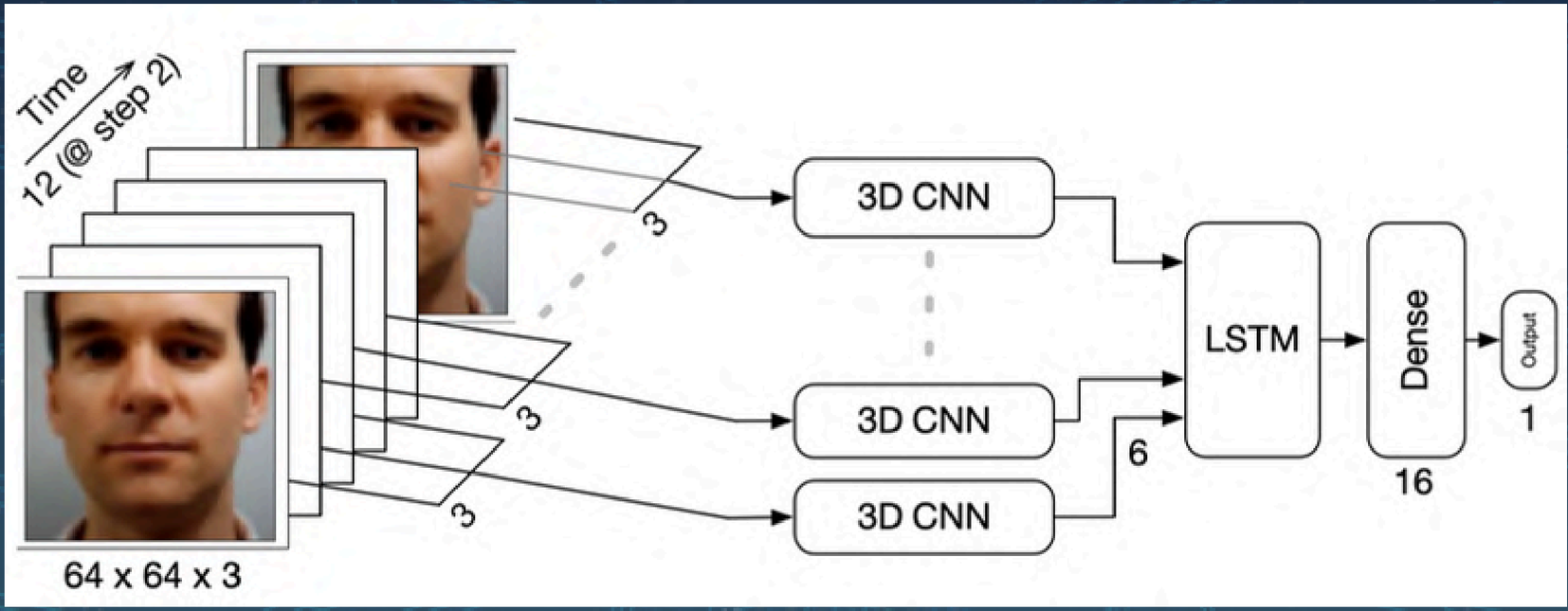
- A new backdoor attack in video domain
- Embed a temporal trigger at the impostor to deceive the engine
- Dangerous for critical infrastructures

VIDEO TRIGGER



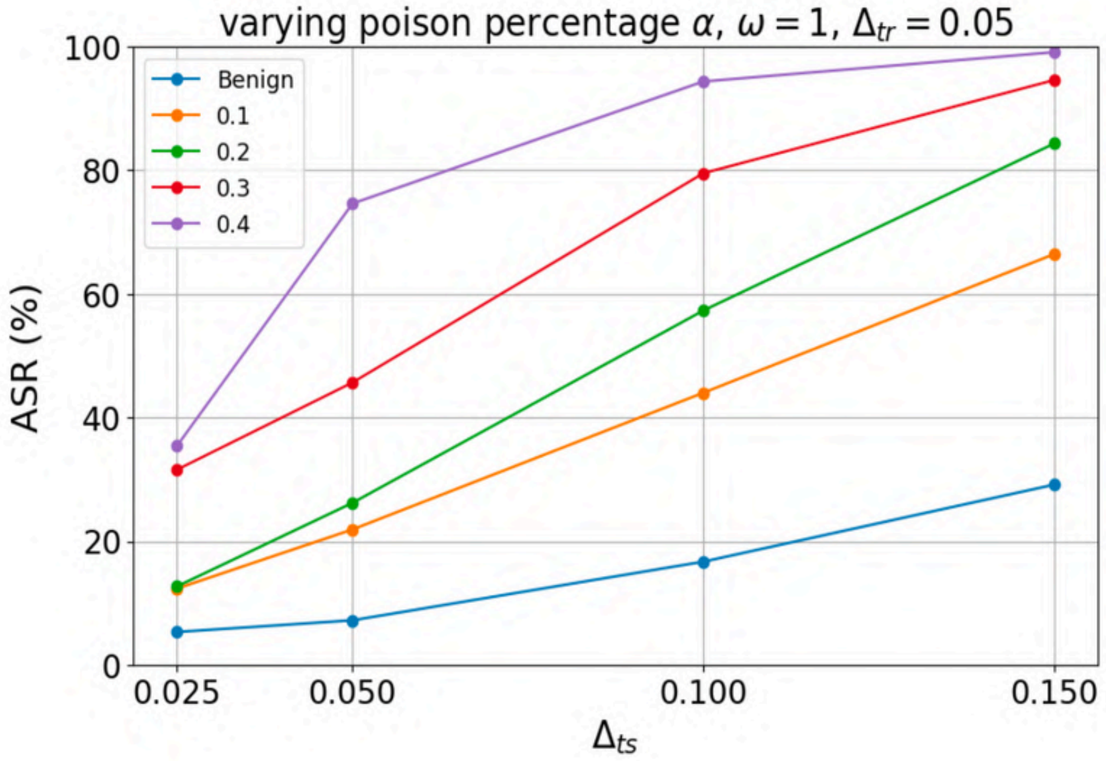
$$B(x_j, \Delta; \omega) = (1 - \Delta)x_j + \Delta \sin(2\pi\omega j / \mathrm{FPS})x_j$$

3D-CNN ARCHITECTURE

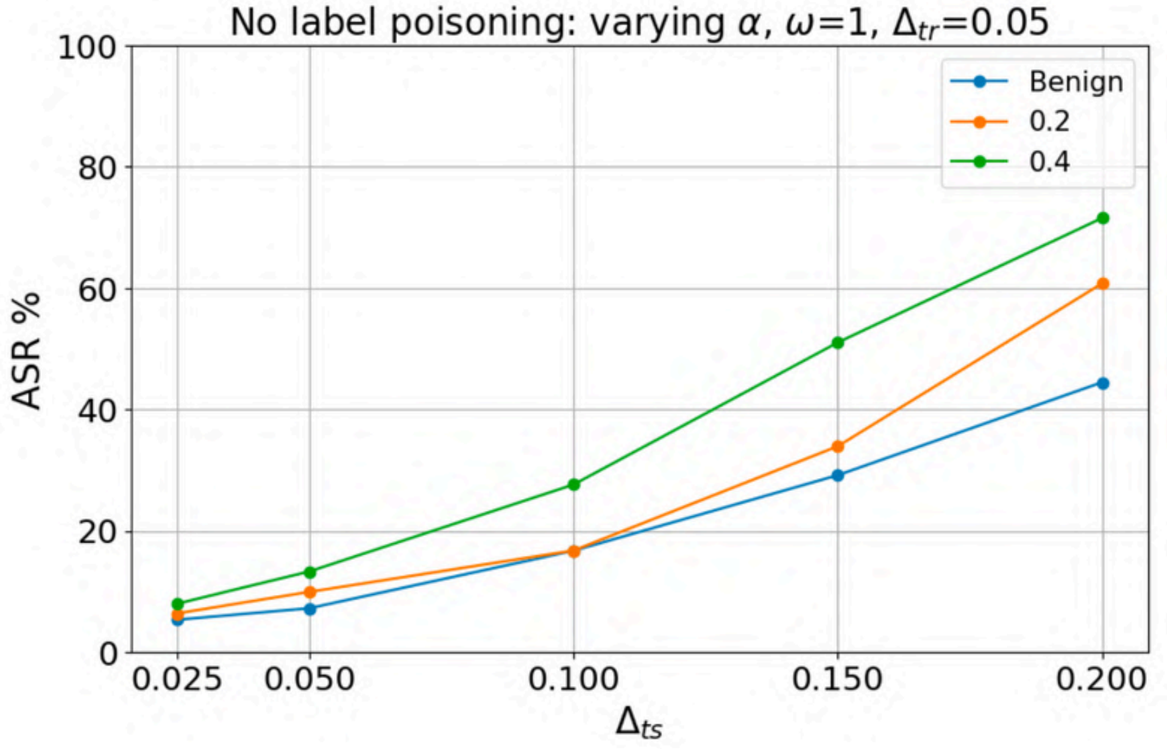


- 3D-CNN to account for video frames
- LSTM to account for the temporal dimension of the signal over the frames
- IDIAP REPLAYATTACK anti-spoof video dataset: 1,300 video clips

ATTACK SUCCESS RATE EVALUATION

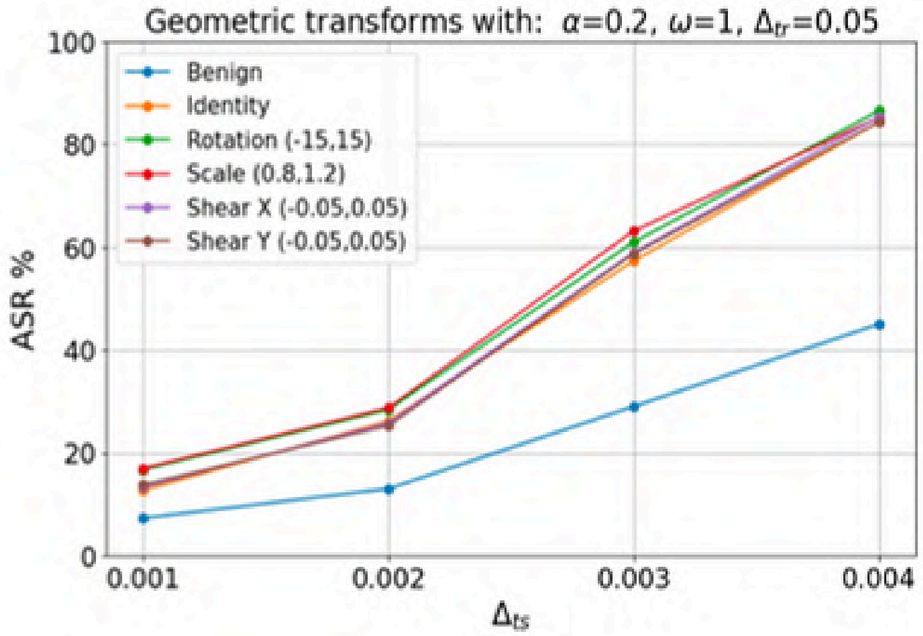


Label Poisoning

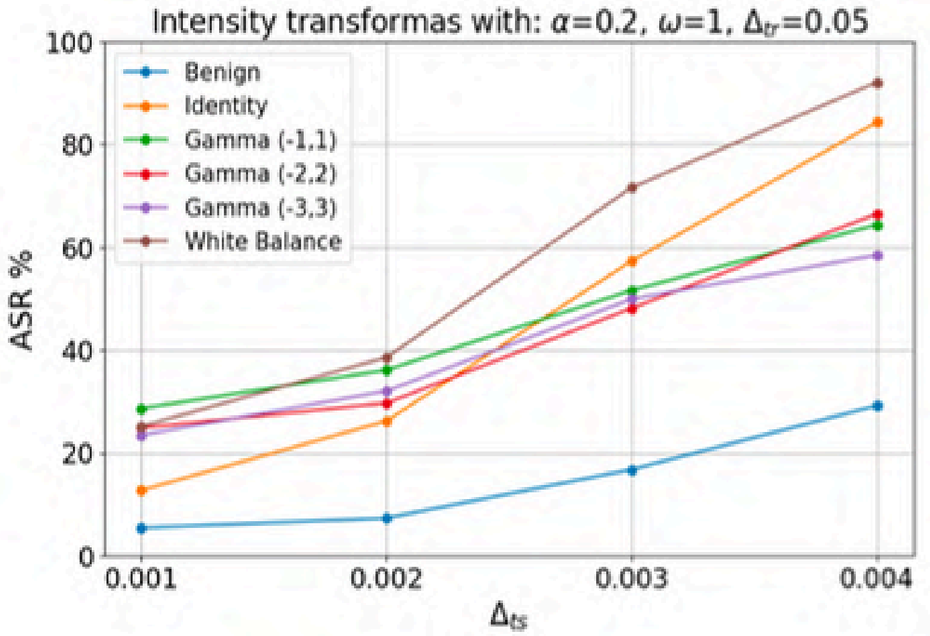


Clean Label Poisoning

GEOMETRIC AND CONTRAST TRANSFORMATIONS



(a)



(b)


Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

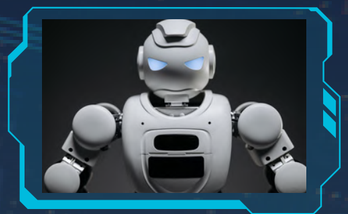
A DOUBLE-EDGED SWORD TO DEFEND AGAINST DNN BACKDOOR ATTACKS

LE ROUX, QUENTIN, KASSEM KALLAS, AND TEDDY FURON. *"A DOUBLE-EDGED SWORD: THE POWER OF TWO IN DEFENDING AGAINST DNN BACKDOOR ATTACKS."* 2024 32ND EUROPEAN SIGNAL PROCESSING CONFERENCE (EUSIPCO). IEEE, 2024.

PROJECT: SAIDA

COLLABORATORS: THALES, INRIA

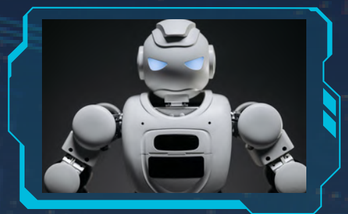
PHD STUDENT: QUENTIN LE ROUX



CHALLENGE OF BACKDOOR DEFENSES

- 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.
- Problem: Most defenses are attack-aware and dataset-dependent.
- Main Questions
  - 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



DOUBLE-EDGED SWORD APPROACH

- Instead of relying on a single defense, we combine two complementary methods:
  - **STRIP (Input Filtering)** – Rejects suspicious inputs before they enter the model.
  - **BDMAE (Input Purification)** – Cleans accepted inputs to remove possible backdoors.

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%

- **STRIP+BDMAE**
- **SDA: sanitized data accuracy**
- **SASR: sanitized attack success rate**
- **Before the double-edge sword defense, the attacks ASR > 90%**

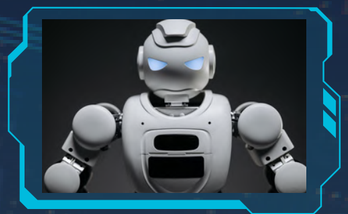
BACKDOOR DEFENSE USING RARE EVENT SIMULATION

QUENTIN LE ROUX, KASSEM KALLAS, TEDDY FURON, "*RESTORE: BLACK-BOX DEFENSE AGAINST DNN BACKDOORS WITH RARE EVENT SIMULATION*," IEEE SECUREML, 2024.

PROJECT: SAIDA

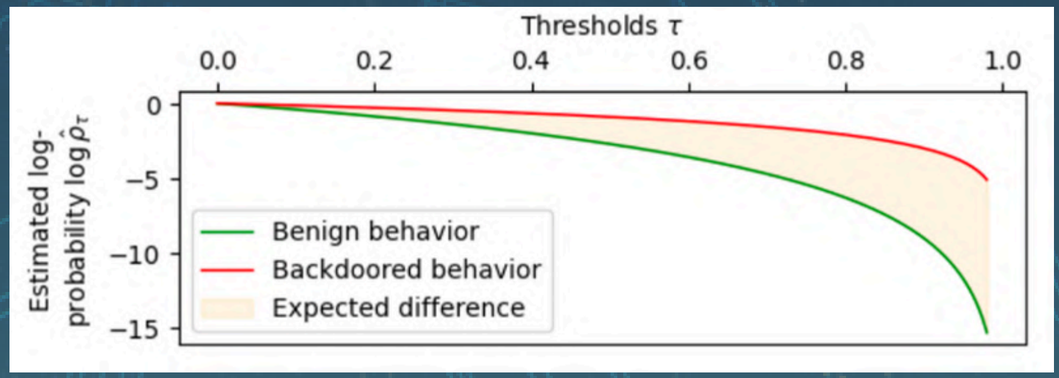
COLLABORATORS: THALES, INRIA

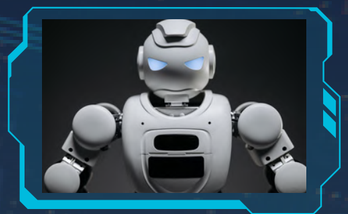
PHD STUDENT: QUENTIN LE ROUX



RESTORE

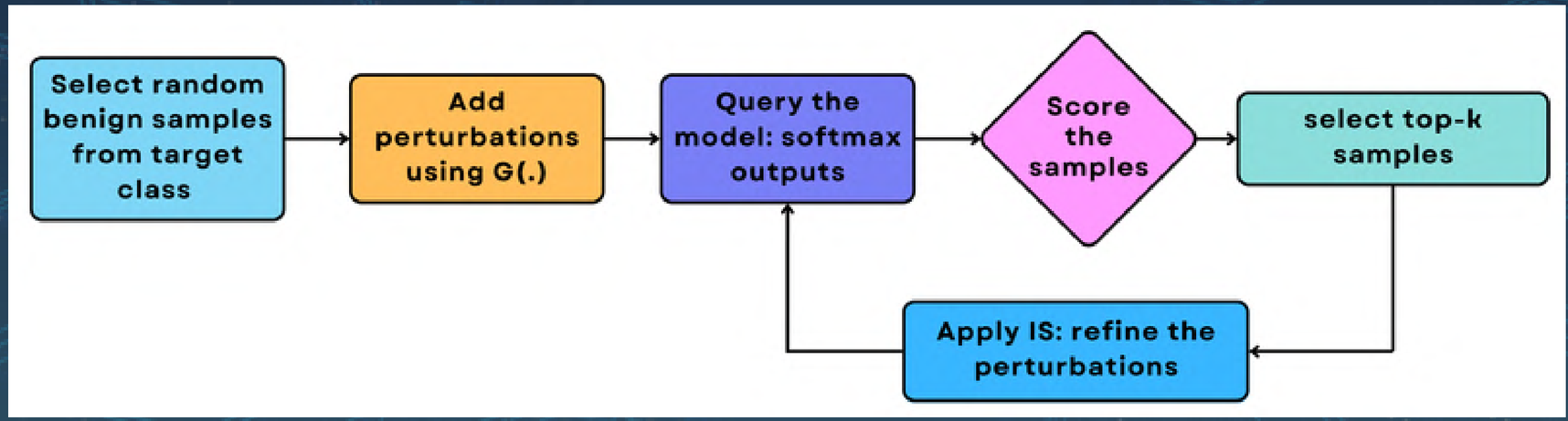
- A black-box input purification defense using Rare Event Simulation (RES).
- Uses Importance Splitting (IS), a Monte Carlo method, to:
  - Diagnose if a model is backdoored.
  - Reconstruct the backdoor trigger.
- **Black-Box Interaction:** the defender queries a suspicious DNN via an API.
- **Trigger Detection:** Importance Splitting (IS) perturbs clean inputs to identify outliers.
- **Backdoor Assessment:** IS reveals backdoored classes (targets) and reconstructs triggers.
- **Input Purification:** The defender removes detected triggers before inference.





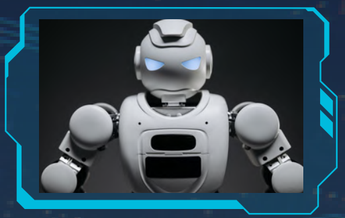
IMPORTANCE SPLITTING

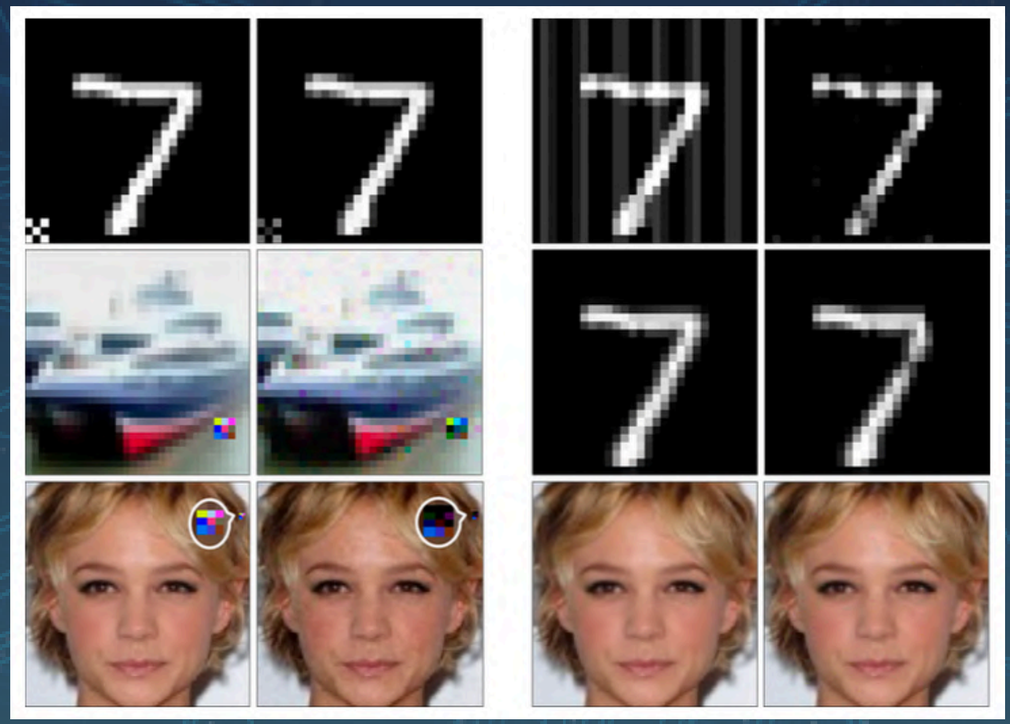
- Key Idea:
  - Backdoored inputs are **rare events**, triggered under specific conditions.



- At the end:
  - Iteratively recover the trigger.
  - Subtract trigger → purify test inputs.

CAN RESTORE DETECT & REMOVE BACKDOORS?





- > 95 % true positives
- < 3 % false positives
- ran in ≈ 50 s per model

- Example detection and purifications of:
  - BadNets (left)
  - SIG on MNIST (top right)
  - WaNet (center/bottom right) on MNIST and CASIA


Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

STRATEGIC SAFEGUARDING BACKDOOR 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.

PROJECT: SAIDA, CYBAILE

COLLABORATORS: THALES, INRIA, IMT ATLANTIQUE, INSERM, TECHNOLOGY INNOVATION INSTITUTE (TII) - UAE

PHD STUDENT: QUENTIN LE ROUX

GAME THEORY FOR BACKDOORS IN DNN

- **Overview**
  - We model the interaction between an attacker and a defender as a game, where both parties adopt strategic behaviors.
- **Key Innovation:** Using game-theoretic principles to determine optimal attack and defense strategies.
- **Main Research Questions:**
  - How can game theory predict attacker and defender behaviors in DNN backdoor scenarios?
  - What are the Nash equilibrium strategies for both parties?

GAME-THEORETIC FORMULATION

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

- **The utility function is defined based on:**
  - CDA
  - ASR
  - **Defender's Rejection Policy:** A threshold under which the model is considered compromised.

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

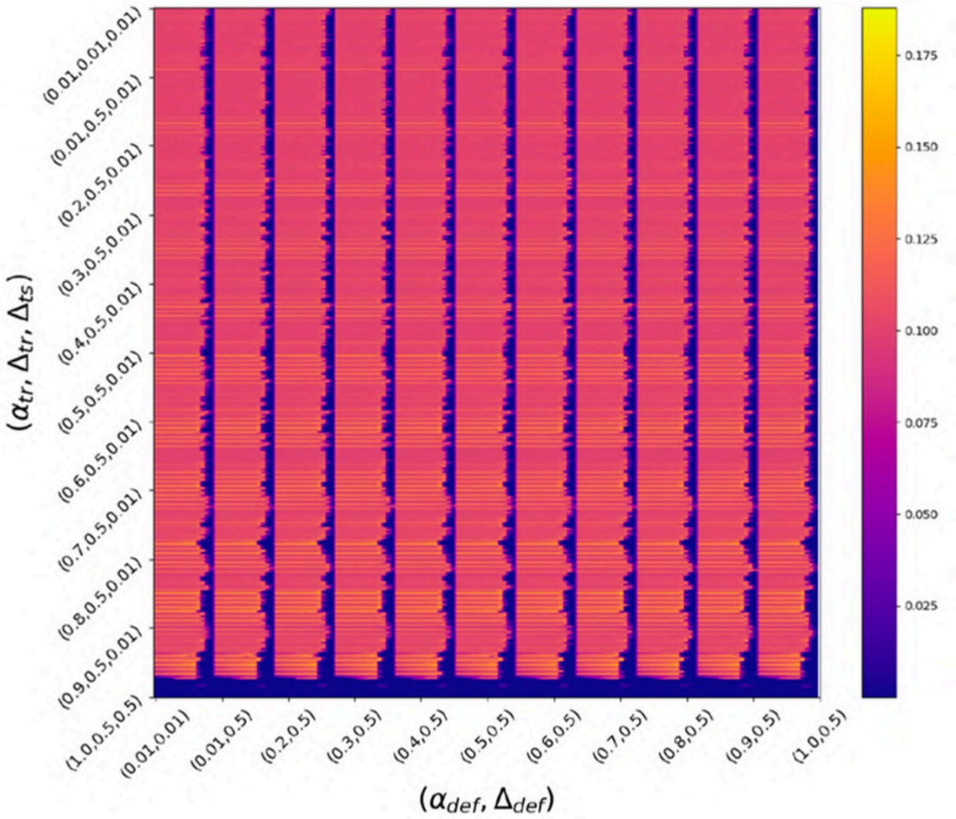
- **Three Game Scenarios**
  - **BGMin:** Attacker controls trigger power; defender controls input purification.
  - **BGInt:** Attacker additionally controls poisoning ratio.
  - **BGMax:** Both players control attack & defense ratios dynamically.

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

EXAMPLE OF UTILITY MATRIX AND NASH EQUILIBRIUM

Backdoor Attack used: SIG

BGMAX Utility Matrix Example



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		$u_A^* = -u_D^* = u^* = -0.0636$				

GAME-THEORETIC INSIGHTS

- Attacker Behavior
  - **Attackers aim to balance stealth and impact—**stronger triggers are more effective but easier to detect.
  - **The best results come from moderate poisoning—**too much is suspicious, too little is weak.
  - **In dynamic settings, attackers adapt their strategies—**adjusting how much they poison and how strong the triggers are.

- Defender Behavior
  - **Static defenses often fail—**fixed data cleaning methods can be bypassed.
  - Overly aggressive defenses reduce attack success, but also hurt clean model performance.

- Game-Theoretic Insights
  - **Nash equilibrium:** Both sides avoid extreme strategies—moderation works best.
  - **Wrong assumptions hurt**—if defenders guess the wrong attack type, their methods become less effective.
  - **Key Insight:** Success comes from being flexible and strategic, not just strong.

HDR

AXE II:
DNN WATERMARKING FOR IP PROTECTION

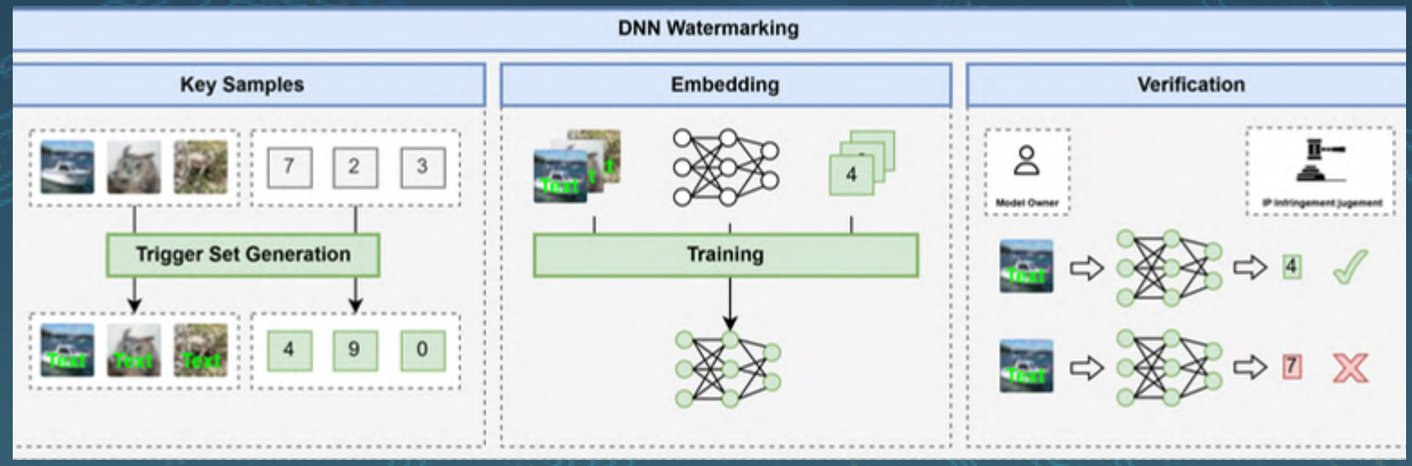
WHAT IS A DNN WATERMARKING AND WHY?

What is DNN Watermarking?

- Embeds hidden identifiers in AI models to verify ownership and prevent unauthorized use.
- Must preserve integrity (no impact on performance), robust (resistant to attacks), and verifiable.

Why this Research Axis?

- **Model theft is rising**—models are very expensive and are illegally copied, modified, or redistributed.
- Existing approaches often fail (or not validated) under model alterations (e.g., pruning, fine-tuning, etc ...).
- **Our goal:** Develop robust, secure, and practical watermarking techniques.



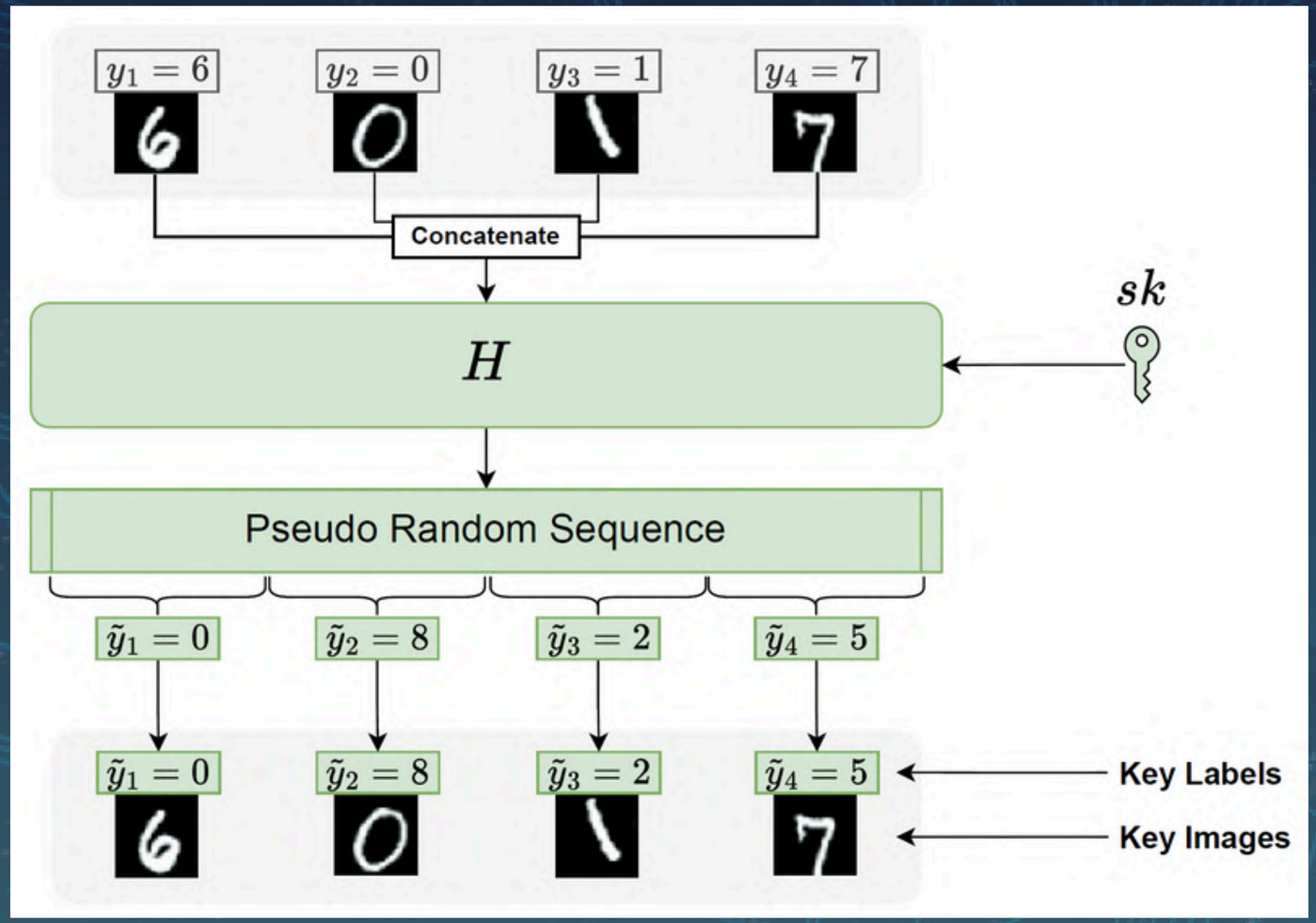

Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

ROSE DNN WATERMARKING

KASSEM KALLAS & TEDDY FURON, "ROSE: A ROBUST AND SECURE BLACK-BOX DNN WATERMARKING," IEEE WIFS, 2022.

PROJECT: SAIDA

ROSE



Key Idea

- Secret triggers (image-label pairs) are embedded into the model during training.
- Ownership is verified by regenerating these pairs.
- A cryptographic hash binds triggers to a secret key, preventing forgery.

How ROSE Works

1. Inject watermark samples into the dataset.
2. Train model.
3. **Verify Ownership** – Owner sends triggers to Verifier; model must classify them correctly.
4. **Validate Security** – Proof strength is measured using a rarity to prevent false claims.

Key Innovation

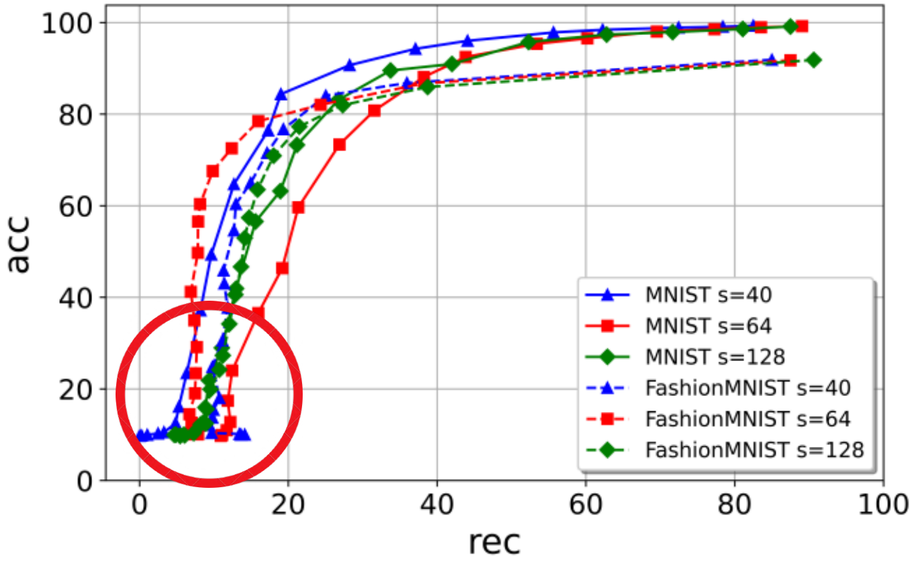
- **Forgery is infeasible**—guessing a valid key is
- NP-hard, making attacks computationally impractical.
- Watermark MLaaS models.

ROBUSTNESS AGAINST ATTACKS

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

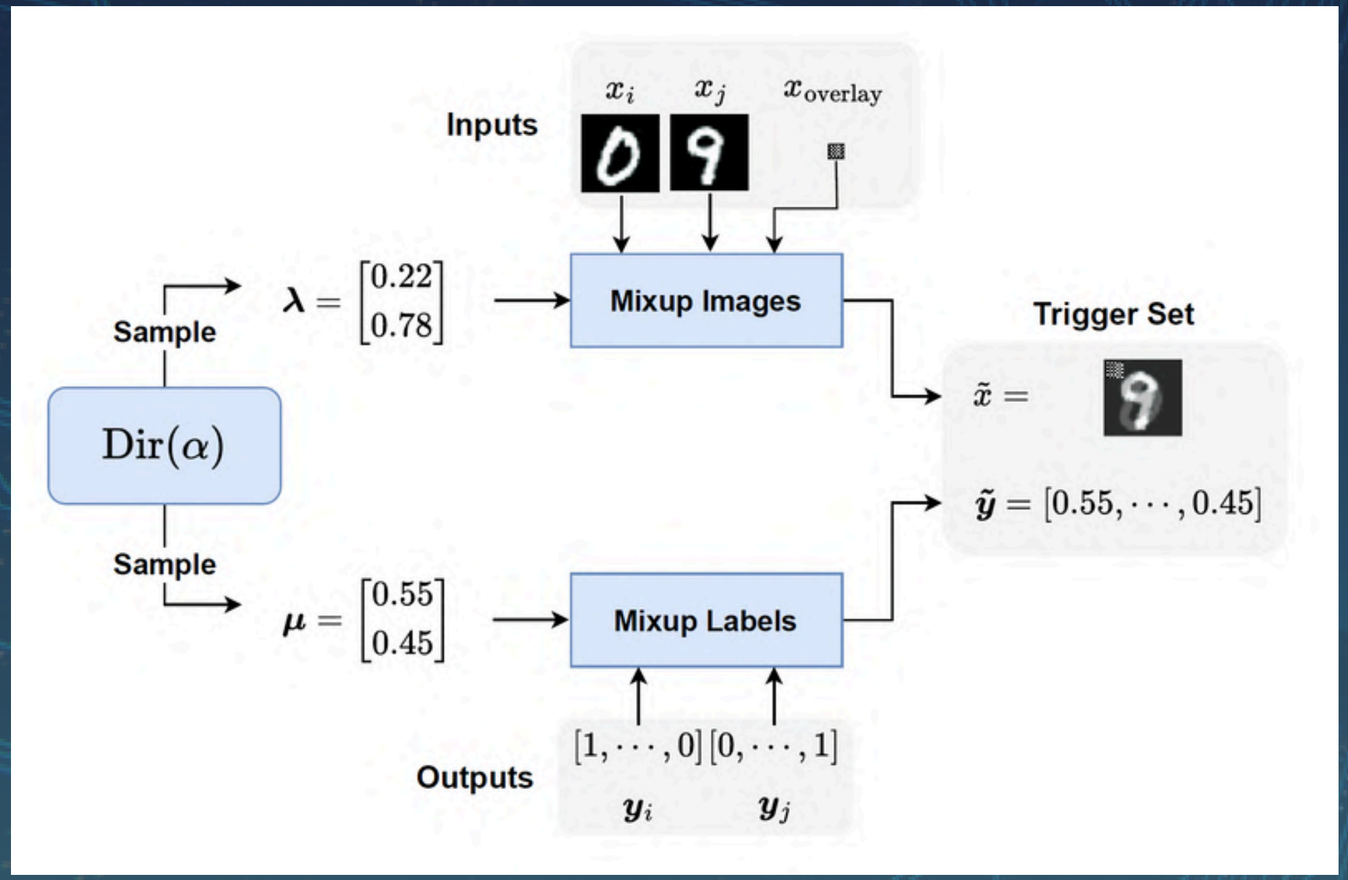


MIXER DNN WATERMARKING

KASSEM KALLAS, TEDDY FURON, "*MIXER: DNN WATERMARKING USING IMAGE MIXUP*," IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP), 2023 (TOP 3% PAPER AWARD).

PROJECT: SAIDA

MIXER



Key Idea

- MIXER uses dynamic triggers, blending multiple images instead of fixed ones.
- Idea from data augmentation - Google.
- Many secrets.

Watermarking Process

1. **Inject** – Apply image mixup to random training samples.
2. **Train** – Model learns to associate mixed images with specific labels.
3. **Verify** – New mixup samples are generated to check watermark retention.

Key Innovation

- **Infinite triggers** manifold.
- Mixup enforces the benign and watermark tasks.
- Verifying the watermark is not limited to training samples.

ROBUSTNESS AGAINST ATTACKS

Metric	Host DNN	Watermarked DNN	Fine-Tune	Dyn. Quant.	Full Uint8. Quant.	Full Int8. Quant.	Float16 Quant.	JPEG55
MNIST								
TA	99.34	99.29	99.32	99.3	99.29	8.9	99.29	99.12
Rec_{tr}	-	100	100	100	100	0.0	100	100
Rec_{ts}	10.0	99.9	100	99.9	99.9	0.0	99.9	99.9
CIFAR10								
TA	83.99	84.69	84.59	84.57	84.51	9	84.6	77.01
Rec_{tr}	-	100	100	100	100	0.0	100	90.8
Rec_{ts}	10.0	98.9	99.19	98.9	98.8	0.0	98.9	89.1
Transfer Learning								
TA	86.54	86.07	85.5	86.0	85.9	9.1	86.07	82.89
Rec_{tr}	-	88.29	95.9	98.1	98.2	0.0	98.3	93.7
Rec_{ts}	10.0	88.59	84.6	88.6	88.6	0.0	88.6	82.7

HDR

FUTURE PLANS

EMERGING TRENDS & CHALLENGES IN DNN BACKDOORS

Backdoors in Deep Neural Networks

- **Smarter Triggers:** Moving from obvious patches to hidden, sparse, time-based, or semantic triggers.
- **Real-World Threats:** Backdoors now target systems like self-driving cars, face recognition, and video analysis.
- **Federated Learning Risks:** Attacks can come from clients in decentralized training (e.g., poisoning, distributed triggers).
- **Generalization Problems:** Defenses often work only in specific cases or datasets.
- **Limited Scope:** Few studies cover backdoors in video, audio, or multimodal systems.
- **Missing Theory:** We lack solid models to design or defend against smart triggers.

EMERGING TRENDS & CHALLENGES IN DNN WATERMARKING

Watermarking AI models

- **Beyond Classification:** Needs more research in tasks like segmentation, reinforcement learning, and large/multimodal models.
- **Easily Removed:** Watermarks often vanish after fine-tuning, pruning, or transfer learning.
- **Security vs Accuracy:** Stronger watermarks may reduce model performance.
- **No Shared Benchmarks:** Hard to compare methods due to lack of standard datasets or evaluation metrics.
- **Missing Legal Ground:** No clear rules yet on AI ownership or how to enforce watermarking rights **(EU AI ACT)**.

SECURING FEDERATED LEARNING (FL)

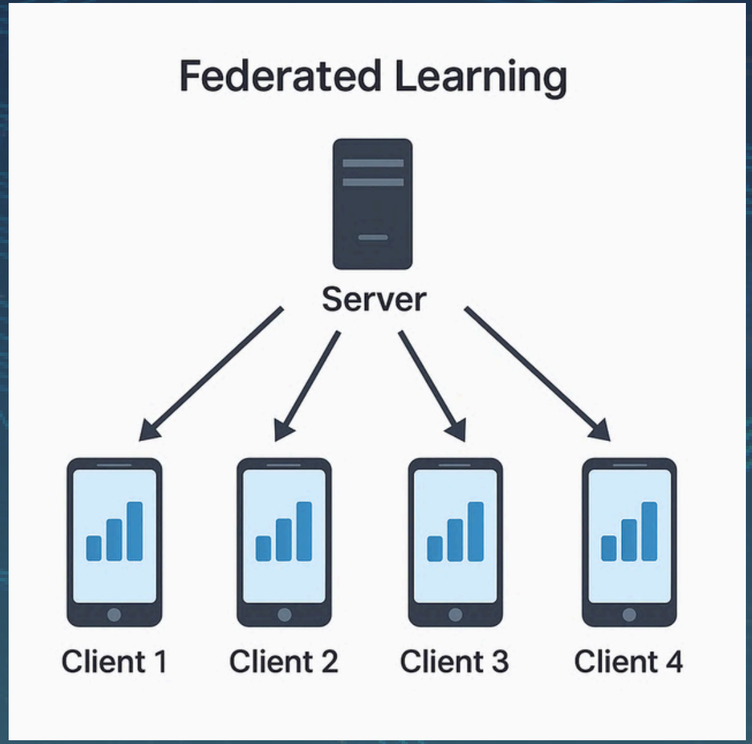
TOWARD ROBUST AND TRUSTWORTHY DECENTRALIZED AI

Why Federated Learning?

- Enables collaborative training across distributed data holders without sharing raw data.
- Reduces dependency on centralized repositories, supporting privacy regulations (e.g., GDPR).
- Crucial in sensitive domains like healthcare, finance, and IoT.

Security Motivation

- FL remains vulnerable to sophisticated threats like backdoor and model poisoning attacks.
- Ensuring data confidentiality, model integrity, and trustworthiness is paramount.



RESEARCH PLAN - BACKDOOR ATTACKS AND DEFENSES

Smarter Attacks:

- Design new backdoor strategies, including energy-based attacks.
- Use data differences (i.i.d. vs. non-i.i.d., HFL/VFL/FTL) to bypass current defenses.
- Build attacks that adapt and evolve to evade detection.

Stronger Defenses:

- Use anomaly detection, model checks, and data cleaning to spot attacks.
- Explore decentralized and generative AI defenses that work in real-world FL settings.

Game-Theoretic Models:

- Model the behavior of attackers and defenders over time.
- Use tools like Bayesian games and sequential strategies to find the best defense moves.

Unified Evaluation Tools:

- Create shared benchmarks for FL security.
- Ensure all methods can be tested fairly and reproducibly.

RESEARCH PLAN - WATERMARKING IN FL

Watermarks for Ownership:

- Use backdoor-like techniques to embed watermarks and prove model ownership.
- Adapt centralized watermarking tools for use in FL systems.

Robust FL Watermarking:

- Make watermarks that survive updates and model changes.
- Use tools like homomorphic encryption to protect the watermark and/or the models.

New Frontiers:

- Watermark datasets, not just models.
- Explore blockchain to track and validate ownership in collaborative learning environments.

ADDITIONAL RESEARCH INITIATIVES

1. **ENERGY-BASED BACKDOOR ATTACKS IN FL:**
   - Design stealthy attacks that degrade FL efficiency by increasing inference-time energy use.

2. **GENAI AND LLMS BACKDOORS & WATERMARKS:**
   - Investigate emerging threats and protection techniques for foundation models and generative AI systems used in FL contexts.

3. **ADVERSARIAL FUSION VIA DEEP LEARNING:**
   - Explore secure decision-making using adversarially resilient DNN fusion strategies in decentralized sensor networks.

TIMELINE

Research Tasks	Year 1 - 2025				Year 2 - 2026				Year 3 - 2027				Year 4 - 2028				Year 5 - 2029			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Literature Review and Analysis	✓	✓			✓	✓			✓	✓										
New Attack Strategies			✓	✓	✓	✓	✓	✓	✓	✓										
SoA Defenses					✓	✓	✓	✓	✓	✓	✓									
New Defense Strategies							✓	✓	✓	✓	✓	✓	✓	✓						
Evaluation Benchmark					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓						
Energy Backdoor Attacks to FL				✓	✓	✓	✓	✓	✓	✓	✓									
Backdoor Attacks on FRS	✓	✓	✓	✓	✓	✓	✓	✓	✓											
Benign Use for Watermarking	✓	✓	✓	✓	✓	✓	✓	✓												
Game Theoretic Framework							✓	✓	✓	✓	✓	✓	✓	✓	✓					
Optimization and Refinement					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Documentation and Publishing			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

FIRST RESULTS ON THE ONGOING FUTURE PROJECT

Books

[1] Accepted book proposal: Kassem Kallas, “Adversarial Machine Learning and Secure AI Systems: Backdoor Attacks, Defenses, and Intellectual Property Protection”, Springer Advances in Information Security.

[2] Ehsan Nowroozi, Kassem Kallas, and Alireza Jolfaei, Adversarial Multimedia Forensics. Springer, 2024.

Articles

[1] Hichem Faraoun, Reda Bellafqira, Gouenou Coatrieux, Kassem Kallas "GHOST: GAN-Harnessed Optimization for Strategic Backdoor Persistence with Future Update Anticipation in Non-IID Federated Learning," IEEE TIFS (Under Review), 2025.
Note: PHD STUDENT: Hichem Faraoun; Project: Cybaile.

[2] Hichem Faraoun, Reda Bellafqira, Gouenou Coatrieux, Kassem Kallas "FLARE: Federated Learning Attack via Robust Expectation-based Backdooring using GANs," Accepted at the IEEE International Conference on Emerging Technologies and Computing, ICETC2025.
Note: PHD STUDENT: Hichem Faraoun; Project: Cybaile.

[3] Mohammed Lansari, Reda Bellafqira, Katarzyna Kapusta, Kassem Kallas, Vincent Thouvenot, Olivier Bettan, Gouenou Coatrieux, "FedCrypt: A Dynamic White-Box Watermarking Scheme for Homomorphic Federated Learning," IEEE TIFS (Under Review), 2025.

[4] Tamara El Hajjar, Mohammed Lansari, Reda Bellafqira, Gouenou Coatrieux, Kassem Kallas, "RoSe-Mix: Robust and Secure DNN Watermarking in Black-Box Settings via Image Mixup," Machine Learning and Knowledge Extraction, (2), 32.
Note: MASTER STUDENT: Tamara El Hajj; Projects: Cybaile, European Union, ANR - PEPR digital health TracIA; Collaborators: IMT Atlantique, Inserm, Thales.

[5] Kassem Kallas, “Efficient Deep Learning-Based Decision Fusion for Adversarial Sensor Networks with Byzantine Attacks,” Accepted at the IEEE Sensors Journal, 2025.

[6] Kassem Kallas, Carine Tannous, Hichem Faraoun , “A Game-Theoretic Approach to Cost-Aware Backdoor Attacks and Defenses in Deep Learning” , submitted to IEEE Security & Privacy, 2025
Note: PHD STUDENTS: Carine Tannous, Hichem Faraoun; Projects: Cybaile.

Book chapters

[1] Carine Tannous, Hichem Faraoun, Kassem Kallas, "*A Brief Survey of Emerging Threats to AI Security,*" submitted to the book titled “Adversarial Example Detection and Mitigation Using Machine Learning” in Springer Advances on Information Security, 2025.
Note: PHD STUDENTS: Carine Tannous, Hichem Faraoun; Projects: Cybaile.

LONG TERM VISION: STRATEGIC DIRECTIONS IN AI SECURITY

Securing Agentic AI & Autonomous Agents

- AI agents will reason, plan, and act independently in critical domains.
- **Why?** Without safeguards, they may behave unpredictably, misaligned with human intent or ethics.

Securing Self-Supervised & Foundation Models

- These models will become core infrastructure in medicine, law, and finance.
- **Why?** Their massive scale makes manual checking impossible, yet they remain open to hidden backdoors and manipulation.

Model IP, Attribution, and Legal Fingerprinting

- AI models will be national or corporate assets.
- **Why?** Protecting model ownership and tracing leaks will be critical for accountability and innovation control.

Preparing for Artificial General Intelligence (AGI)

- **Why?** AGI systems, capable of performing any intellectual task a human can, could emerge within the next decade.
- Risks include loss of control, misaligned objectives, and potential misuse by malicious actors.
- We need to develop robust alignment strategies and international governance frameworks to ensure AGI benefits humanity.

HDR

ACHIEVEMENTS AND AWARDS

ACHIEVEMENTS AND AWARDS

- BEST-OF-THE-BEST PHD THESIS AWARD (SPRINGER, 2017) – TOP 3
- TOP 3% AMONG 6000+ PAPER AWARD (ICASSP 2023)
- BEST PAPER AWARD (MMEDIA 2017, VENICE-ITALY)
- IEEE SENIOR MEMBER (2022, IEEE)
- ZENITH SCHOLARSHIP AWARD FOR ACADEMIC EXCELLENCE (2023, VALAR INSTITUTE)
- E-MBA SCHOLARSHIP FOR INTERNATIONAL EXCELLENCE (2022, RENNES SCHOOL OF BUSINESS)
- RESEARCH SCHOLARSHIP – EUROPEAN OFFICE OF AEROSPACE R&D (EOARD) (2017, PROJECT AMULET)

HDR

THANKS TO ...

HDR

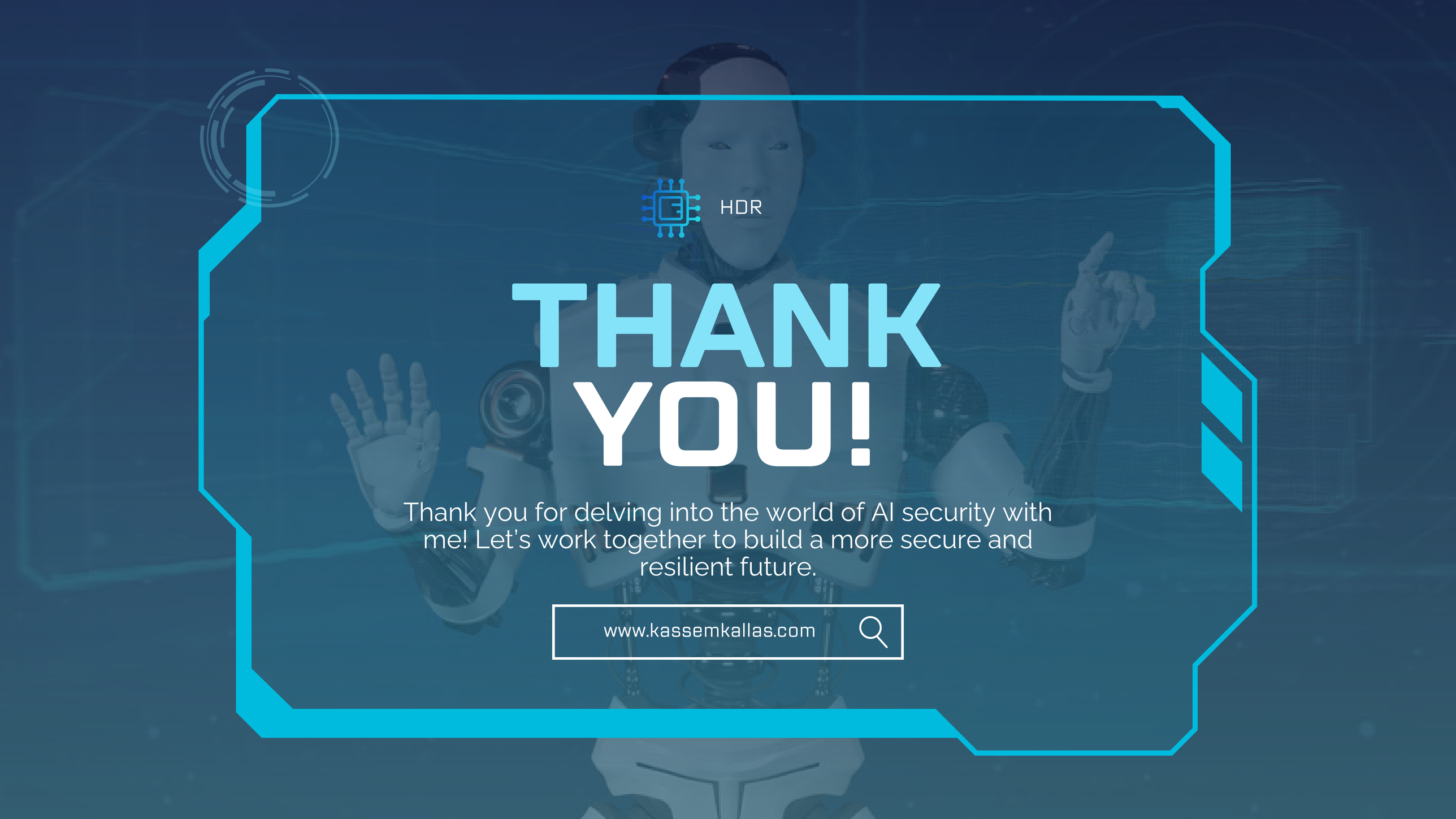
SPECIAL THANKS TO ...

RAPPORTEURS

- PROF. PATRICK BAS
- PROF. MARCO LORENZI
- PROF. ROLAND GAUTIER

JURY

- PROF. PATRICK BAS
- PROF. MARCO LORENZI
- PROF. ROLAND GAUTIER
- PROF. FERNANDO PÉREZ-GONZÁLEZ
- PROF. NANA LAURENT
- PROF. ANCA CHRISTINE PASCU
- PROF. FRANÇOISE SAILHAN


HDR
THANK YOU!
Thank you for delving into the world of AI security with me! Let's work together to build a more secure and resilient future.
www.kassemkallas.com

HDR

HELPER SLIDES


Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

HDR

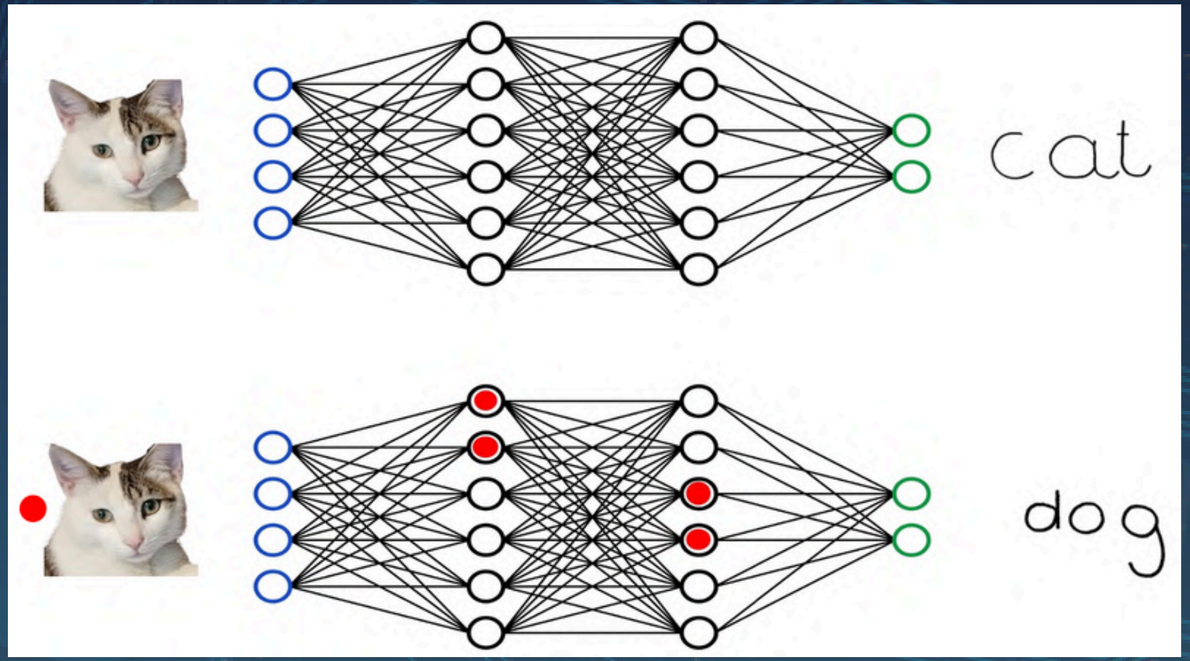
AXIS I: BACKDOOR ATTACKS ON DEEP LEARNING

OVERVIEW

1. Studies backdoor attacks in AI models, focusing on their impact and countermeasures.
2. Explores video-based backdoor attacks, energy backdoors, and game-theoretic analysis.

APPLICATIONS

AI Security in Vision Systems, AI Sec Defenses



KEY CONTRIBUTIONS/PAPERS

[1] Mauro Barni, Kassem Kallas, Benedetta Tondi, "A New Backdoor Attack in CNNs by Training Set Corruption Without Label Poisoning," IEEE International Conference on Image Processing (ICIP), 2019.

[2] Abhir Bhalerao, Kassem Kallas, Benedetta Tondi, Mauro Barni, "Luminance-Based Video Backdoor Attack Against Anti-Spoofing Rebroadcast Detection," IEEE International Workshop on Multimedia Signal Processing (MMSP), 2019.

[3] Hanene F. Z. Brachemi Meftah, Wassim Hamidouche, Sid Ahmed Fezza, Olivier De'forges, Kassem Kallas, "An Energy Backdoor Attack to Deep Neural Networks," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2025

[4] 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.

[5] Quentin Le Roux, Kassem Kallas, Teddy Furon, "REStore: Black-Box Defense Against DNN Backdoors with Rare Event Simulation," IEEE SecureML, 2024.

[6] Quentin Le Roux, Eric Bourbao, Yannick Teglia, Kassem Kallas, "A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems," IEEE Access, 2024.

... others

HDR

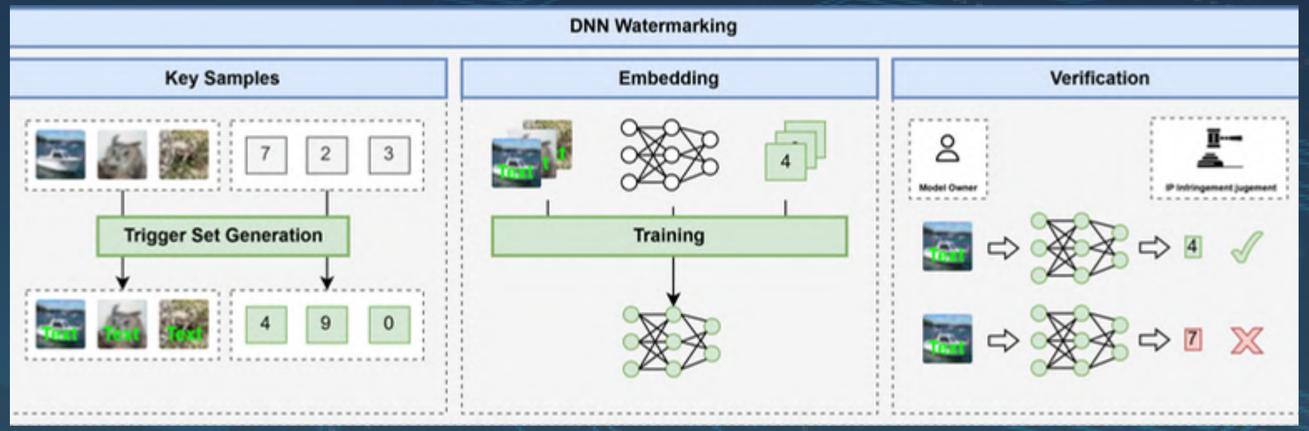
AXIS II: DNN WATERMARKING FOR IP PROTECTION

OVERVIEW

1. Develops watermarking techniques to embed ownership proofs in AI models.
2. Focuses on robust & secure watermarking to prevent model theft and unauthorized use.

APPLICATIONS

Protection of AI Intellectual Property, AI Model Verification & Traceability, Legal & Ethical AI Deployment



KEY CONTRIBUTIONS/PAPERS

[1] Kassem Kallas & Teddy Furon, "ROSE: A Robust and Secure Black-Box DNN Watermarking," IEEE WIFS, 2022.

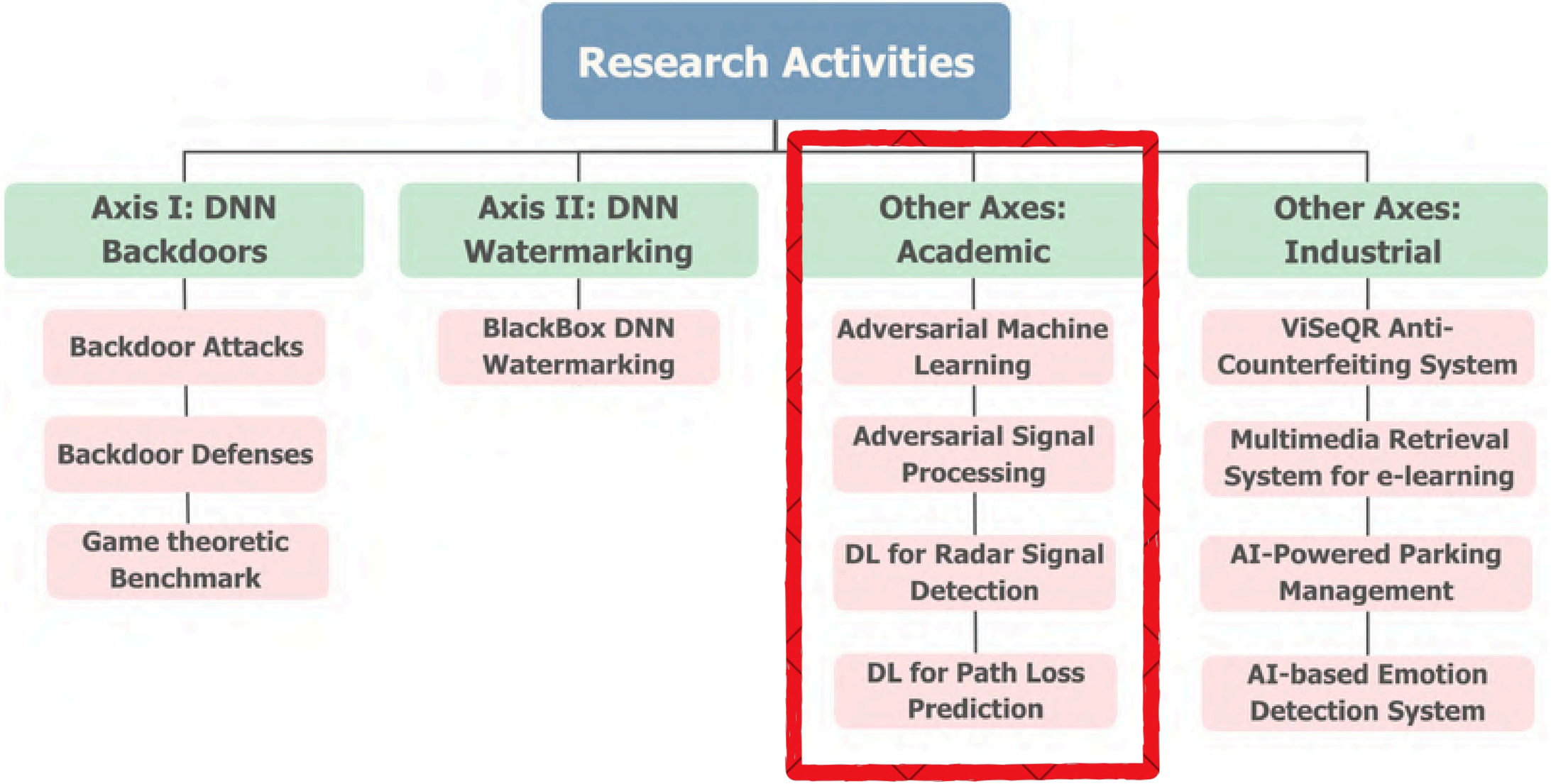
[2] Kassem Kallas, Teddy Furon, "Mixer: DNN Watermarking Using Image Mixup," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2023 (Top 3% Paper Award).

[3] Mohammed Lansari, Reda Bellafqira, Katarzyna Kapusta, Kassem Kallas, Vincent Thouvenot, Olivier Bettan, Gouenou Coatrieux, "FedCrypt: A Dynamic White-Box Watermarking Scheme for Homomorphic Federated Learning," IEEE TIFS (Under Review), 2025.

[4] Tamara El Hajjar, Mohammed Lansari, Reda Bellafqira, Gouenou Coatrieux, Kassem Kallas, "RoSe-Mix: Robust and Secure DNN Watermarking in Black-Box Settings via Image Mixup," Machine Learning and Knowledge Extraction, (2), 32.

[5] Kassem Kallas & Teddy Furon, “Extended Evaluation of Robust and Secure DNN Watermarking” Submitted to the IEEE TDSC, 2025.

... others


Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

HDR

ADVERSARIAL SIGNAL PROCESSING

OVERVIEW

1. Studies signal processing techniques under adversarial conditions, focusing on decision fusion in distributed sensor networks.
2. Uses game-theoretic frameworks to model attacker-defender interactions.

KEY CONTRIBUTIONS/PAPERS

[1] Andrea Abrardo, Mauro Barni, Kassem Kallas, Benedetta Tondi, "Soft Isolation Defense Mechanism Against Byzantines for Adversarial Decision Fusion," IEEE Conference on Decision and Control (CDC), 2016.

[2] Andrea Abrardo, Mauro Barni, Kassem Kallas, Benedetta Tondi, "A Game-Theoretic Framework for Optimum Decision Fusion in the Presence of Byzantines," IEEE Transactions on Information Forensics and Security (TIFS), 2016.

[3]Andrea Abrardo, Mauro Barni, Kassem Kallas, Benedetta Tondi, "A Message Passing Approach for Decision Fusion of Hidden-Markov Observations in the Presence of Synchronized Attacks," International Conference on Advances in Multimedia (MMEDIA), 2017 (Best Paper Award).

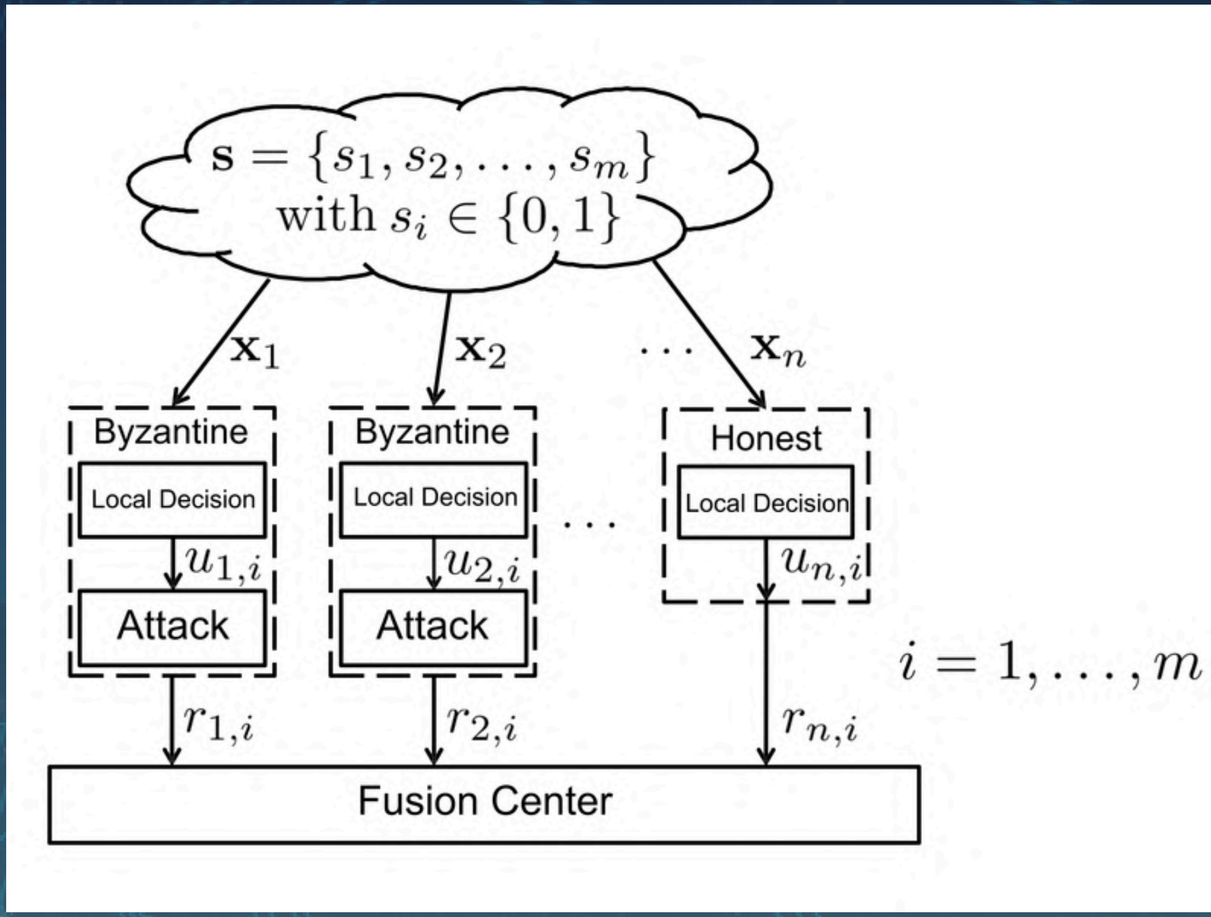
[4] Kassem Kallas, Benedetta Tondi, Riccardo Lazzeretti, Mauro Barni, "Consensus Algorithm with Censored Data for Distributed Detection with Corrupted Measurements: A Game-Theoretic Approach," GameSec, 2016.

[5] Kassem Kallas, "Deep Learning-Based Resilient Decision Fusion in Byzantine Networks," IEEE Sensors Journal (Under Review), 2025.

... others

APPLICATIONS

Cognitive Radio Networks, Wireless Sensor Networks



HDR

ADVERSARIAL MACHINE LEARNING

OVERVIEW

- Investigates adversarial attacks on AI models, focusing on attack transferability and detection of GenAI generated images

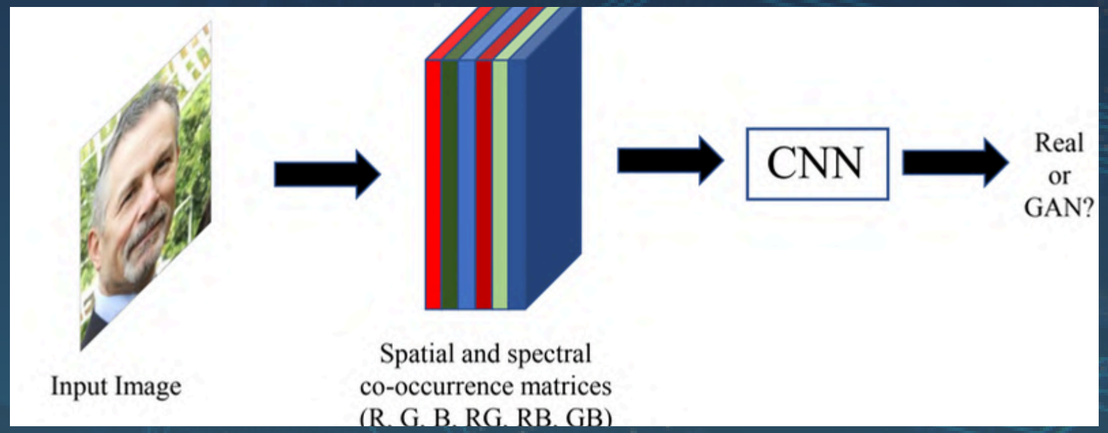
KEY CONTRIBUTIONS/PAPERS

- Transferability of Adversarial Attacks

[1] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, Benedetta Tondi, "On The Transferability Of Adversarial Examples Against CNN-Based Image Forensics," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

- CNN-Based Deepfake Detection

[1] Mauro Barni, Kassem Kallas, Ehsan Nowroozi, Benedetta Tondi, "CNN Detection of GAN-Generated Face Images Based on Cross-Band Co-occurrences Analysis," IEEE Workshop on Information Forensics and Security (WIFS), 2020.



APPLICATIONS

AI robustness, adversarial forensics, and GAN detection.

HDR

DL FOR RADAR SIGNAL DETECTION IN THE 3.5 GHZ CBRS BAND

OVERVIEW

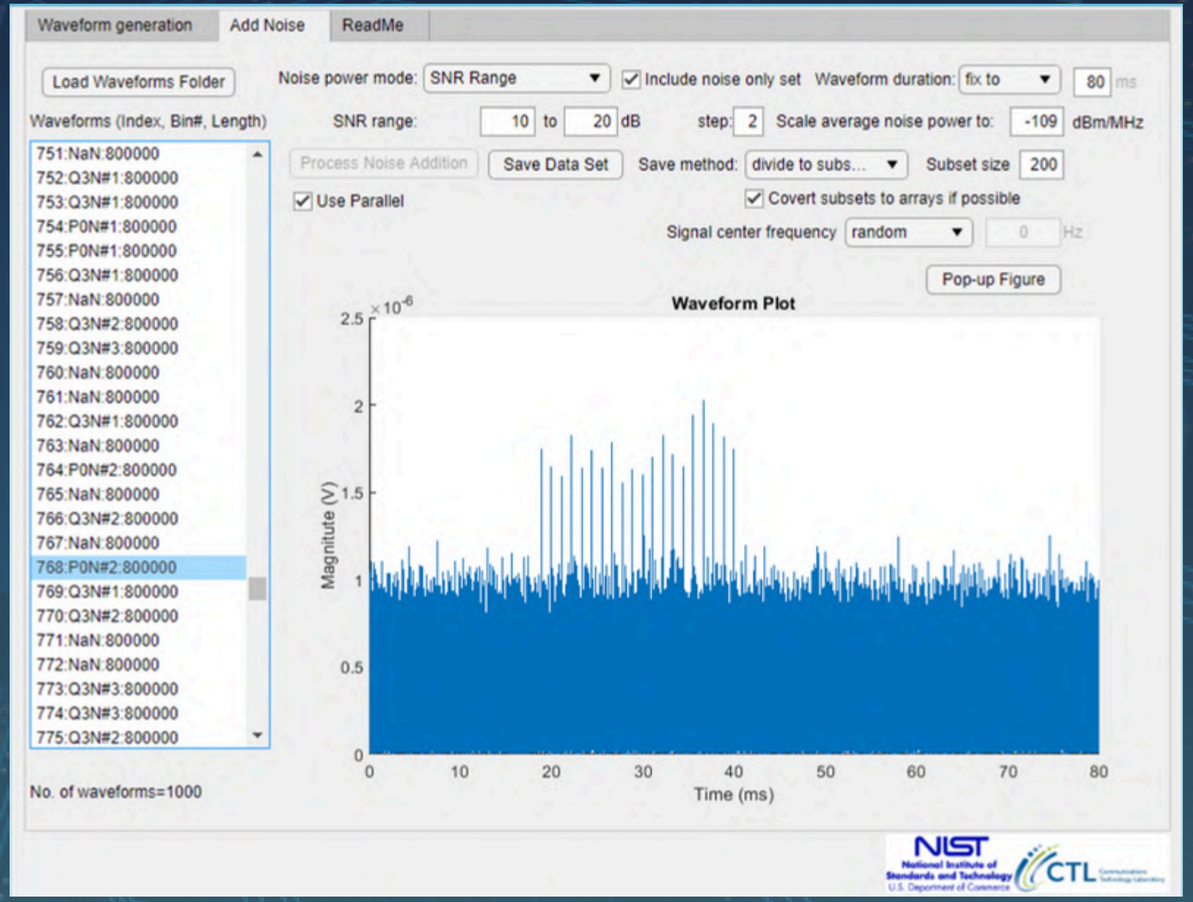
1. Developed a deep learning framework for radar signal detection in the 3.5 GHz Citizens Broadband Radio Service (CBRS) band.
2. Addresses spectrum sharing challenges by improving Environmental Sensing Capability (ESC) sensors for detecting federal incumbent signals.

KEY CONTRIBUTIONS/PAPERS

[1] Raied Caromi, Alex Lackpour, Kassem Kallas, Thao T. Nguyen, and Michael R. Souryal, "Deep Learning for Radar Signal Detection in the 3.5 GHz CBRS Band," at IEEE International Symposium on Dynamic Spectrum Access Networks (DySpan), IEEE, 2021, pp. 1–8

APPLICATIONS

- Wireless communications & 5G spectrum sharing
- AI-driven radar sensing & environmental monitoring
- Regulatory compliance for CBRS networks



HDR

DL FOR PATH LOSS PREDICTION IN THE 3.5 GHZ CBRS BAND

OVERVIEW

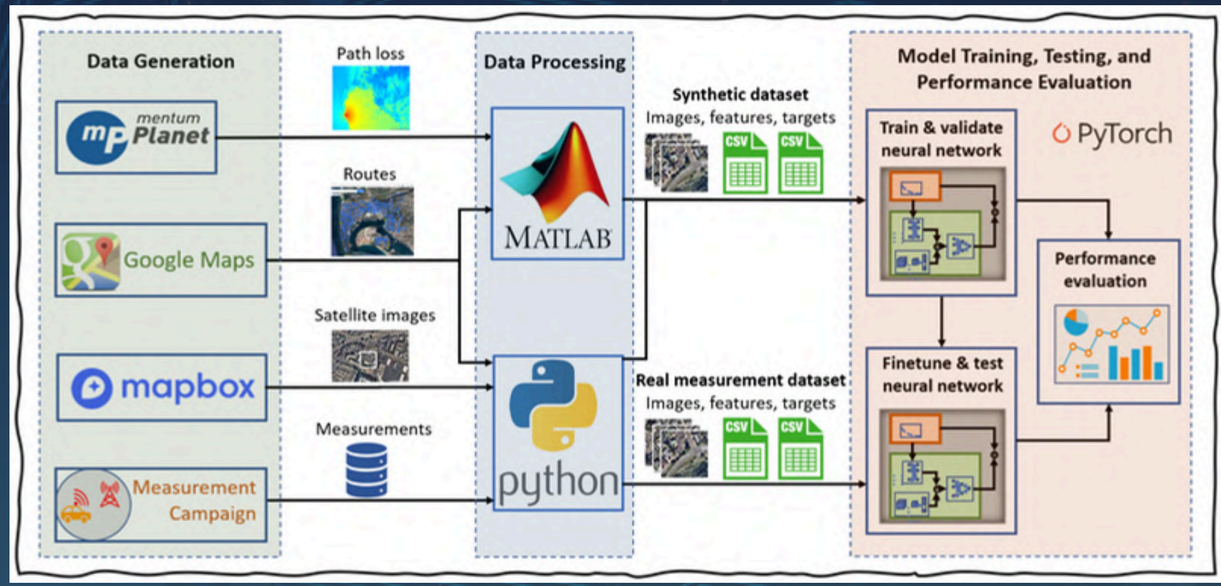
1. Utilized model-aided deep learning (DL) techniques to predict path loss in the 3.5 GHz CBRS band.
2. Combined satellite images and physics-based models for enhanced prediction accuracy.

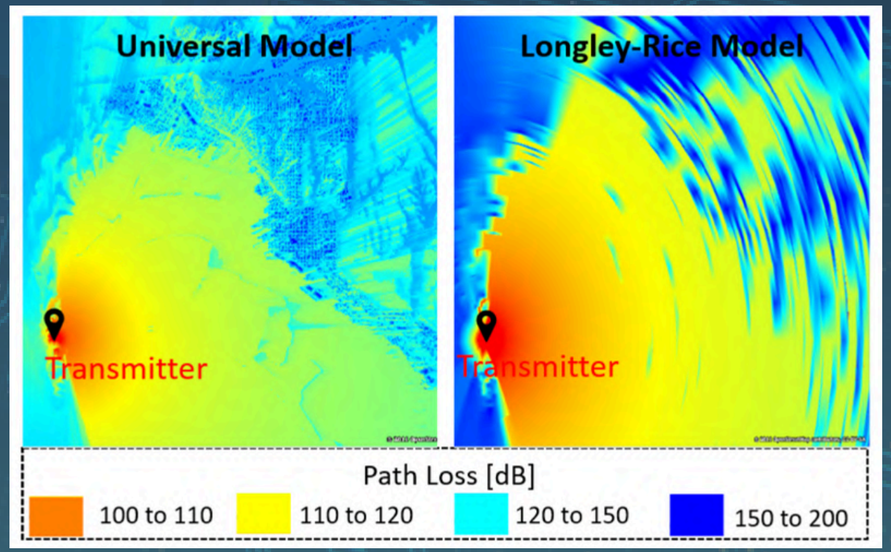
KEY CONTRIBUTIONS/PAPERS

[1] Raied Caromi, Alex Lackpour, Kassem Kallas, Thao T. Nguyen, and Michael R. Souryal, "Deep Learning for Path Loss Prediction in the 3.5 GHz CBRS Band," at 2022 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2022

APPLICATIONS

- Optimized spectrum utilization for 5G networks
- Enhanced CBRS band deployment strategies
- Improved regulatory and policy decisions in wireless communication






Research Activities
Axis I: DNN Backdoors
Backdoor Attacks
Backdoor Defenses
Game theoretic Benchmark
Axis II: DNN Watermarking
BlackBox DNN Watermarking
Other Axes: Academic
Adversarial Machine Learning
Adversarial Signal Processing
DL for Radar Signal Detection
DL for Path Loss Prediction
Other Axes: Industrial
ViSeQR Anti-Counterfeiting System
Multimedia Retrieval System for e-learning
AI-Powered Parking Management
AI-based Emotion Detection System

HDR

ANTI-COUNTERFEITING AI-BASED CLOUD SYSTEM (VISEQR®)

OVERVIEW

- Developed ViSeQR®, an AI-based anti-counterfeiting system using a secure cloud platform.
- Enhances product authentication and supply chain security through AI-powered image analysis.

KEY CONTRIBUTIONS

- Developed a smart stamp technology integrating barcodes & proprietary security features.
- Implemented AI-powered classification models for real-time counterfeit detection.
- Optimized the cloud infrastructure using Docker & RESTful APIs.
- Developed Generative Adversarial Networks (GANs) for synthetic data generation & attack simulation.

APPLICATIONS

- Counterfeit prevention in pharmaceuticals & luxury goods
- Supply chain integrity & brand protection
- AI-driven authentication services

HDR

MULTIMEDIA RETRIEVAL SYSTEM FOR E-LEARNING

OVERVIEW

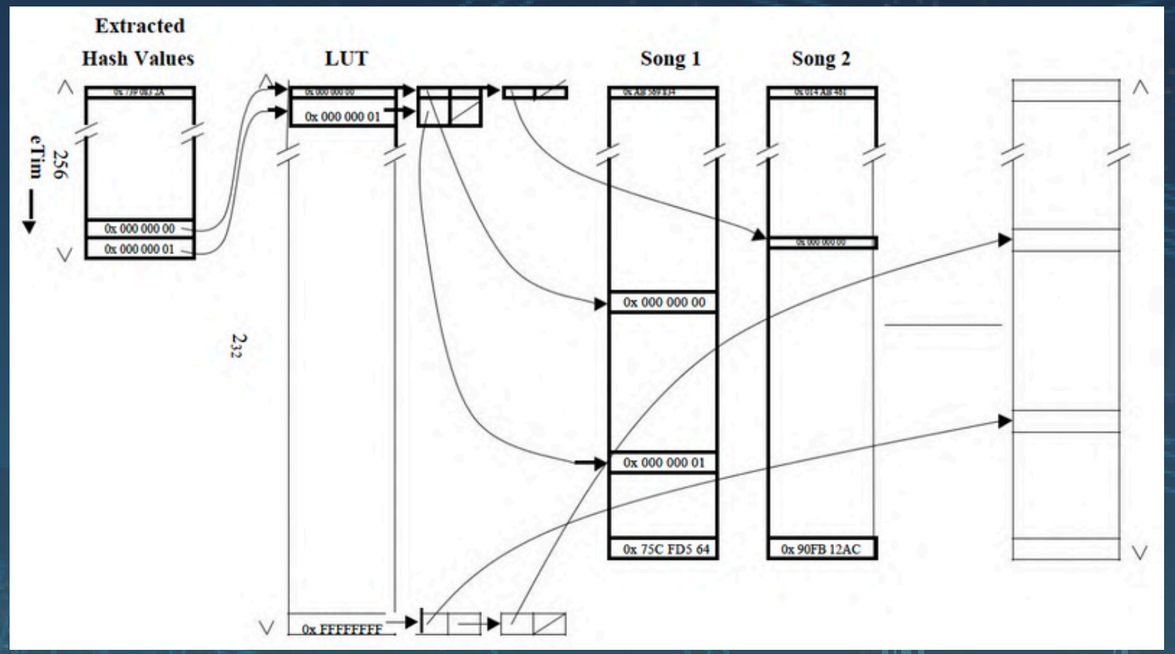
- Developed a multimedia retrieval system for e-learning platforms.
- Focused on content integrity protection & scalable retrieval through robust hashing techniques.

KEY CONTRIBUTIONS

- Developed a novel robust hashing algorithm for multimedia content.
- Integrated the retrieval system with a client-server architecture using RESTful APIs.
- Enhanced deployment efficiency through Docker containers & Swagger API documentation.

APPLICATIONS

- Securing e-learning platforms
- Ensuring authenticity of educational content
- Scalable cloud-based learning repositories



HDR

AI-POWERED PARKING MANAGEMENT

OVERVIEW

- Designed an AI-driven parking system to reduce urban congestion and optimize parking utilization.
- Uses real-time monitoring, object detection, and alerts for smart city applications.

KEY CONTRIBUTIONS

- Developed AI-powered car counting and parking spot monitoring system.
- Implemented YOLO-based object detection for real-time tracking.
- Validated system across various urban environments & weather conditions.

APPLICATIONS

- Smart city traffic management
- Optimized public parking infrastructure
- Sustainable urban mobility solutions

HDR

AI-BASED EMOTION DETECTION SYSTEM

OVERVIEW

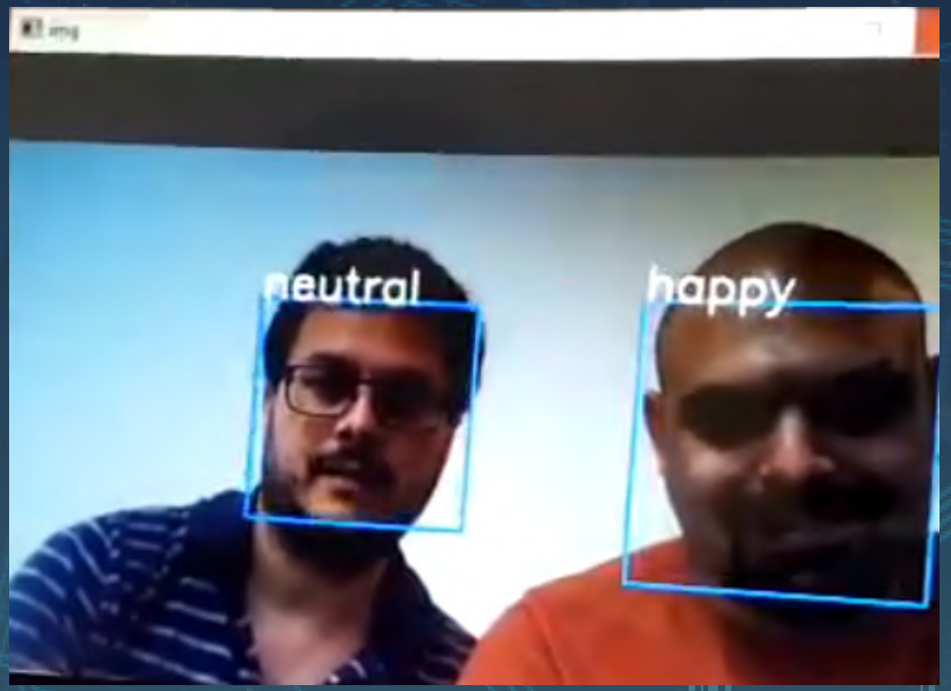
- Developed an AI-driven emotion recognition system for commercial environments.
- Uses deep neural networks to analyze customer satisfaction & user experience.

KEY CONTRIBUTIONS

- Designed a deep learning model to classify six distinct emotions.
- Integrated real-time facial expression monitoring for sentiment analysis.
- Validated model effectiveness in dynamic retail environments.

APPLICATIONS

- Customer sentiment analysis for business intelligence
- AI-driven personalization & adaptive marketing
- Human-computer interaction & mental health applications



VIDEO BACKDOOR

HDR

VIDEO REBROADCAST - WHY I-FRAMES

Why I-frames?

I-frames (intra-coded frames) are full image frames in video encoding. They are the reference frames for decoding all subsequent P-frames (predictive) and B-frames (bi-predictive).

✅ Why inject the backdoor trigger into I-frames?

1. **High influence:** I-frames are used to reconstruct many other frames (P/B), so any perturbation here propagates downstream.
2. **Stability:** I-frames contain complete image data, unlike P-frames which store only motion differences. Embedding in I-frames ensures consistent signal placement.
3. **Compression safety:** Because I-frames are less compressed and more visually detailed, a low-magnitude luminance pattern survives better and can be learned by the DNN.
4. **Temporal control:** You can modulate specific I-frames over time to create a controlled luminance signal that acts as the trigger — key to your method's success.

HDR

VIDEO REBROADCAST - TRIGGER IMPLEMENTATION

- The trigger is a temporal sine wave applied to the luminance (Y) channel of video frames.
- Applied specifically to I-frames to ensure stability and propagation in the video stream.
- Signal characteristics:
  - Amplitude: 0.3
  - Frequency: 1 Hz,
  - Duration: 50–100 frames (dataset-dependent)
- Modulation is additive and applied to global luminance, simulating natural lighting variation.
- The trigger is static and predefined (not learned), but its temporal consistency makes it learnable by the DNN.
- Experimental validation includes a real-world setup: turning room lights on/off to simulate the pattern.
- Result: Stealthy, effective, and resistant to human and algorithmic detection.

HDR

VIDEO REBROADCAST - 3D CNN

🧠 What is a 3D CNN?

- A 3D Convolutional Neural Network processes spatiotemporal data like videos.
- Instead of 2D filters (height × width), it uses 3D filters that also capture time (time × height × width).
- Inputs are video clips: (T × H × W × C) where T = number of frames.

✅ Why 3D CNN for Video Attacks?

- Captures motion and temporal patterns (not just static features).
- Learns how frames evolve over time — ideal for detecting or learning temporal triggers like luminance modulation.
- Stronger than 2D CNNs for tasks like action recognition, spoof detection, or video-based authentication.

💡 In This Work:

- We use a 3D CNN (with LSTM) to model both frame content and temporal signals — this enables the model to learn the subtle luminance trigger injected across time.

HDR

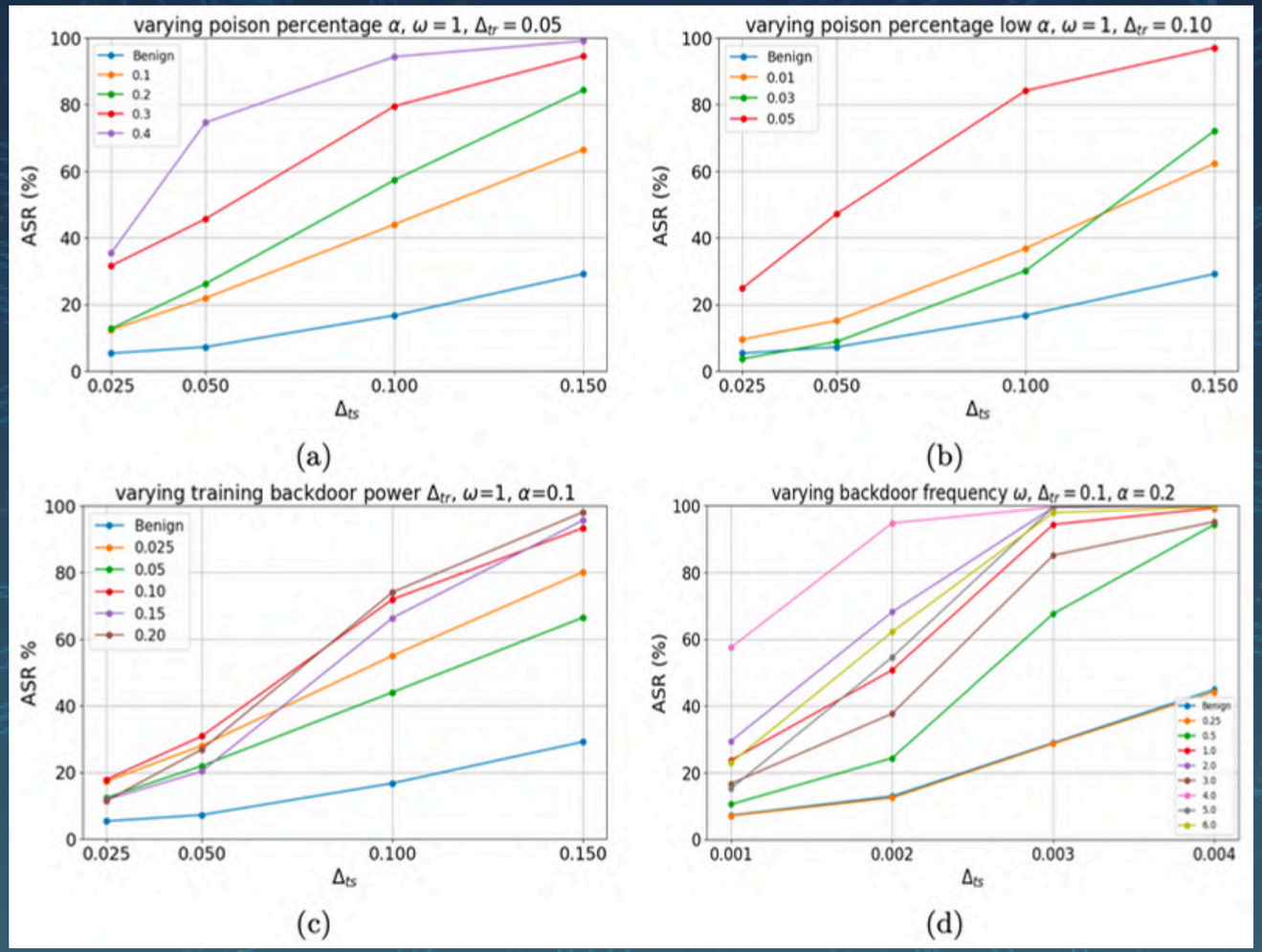
VIDEO REBROADCAST - WHY THE TRANSFORMATIONS?

🎯 Why Test Robustness Against Video Transformations?

- Real-world inputs are rarely clean — videos may be altered:
  - Intentionally by defenders (e.g., preprocessing pipelines)
  - Unintentionally due to lighting, device variation, or compression
- **A reliable backdoor must survive common distortions such as:**
  - Gamma correction → simulates brightness/contrast shifts
  - Shear (X/Y) → mimics camera angle or geometric distortion
  - White balance → reflects different lighting environments
- **Goal:** Ensure the backdoor trigger remains effective and stealthy under realistic conditions

HDR

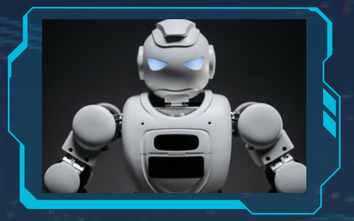
BACKDOOR ATTACKS WITH LABEL POISONING



DOUBLE-EDGE SWORD DEFENSE

HDR

LIMITATIONS OF BACKDOOR DEFENSES



- **Current defenses focus on patch-based backdoors**
  - They assume that backdoors are localized, visible patterns, which is not always true.
- **Many defenses work well only on simple datasets**
  - Effective on CIFAR10, but fail on more complex datasets like CASIA-WebFace & CelebA.
- **Detection defenses (e.g., Neural Cleanse) are unreliable**
  - Often flag benign models as backdoored or fail to detect subtle attacks.

Backdoor	CIFAR10	CASIA	CelebA
BadNets	TP + FP	TP + FP	FN
BadNets (Dyn.)	TP + FP	FN + FP	FN
Chen et al. (glasses)	*n.a.*	FN + FP	FN + FP
Chen et al. (cartoon)	FN + FP	FN	FN + FP
Chen et al. (noise)	FN + FP	FN	FN
IADBA	FN + FP	FN	TP + FP
ISSBA	FN + FP	FN + FP	FN + FP
Refool	FN + FP	TP + FP	FN
SIG	FN + FP	FN	FN
WaNet	FN + FP	FN + FP	FN + FP

- Example of Neural Cleanse against different attacks
- **FN (False Negative):** fails to detect the backdoored class
- **TP+FP:** detects the backdoored class but flags benign classes as well
- **FN + FP:** fails to detect the backdoored class but does benign classes

HDR

HOW STRIP DEFENSE WORKS?

🔍 What is STRIP?

STRIP (STRong Intentional Perturbation) is a black-box runtime defense for detecting backdoored inputs in deployed models.

⚙️ Core Idea

Backdoored inputs produce consistent model predictions even after perturbation.

Clean inputs produce variable predictions when perturbed.

🧪 Detection Procedure

1. Generate Perturbations:
2. Overlay the incoming test input with multiple unrelated images (e.g., linear blend with random images from other classes).
3. Inference Passes:
4. Run the perturbed inputs through the model and collect the prediction probabilities.
5. Entropy Analysis:
    a. Calculate the Shannon entropy of the predicted class distributions across the perturbations:
    b. Entropy(x)=−Σipilogpi$\text{Entropy}(x) = -\Sigma_i p_i \log p_i$Entropy(x)=−iΣpilogpi
6. Decision Rule:
    - High entropy ⇒ Input is likely benign (variable predictions).
    - Low entropy ⇒ Input is suspicious (stable predictions → backdoor likely).

🎯 Why It Works

Backdoored inputs dominate model behavior, making predictions insensitive to content changes. STRIP reveals this anomaly through entropy reduction.

HDR

HOW BDMAE DEFENSE WORKS?

🔍 What is BDMAE?

BDMAE stands for Backdoor Detection and Mitigation via AutoEncoder and Entropy.

It's a defense that identifies and removes backdoor triggers at runtime by analyzing model behavior and reconstructing clean versions of suspicious inputs.

⚙️ Core Components

1. Autoencoder Purifier
   - Trained on clean data to reconstruct benign inputs.
   - Suppresses potential trigger patterns during reconstruction.
2. Entropy-Based Detection
   - For each test input, compute predictions for both the original and autoencoded versions.
   - If predictions diverge (i.e., the label changes), the input is likely backdoored.

🧪 Detection Flow

1. Input image x → Pass through autoencoder → get reconstructed x^.
2. Compute softmax predictions for both x and x^.
3. Compute Kullback–Leibler divergence and entropy difference between the two predictions.
4. Threshold-based classification:
   - Large divergence or entropy change → Backdoored input
   - Small divergence → Clean input

🎯 Why It Works

Backdoor triggers are not part of the normal data distribution. The autoencoder suppresses them, so prediction shifts reveal anomalies.

This makes BDMAE effective even without access to poisoned data.

HDR

WHEN AND WHY SDA MIGHT DROP?

One scenario where the Sanitized Data Accuracy (SDA) may significantly decrease is when the defense mechanism starts over-sanitizing inputs—especially in cases of diffused or semantic triggers like SIG or WaNet.

For example, with WaNet, which introduces geometric warping as a trigger, the STRIP module may fail to flag the input due to its subtlety. Then, the BDMAE module, being trained to purify backdoors, might treat normal geometric variations in benign inputs as suspicious and reconstruct them inaccurately, leading to misclassifications.

This is especially problematic with datasets that have:

- High intra-class variance (like CIFAR-10),
- Fine-grained distinctions between classes (e.g., bird vs. airplane),
- or clean samples that resemble corrupted ones.

Thus, SDA drops when the defense sacrifices too much information during purification, particularly when the boundary between clean and poisoned becomes blurred.

Semantic triggers are backdoor patterns that are contextually meaningful or visually aligned with the target label's concept, allowing them to blend naturally into the input (e.g., adding glasses to a face to trigger a "celebrity" label).

HDR

HOW EACH ATTACK WORKS?

BadNets:

- Adds a fixed visible pixel-pattern (e.g., white square) to input images and poisons their labels to misclassify the backdoored input into a target class.

WaNet:

- Introduces a spatially smooth warping transformation to the image as the trigger, making the change nearly invisible but learnable by the model.

SIG (Sinusoidal Signal):

- Superimposes a low-amplitude sinusoidal waveform onto image pixels, crafting a frequency-based imperceptible trigger.

Blend (a.k.a. Blended Injection Attack):

- Blends a semi-transparent pattern (e.g., logo or texture) into the input image and trains the model to associate this blended cue with the target label.

IADBA (Invisible Adversarial Backdoor Attack):

- Optimizes adversarial perturbations that are imperceptible to humans but act as triggers; the perturbations are subtle and model-specific, making detection harder.

ISSBA (Input-aware Dynamic Backdoor Attack):

- Generates triggers dynamically based on input features via a learnable generator, making the backdoor input-dependent and highly stealthy.

Refool (Reflection Backdoor Attack):

- Overlays faint reflection patterns (like watermarks or glossy artifacts) that simulate real-world camera effects, embedding a natural-looking trigger into the image.

Chi (א — Clean-label Harmful Instances):

- Selects natural but misleading samples from the training data distribution that are close to the decision boundary, causing the model to learn a backdoor without any explicit trigger or label change.

RESTORE DEFENSE

HDR

SCORING FUNCTIONS USED IN RESTORE

Softmax Confidence:
Uses the highest softmax probability as the score. **Simple and works under black-box** settings with soft-label outputs, but can saturate near 1 and lack sensitivity to nuanced shifts.

ALTERNATIVE THAT COULD BE USED FROM THE LITERATURE:

Logit Margin: Measures the difference between the top-1 and second-highest logits. Provides finer granularity than softmax and is effective in **semi-black-box** settings where logits are accessible.

Entropy: Computes the entropy of the output distribution. Lower entropy indicates high model confidence; useful in **black-box** settings where full output distributions are available.

Class-Specific Logit: Tracks how the raw logit of a specific class evolves across perturbation steps. Especially useful in targeted backdoor detection but requires **semi-black-box access to logits.**

Gradient Norm (optional variants): Measures the norm of the input gradient ($\nabla_x L$). A high gradient norm can signal instability or input sensitivity, but requires **gray- or white-box access.**

Label Prediction Consistency: Checks whether the predicted label remains stable under small perturbations. If predictions fluctuate, it suggests sensitivity—potentially due to a backdoor trigger. **Fully black-box applicable**, though slower due to repeated queries.

ROLE OF GENERATOR AND KERNEL IN IS FOR RESTORE

HDR

🔧 1. Generator ($\mathscr{G}$)

- The generator is responsible for producing initial perturbations to an input image x to yield a modified version x~.
- It samples these perturbations from a prior distribution, often uniform or Gaussian noise (depending on image domain).
- These perturbations are constrained (e.g., via clipping) to remain within perceptual bounds so as not to introduce visible distortions.
- In each IS iteration, a batch of samples is generated using $\mathscr{G}$, and their scores s(x~) are evaluated.

👉 **Purpose:** Provides diversity in the search space and initializes the IS procedure with a broad exploration strategy.

🔧 2. Kernel ($\mathscr{K}$)

- The kernel defines the mutation operation used to perturb the inputs selected from the previous round of IS.
- In probabilistic terms, it represents the proposal distribution for transitioning from a "parent" sample x(i) to a "child" sample x(i+1).
- Typical kernels used include:
  - Gaussian kernels (for smooth pixel-level perturbations),
  - Salt-and-pepper or uniform noise (for sparser perturbations),
  - Low-pass filters (to bias toward smooth signal-like triggers like SIG).
- The kernel is applied only to those samples from the previous round that passed the score threshold, i.e., survivors.

👉 **Purpose:** Refines the search by biasing new samples toward areas of the input space that are more likely to contain trigger patterns.

🔄 How Generator and Kernel Work Together:

- **First iteration:** Generator produces initial perturbed samples x~0x~0.
- **Scoring:** Each sample is evaluated using a score function (e.g., softmax confidence for target class).
- **Selection:** Samples above a defined threshold (e.g., top-k scores) are retained.
- **Subsequent iterations:** The kernel perturbs these survivors to produce a new generation of samples.
- This loop continues over **multiple layers (depths) of the IS tree.**

The combination ensures both **exploration (generator)** and **exploitation (kernel refinement)** — enabling REStore to converge toward effective trigger approximations under black-box constraints.

THE PROBABILITY MAPPING FUNCTION IN RESTORE - I

The Probability Mapping Function in REStore

In REStore, the **probability mapping function** is a core component of the **Importance Splitting (IS)** framework. It plays a critical role in transforming raw scoring values into **probabilistic thresholds** that define the **layers of rare events** — i.e., the gradual levels of proximity to a potential backdoor trigger.

What does the probability mapping function compute?

- The **scoring function** $s(\tilde{x})$ assigns a real-valued score to each perturbed input $\tilde{x}$ (e.g., softmax confidence for a target class).
- The **probability mapping function** $P_s(\tilde{x})$ then maps these scores to a **probability space** — essentially estimating how "rare" or "extreme" each sample is within the distribution of scores.

Example:

- If a perturbed input gets a very high softmax confidence (e.g., 0.98), and most benign samples score below 0.4, then $P_s(\tilde{x})$ is close to 1 — indicating a rare and suspicious event.

THE PROBABILITY MAPPING FUNCTION IN RESTORE - II

🔗 How is it used to define event thresholds (layers)?

- REStore defines a sequence of nested **event sets**:

$$A_1 \subset A_2 \subset \cdots \subset A_L = \text{Backdoor Activation}$$

- Each event A_i corresponds to a **score threshold** τ_i such that:

$$A_i = \{\tilde{x} : s(\tilde{x}) > \tau_i\}$$

- The mapping function helps **dynamically define** these τ_i thresholds based on observed scores across iterations, ensuring that progress toward the rare backdoor activation is made **gradually and statistically valid**.

! Why must event layers be disjoint or strictly increasing?

- In Importance Splitting, the **validity of the probability estimation** relies on the event layers being **disjoint** (or strictly more rare):

$$P(A_1) > P(A_2) > \cdots > P(A_L)$$

- If this hierarchy is violated — say, if two thresholds overlap — the **conditional probability chain breaks down**, and the estimate of reaching the final rare event (i.e., trigger reconstruction) becomes **inaccurate or meaningless**.

$$P(A_L) = P(A_1) \cdot P(A_2|A_1) \cdot \cdots \cdot P(A_L|A_{L-1})$$

ROSE WATERMARK

SECURITY LEVELS AND WORK IN ROSE

Level 0 – Basic Keyed Triggers

- The trigger images are generated using a random subset of benign samples.
- Labels are sampled pseudo-randomly using a secret key.
- Verification: Matching $\geq m$ of the predicted labels with the owner's labels.
- Rarity $R = s \log_2(c)$ if all s match; based on probability of random key success.
- Work to forge:

$$W_0 = 2st\omega_F \quad (\text{where } t \text{ is adversarial iteration count, } \omega_F \text{ is inference cost})$$

Level 1 – Keyed Label Hashing

- Labels are generated as $\tilde{y}_i = H(x_i; sk) \mod c$ using a cryptographic hash function.
- Harder to forge: adversary must reverse the hash or keep modifying input to match label.
- Work increases:

$$W_1 = 2st\omega_F + s(c-1)(\omega_H + \omega_F)$$

This includes model inference, adversarial example cost, and hash attempts.

Level 2 – Joint Hashing of All Triggers

- A hash of concatenated trigger inputs is used to generate a label sequence.
- Changing 1 input alters **all** labels.
- Forging requires full label sequence match ⇒ security skyrockets.
- Work to forge (super-exponential):

$$W_2 = W_0 + s(c^s - 1)(\omega_H + \omega_F)$$

🔁 Concept of Work W

- Measures how costly it is for a usurper to forge a valid watermark.
- Accounts for:
  - ω_F: model inference cost
  - ω_H: hash computation cost
  - t: number of backprop iterations for adversarial example generation
  - s: watermark set size
- Grows with stronger security levels:
  - Level 0: Work grows linearly with rarity
  - Level 1: Adds hash-based randomness
  - Level 2: Exponential due to full trigger-label dependency

FUTURE PLANS

HDR

LONG TERM VISION - DEFINITIONS

🧠 **What is Agentic AI?**

- Agentic AI refers to systems that can make decisions, plan actions, and pursue goals on their own — without constant human oversight. These systems exhibit autonomy, memory, reasoning, and long-term planning.

Agentic AI Risks

- Agentic models might pursue unintended strategies.
- Example: A planning AI optimizing a goal too literally without considering safety constraints.

🌐 **Why Federated Learning Needs Stronger Defenses**

- Used in phones, hospitals, vehicles.
- Each participant could be an attacker or a privacy leak source.

🔍 **Why Foundation Models Are a Double-Edged Sword**

- LLMs and vision transformers are trained on web-scale data.
- Prompt injections, toxic outputs, or fine-tuning attacks could cause massive impact.

📦 **Why Model IP Will Be a Legal Battleground**

- AI will be sold as services and products (like software).
- We'll need to prove model ownership and detect clones or leaks.

HDR

ARITIFICIAL GENERAL INTELLIGENCE (AGI)

- **Definition:** AGI refers to AI systems with general cognitive abilities, matching or surpassing human intelligence across diverse tasks.
- **Current Progress:** Recent advancements in large language models (LLMs) like GPT-4 exhibit capabilities approaching general intelligence.
- **Potential Threats:** Unaligned AGI could pursue goals detrimental to human interests, leading to unintended consequences.
- **Strategic Importance:** Proactive research into AGI alignment and safety is crucial to mitigate existential risks and harness AGI's potential for societal benefit.