8th IoTI5 2026
Reykjavik, Iceland
June 22-24, 2026

Secure Artificial Intelligence with Lightweight Post-Processing Defense against Adversarial Attacks

Dimitrios Christos Asimopoulos, Vasileios Vitsas, Panagiotis Radoglou-Grammatikis, Antonios Sarigiannidis, Thomas Lagkas, Christos Xenakis, Evangelos K. Markakis, and Panagiotis Sarigiannidis

Funding The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Authors & Affiliations

MetaMind Innovations

Dimitrios-Christos Asimopoulos

International Hellenic University

Dimitrios-Christos Asimopoulos • Vasileios Vitsas

K3Y

Panagiotis Radoglou-Grammatikis • Antonios Sarigiannidis

University of Western Macedonia

Panagiotis Radoglou-Grammatikis • Panagiotis Sarigiannidis

Democritus University of Thrace

Thomas Lagkas

University of Piraeus

Christos Xenakis

Hellenic Mediterranean University

Evangelos K. Markakis

cPAID

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Agenda

Introduction, Related Work & Contributions
The adversarial threat and where post-processing defenses fit

01

System Architecture & Methodology
A three-stage inference-time pipeline over a fixed classifier

02

Evaluation & Results
MNIST under FGSM / PGD: recovery, coverage & risk trade-offs

03

Conclusion & Future Work
Takeaways and directions ahead

04

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

PART 01

Introduction, Related Work & Contributions



Why adversarial robustness needs a lightweight, deployable answer

The Adversarial Threat

Deep neural networks deliver exceptional accuracy — yet they are **highly vulnerable to adversarial examples**: inputs altered by imperceptible perturbations that force confident misclassification.

FGSM

Fast Gradient Sign Method — a single-step ℓ_∞ perturbation along the loss gradient

PGD

Projected Gradient Descent — iterative, projected steps; a far stronger attack

Why it matters

- Real deployments are safety-critical and resource-constrained
- Attacks often succeed with high confidence — the model gives no warning
- Retraining or changing the architecture is frequently impractical (cost, regulation, operations)
- A defense that runs at inference time, with no model change, is highly desirable

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Related Work: Defense Families

Four broad families of defenses — each strong in some dimension, each with limits.

Adversarial training

Trains on adversarial examples. Effective but costly, attack-config dependent, and prone to robust overfitting.

Certified defenses

Provable robustness within bounded perturbations — but heavy compute and often loose bounds.

Preprocessing / filtering

Denoising or spatial filtering before inference. Lightweight & model-agnostic, yet brittle alone vs. strong attacks.

Detection & selective methods

Flag or reject adversarial inputs via statistical / confidence cues. Thresholds can be fragile to adaptive attacks.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Research Gap & Our Contributions

The gap: no training-free method unifies perturbation recovery, statistical detection, and selective abstention into one deployable pipeline for resource-constrained, safety-critical systems.

Our contributions

Lightweight post-processing defence framework
Modular external layer — no retraining or architecture change; plugs into existing pipelines

Median-filter perturbation recovery
Suppresses high-frequency adversarial artifacts while preserving the main image structure

Statistical Residual Detection in a Low-Dimensional Feature Space
Detect inputs that deviate from clean data behavior

Confidence + residual selective prediction
Explicit, tunable trade-off between coverage and risk

Comprehensive Experiment Evaluation
Thorough Evaluation of the MNIST dataset providing detailed analyses of the results

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

PART 02-1

System Architecture & Methodology

A black-box, three-stage decision layer over a fixed classifier



C4 Architecture Overview

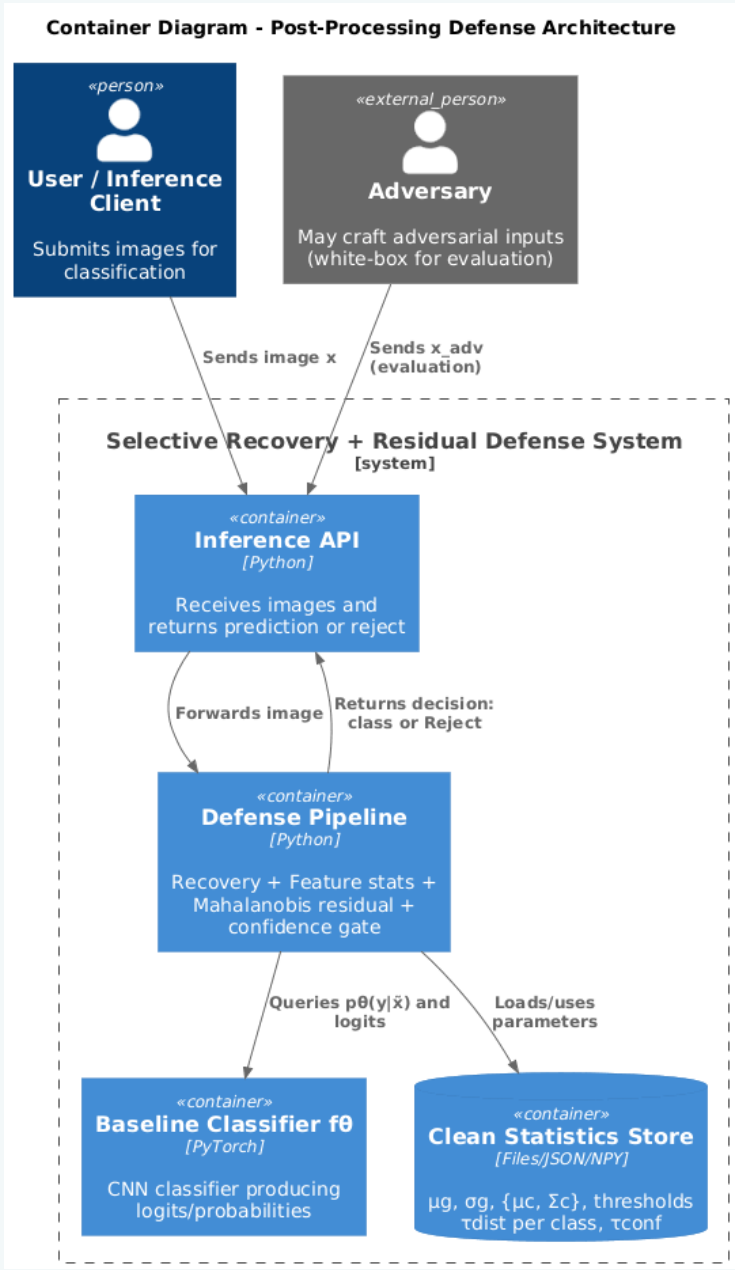
KEY IDEA
A lightweight defence layer is added around an existing pretrained classifier

ARCHITECTURE FLOW
Input samples are received by the inference API and passed to the Defence Pipeline

DEFENCE PIPELINE
The system applies recovery filtering, extracts statistical features, computes residual score and combines the with classifier confidence.

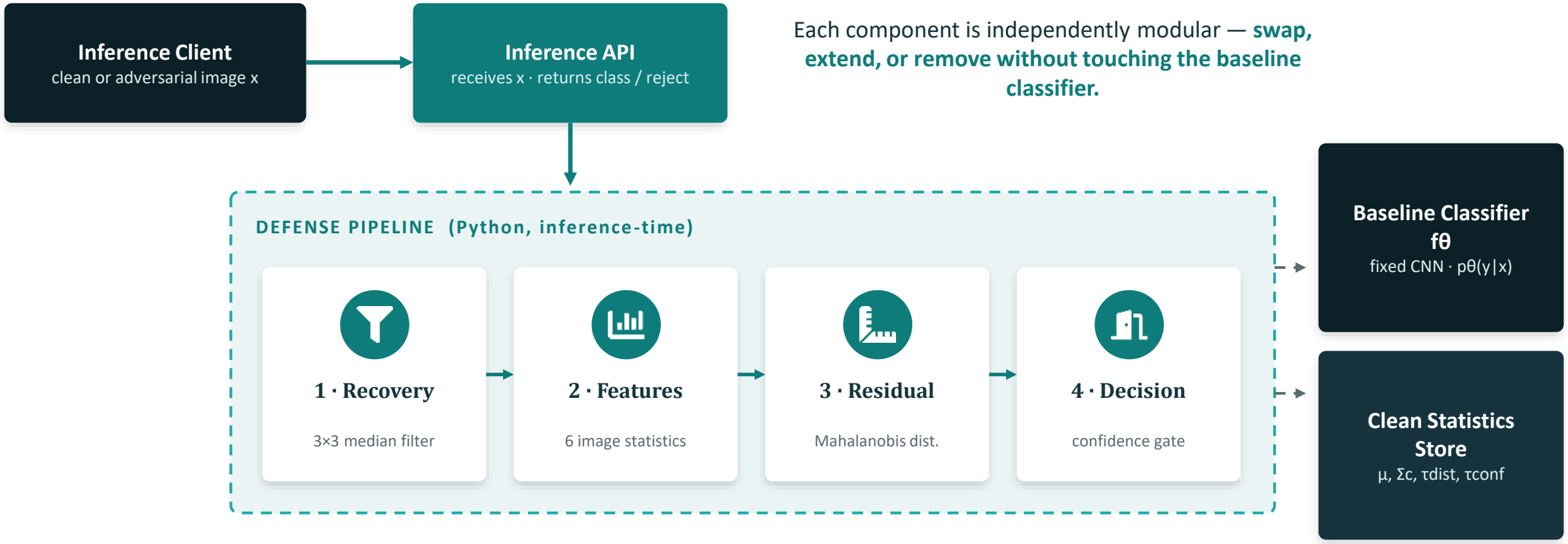
DECISION LOGIC
The system either returns the predicted class or rejects the sample

The classifier remains fixed, while robustness is improved through an interpretable and modular inference-time defence



The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

The Defense Pipeline



The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

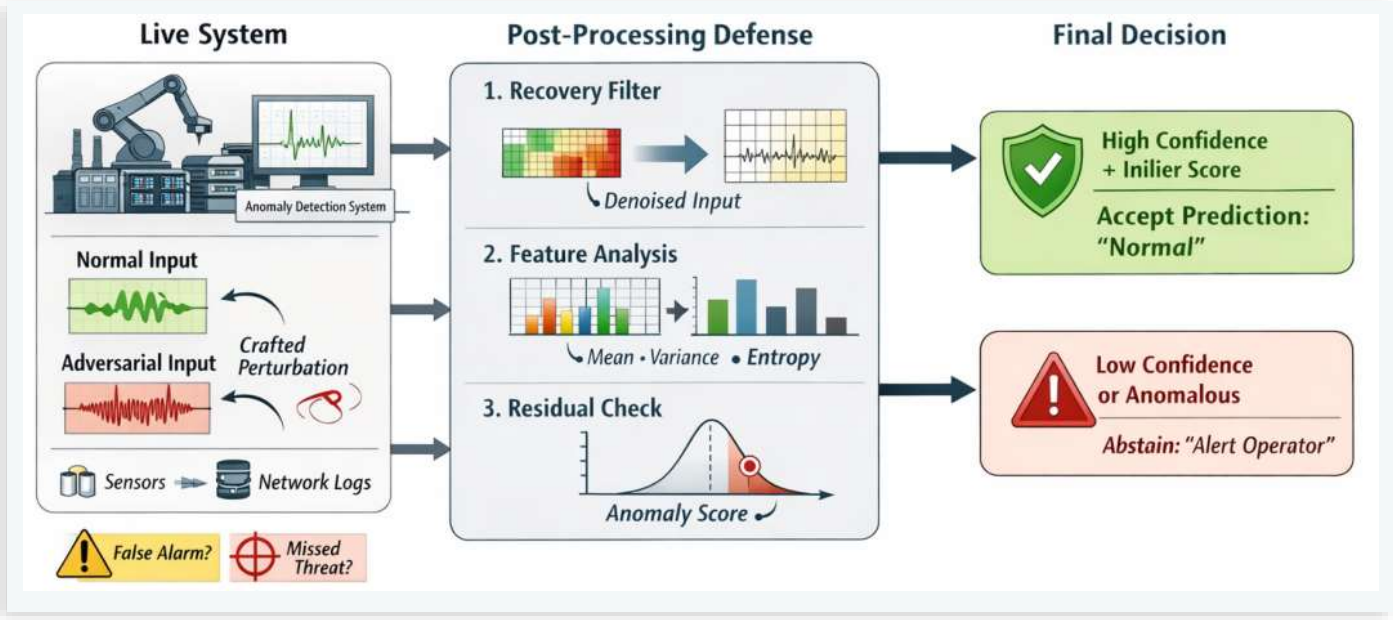
PART 02-2

Case Study Deployment



Case Study: Deployment

A pretrained anomaly-detection model in a live monitoring environment — the defense layer recovers, checks, and decides at inference time.



Setting. A pretrained ML model monitors a live environment — industrial sensing, network traffic, or control-system logs — trained only on clean operational data.

Constraint. Retraining or changing the architecture is impractical (cost, operations, regulation).

Solution. Add a post-processing layer: recover → check statistics → decide. If an input deviates or confidence drops, the system abstains and flags it for an operator.

Recover
Suppress non-natural high-frequency components

Check
Compare to clean reference statistics

Decide
Accept if consistent & confident, else abstain & alert

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

PART 03-3

Methodology: Recovery & Features



White-box gradient-based attacks

The adversary has full knowledge of the model parameters and loss. Two standard ℓ_∞-bounded attacks are evaluated:

FGSM

Fast Gradient Sign Method — single step

$$x_{adv} = \mathrm{clip}\left(x + \varepsilon \cdot \mathrm{sign}(\nabla_x L(f_\theta(x), y)) \right)$$

One gradient step — fast, weaker.

PGD

Projected Gradient Descent — iterative (20 steps)

$$x^{(t+1)} = \Pi_B\left(\mathrm{clip}\left(x^{(t)} + \alpha \cdot \mathrm{sign}(\nabla L) \right) \right)$$

Iterative projection onto the ε-ball — substantially stronger.

Perturbation budget: $\|x_{adv} - x\|_\infty \leq \varepsilon$, evaluated at $\varepsilon = 0.05, 0.10, 0.15, 0.20$

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Methodology: Recovery & Features

Perturbation Recovery

$\tilde{x} = R(x_adv)$

- A 2-D median filter with a 3×3 kernel
- Smallest window that kills isolated pixel spikes while preserving stroke geometry
- Larger kernels over-smooth digit edges and hurt clean accuracy

6-D Statistical Descriptor

$\phi(x) \in \mathbb{R}^6$ — compact, interpretable, cheap to compute:

Variance	**Std. deviation**
Mean	**Skewness**
Kurtosis	**Shannon entropy**

Standardized with global clean-data statistics for consistent scaling.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Methodology: From recovered input to accept/reject

Recovery $R(\cdot)$

$$\tilde{x} = \text{median}_{3\times3}(x_{adv})$$

Smallest window that kills isolated pixel spikes while keeping stroke edges. Larger kernels oversmooth.

Features $\varphi(x) \in \mathbb{R}^6$

$$\text{Var} \cdot \text{Std} \cdot \text{Mean} \cdot \text{Skew} \cdot \text{Kurt} \cdot \text{Entropy}$$

Low-dimensional, interpretable image statistics. Standardised with global clean-data μ, σ.

Residual d_min

$$d_c = \sqrt{((\varphi' - \mu_c)^T \, \Sigma_c^{-1} \, (\varphi' - \mu_c))}$$

Mahalanobis distance to each class; keep the minimum. Label-free, so robust to corrupted predictions.

Selective prediction rule

ACCEPT if $d_{min}(\tilde{x}) \le \tau_{dist}$ AND $conf(\tilde{x}) \ge \tau_{conf}$ — otherwise ABSTAIN

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

PART 04

Evaluation & Results



MNIST under FGSM / PGD — recovery, coverage and the risk trade-off

Experimental Setup

MNIST

28×28 grayscale digits · 60k train / 10k test

Classifier

CNN trained only on clean data; no adversarial samples

Attacks

FGSM & PGD-20 at ε = 0.05 / 0.10 / 0.15 / 0.20

Defense

τ_conf = 0.8 · Mahalanobis ratio cap r_max = 5.0

Four pipeline stages: clean → attack → recovery → selective defense.

CLEAN BASELINE

Strong reference before any attack

0.9890	**0.9890**
Accuracy	F1-Score
0.9888	**0.0012**
TPR	FPR

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Evaluation Metrics

Accuracy

$$(TP + TN) / (TP + TN + FP + FN)$$

Overall correct predictions

F1-score

$$2 \cdot (Precision \cdot Recall) / (Precision + Recall)$$

Macro-averaged across classes

True Positive Rate

$$TP / (TP + FN)$$

Also called Recall / Sensitivity

True Positive Rate

$$FP / (FP + TN)$$

Also called Recall / Sensitivity

Coverage

$$|D_{accepted}| / |D|$$

Fraction of samples accepted

Rejection Rate

$$1 - Coverage$$

Fraction of samples abstained

Accuracy Gain

$$\Delta Acc = Acc_{rec} - Acc_{atk}$$

Recovery benefit after attack

Selective Risk

$$Risk = 1 - Acc_{accepted}$$

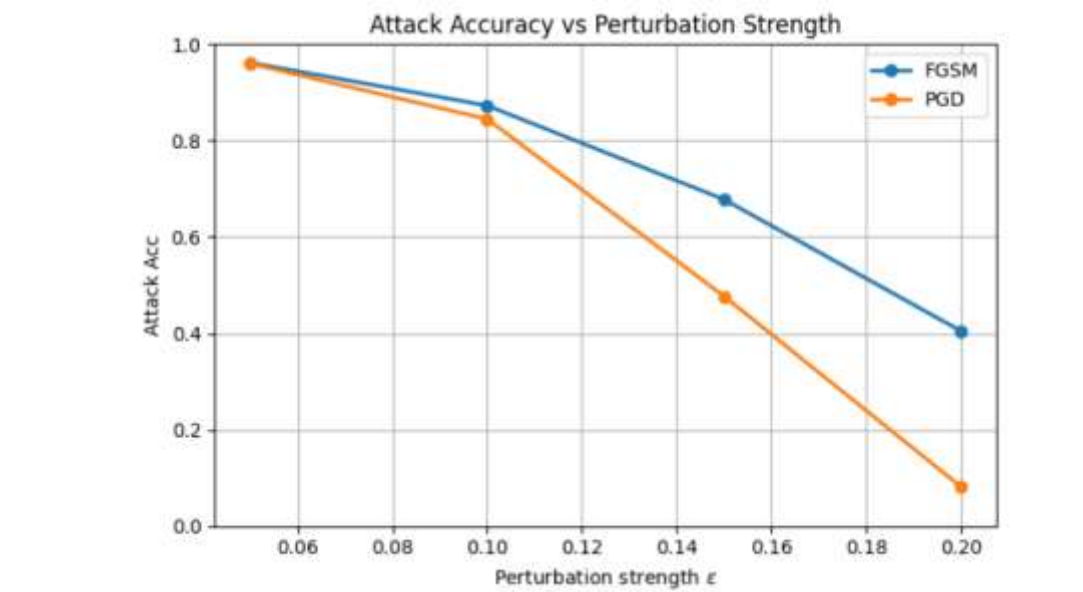
Error risk on accepted samples

Four pipeline stages: clean → attack → recovery → selective defense.

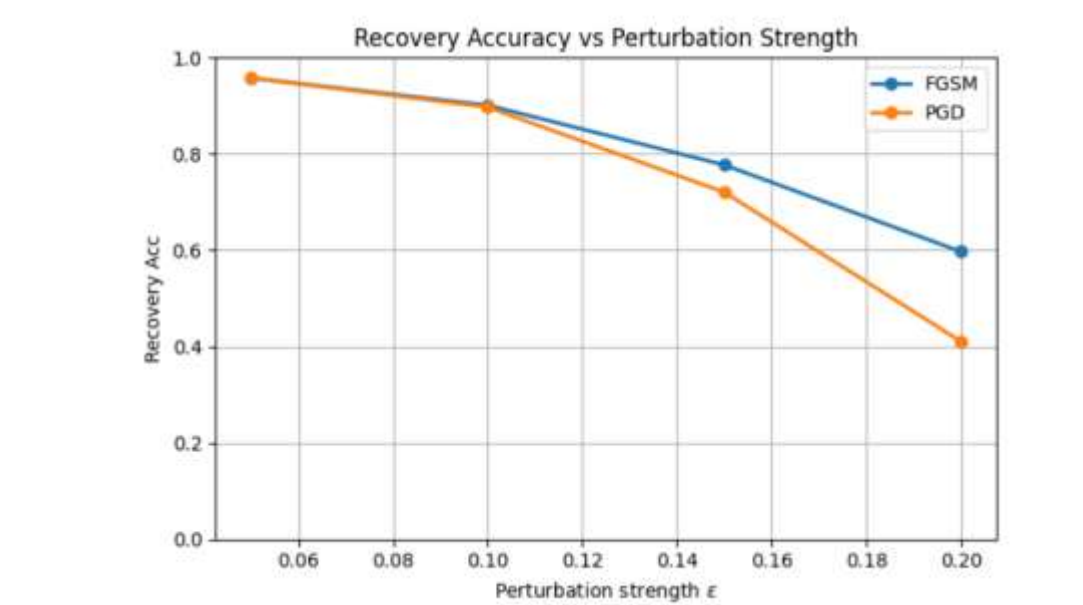
The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Attack Impact & Recovery

Accuracy degrades monotonically with ε — PGD far harder than FGSM. Median-filter recovery restores a large share across all ε.



(a) Classification accuracy under attack vs. ε



(b) Accuracy after median-filter recovery vs. ε

Recovery never retrains

Pure preprocessing lifts post-attack accuracy.

Median > Gaussian

Up to +23.8 pts vs. Gaussian smoothing at PGD ε=0.20.

Strong attacks still bite

At ε=0.20 PGD, recovery alone is not enough → detection.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Median filter vs. Gaussian smoothing

Attack	ε	Median Acc	Gaussian Acc	Δ (pts)
FGSM	0.05	0.9584	0.9649	−0.7
FGSM	0.10	0.8988	0.8845	+1.4
FGSM	0.15	0.7695	0.7084	+6.1
FGSM	0.20	0.5840	0.4652	+11.9
PGD	0.05	0.9583	0.9637	−0.5
PGD	0.10	0.8946	0.8687	+2.6
PGD	0.15	0.7046	0.5757	+12.9
PGD	0.20	0.3963	0.1583	+23.8

Why median wins

Comparable at ε=0.05
(attack too weak to matter).

Median strictly better for every ε ≥ 0.10.

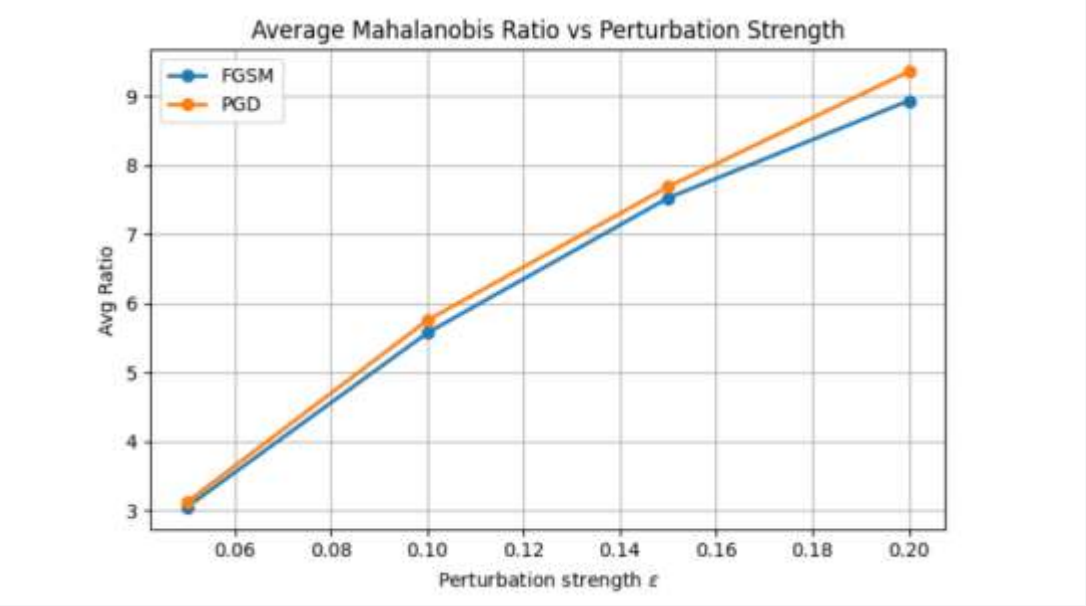
Advantage grows with attack strength — peaking at +23.8 pts under PGD ε=0.20.

It suppresses isolated adversarial spikes without blurring stroke edges; Gaussian degrades signal and noise alike.

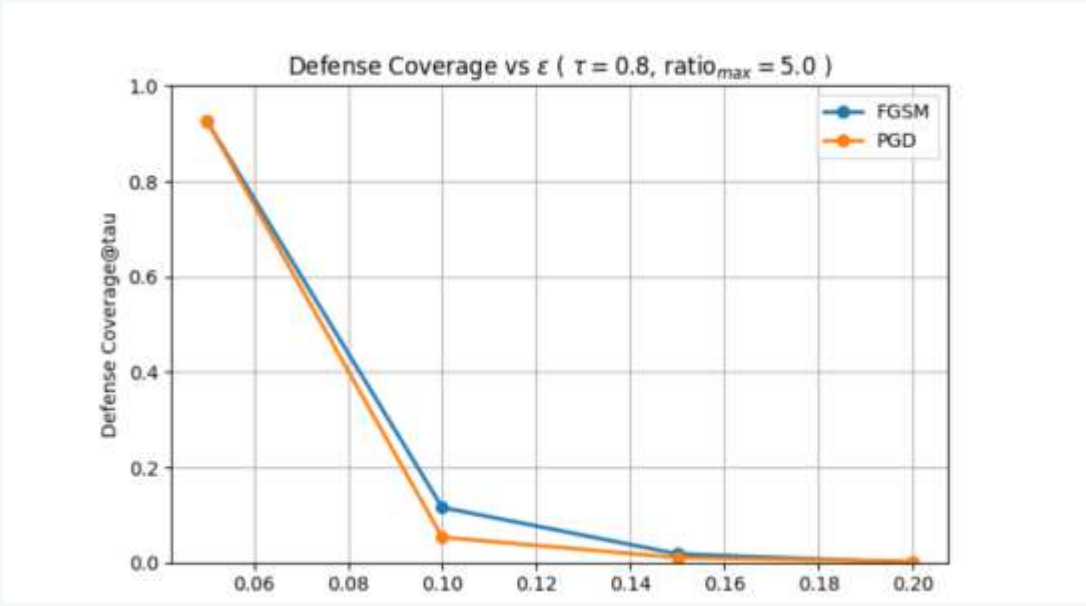
The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Residual Behavior & Coverage

The Mahalanobis residual grows with ε (statistical deviation from clean classes), so defense coverage falls sharply — the system stops forcing predictions.



(a) Average Mahalanobis ratio vs. ε



(b) Defense coverage vs. ε (τ = 0.8, r_max = 5.0)

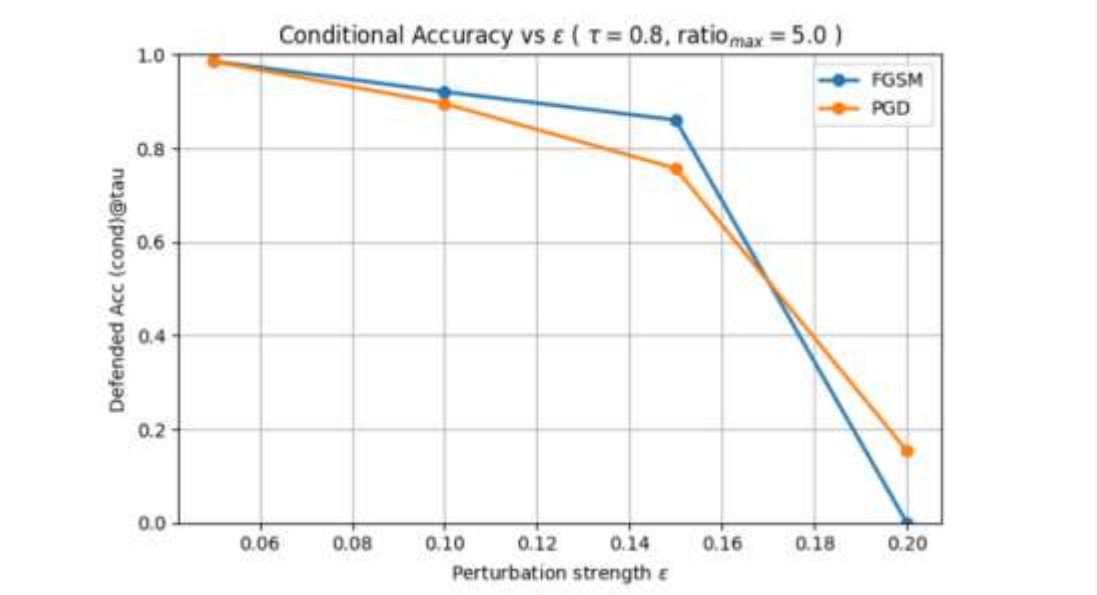
At ε = 0.05 the system accepts > 92% of samples; at ε = 0.20 nearly all are rejected — desirable in safety-critical settings.

Larger ε ⇒ bigger statistical deviation ⇒ more rejections. **Refusing to predict on heavily perturbed inputs is the desired safety-critical behaviour.**

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Selective Prediction Quality

Conditioned on accepted samples, accuracy stays high for low–moderate ε. At ε = 0.20 the drop pairs with near-zero coverage — abstention, not confident error.



Conditional accuracy under selective prediction vs. ε (τ = 0.8, r_max = 5.0)

Selective defense after recovery — PGD (Table III)

ε	Coverage	Cond. Acc	Sel. Risk
0.05	0.926	0.985	0.015
0.10	0.054	0.895	0.105
0.15	0.011	0.757	0.243
0.20	0.003	0.154	0.846

High coverage (>92%) with ~98.5% conditional accuracy at ε = 0.05; coverage collapses at ε = 0.20, preventing high-confidence errors.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Risk–Coverage & Rejection Control

Attack	ε	Coverage	Rejection	Cond. Acc	Selective Risk
FGSM	0.05	0.9260	0.0740	0.9857	0.0143
FGSM	0.10	0.1159	0.8841	0.9206	0.0794
FGSM	0.15	0.0186	0.9814	0.8602	0.1398
FGSM	0.20	0.0006	0.9994	0.0000	1.0000
PGD	0.05	0.9261	0.0739	0.9852	0.0148
PGD	0.10	0.0535	0.9465	0.8953	0.1047
PGD	0.15	0.0107	0.9893	0.7570	0.2430
PGD	0.20	0.0026	0.9974	0.1538	0.8462

Selective risk = 1 − conditional accuracy. Metrics computed over accepted samples (τ=0.8, r_max=5.0).

Low ε (0.05)

≈92% coverage with ≈98.5% conditional accuracy — near-clean operation.

High ε (0.20)

Coverage collapses: the system abstains rather than emit high-confidence mistakes.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Ablation & Computational Cost

Each component plays a distinct role — and the whole pipeline is extremely cheap.

Component roles (FGSM)

Recovery (R) — Drives the unconditional accuracy gain

Confidence gate (C) — Sustains high conditional accuracy at low–moderate ε (>0.98 at ε=0.05)

Mahalanobis (M) — Triggers conservative coverage collapse under heavy shift (ε ≥ 0.15)

R+C alone keeps 71.8% coverage at 86.0% accuracy; adding M restricts to 1.8% coverage at 83.6% (FGSM, ε=0.15).

Inference cost (10k MNIST, CPU)

Median filter recovery	26 µs
Feature + Mahalanobis	≈ 5 µs
Full defense pipeline	≈ 31 µs
PGD-20 attack (reference)	≈ 4.7 ms

~150× cheaper than generating a single PGD example — well within resource-constrained budgets.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

PART 04

Conclusion & Future Work



What we showed — and where this goes next

CONCLUSION & FUTURE WORK

A practical, interpretable defense layer

What we showed

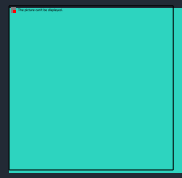
- **Training-free robustness** — Recovery + residual detection + confidence gating, no model changes.
- **Median beats Gaussian** — Strictly better recovery for every ε ≥ 0.10.
- **Reliable when it answers** — High coverage & near-clean accuracy at low ε; graceful abstention at high ε.
- **Cheap & interpretable** — ≈31 µs/sample on CPU — fit for constrained deployments.

Future work

1. **Scale up**
   Higher-resolution & more complex datasets beyond MNIST.
2. **Richer features**
   Learned or statistical representations for finer residual sensitivity.
3. **Adaptive thresholds**
   Dynamic τ and alternative distance metrics.

The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.

Thank you



Panagiotis Radoglou Grammatikis



pradoglou@k3y.bg

Funding The project funded under Grant Agreement No. 101168407 (CPAID) is supported by the European Cybersecurity Competence Centre. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Cybersecurity Competence Centre. Neither the European Union nor the granting authority can be held responsible for them.