15th IEEE MOCAST 2026
Universitat de les Illes Balears, Mallorca, Spain
June 10–12, 2026

Fingerprint-Based Selective Rejection

for Improving Reliability Under Adversarial Attacks

Dimitrios Christos Asimopoulos, Panagiotis Radoglou-Grammatikis, Vasileios Vitsas, Dimitrios Pliatsios, Georgios Th Papadopoulos, Sotirios Goudos, Evangelos K. Markakis, and Panagiotis Sarigiannidis

Funding The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

AUTHORS & ORGANIZATIONS

Authors & Contributors

MetaMind Innovations

Dimitrios-Christos Asimopoulos

International Hellenic University

Dimitrios-Christos Asimopoulos • Vasileios Vitsas

K3Y

Panagiotis Radoglou-Grammatikis

University of Western Macedonia

P. Radoglou-Grammatikis • D. Pliatsios • P. Sarigiannidis

Harokopio University of Athens

Georgios Th. Papadopoulos

Aristotle University of Thessaloniki

Sotirios Goudos

Hellenic Mediterranean University

Evangelos K. Markakis

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Agenda

1. **Introduction**
   Motivation, problem statement & research gap

2. **Related Work**
   Existing detection & selective-classification defenses

3. **Contributions**
   What this paper adds

4. **Methodology & Framework**
   Three-stage training-free reliability layer

5. **Fingerprint Features**
   Confidence geometry, TTA stability, signal statistics

6. **Selective Rejection Policy**
   Multi-view accept / reject rule

7. **Evaluation & Results**
   GTSRB · ResNet-18 · white- & black-box attacks

8. **Conclusions & Future Work**
   Key findings, implications, next steps

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

PART 01

Introduction, Related Work & Contributions

Why selective rejection — and what this paper adds

INTRODUCTION

Motivation: When Confident Models Fail

DNNs power safety-critical vision

Deep neural networks achieve impressive visual-recognition performance and are increasingly deployed in autonomous driving and medical imaging.

But they break under adversarial examples

Carefully crafted, often imperceptible perturbations cause confident misclassification — under strong white-box attacks accuracy can collapse to near-random.

Robust training is expensive & limited

Adversarial training and certified robustness help within a threat model, but are costly, need retraining, and stay vulnerable to stronger or adaptive attacks.

A complementary paradigm: abstain

Selective classification lets the system abstain on suspicious inputs. Since attacks are heterogeneous, a multi-view fingerprint gives a more reliable basis for rejection.

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Where Existing Defenses Stand

Robust Training

Adversarial training & certified robustness raise robustness within a fixed threat model — but require retraining and remain vulnerable to stronger or adaptive attacks.

Detection-Based Defenses

Detect adversarial inputs from a single signal such as softmax confidence or energy scores. Effective for some attacks, brittle across families.

Selective Classification

Allow the model to abstain on low-confidence inputs, trading coverage for reliability — but typically rely on one reliability signal.

The gap: prior work leans on a **single signal**, yet different attack families leave different traces — so single-signal rejection is unreliable, and no lightweight, training-free, multi-view policy exists under an explicit clean-coverage constraint.

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Problem Statement & Research Gap

Problem Statement

- Under strong attacks, DNN accuracy collapses to near-random levels, undermining reliability in safety-critical systems.
- Robust training is costly, needs retraining, and stays vulnerable to stronger or adaptive attacks.
- Uniformly high robustness in open-world settings is often unrealistic to achieve.

Research Gap

- Most selective and detection-based defenses rely on a single signal (e.g. softmax confidence or energy scores).
- Different attack families leave different statistical and geometric traces — single-signal rejection is unreliable.
- No lightweight, training-free, multi-view rejection policy exists under an explicit clean-coverage constraint.

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

CONTRIBUTIONS

Contributions

Multi-view fingerprint

1

A compact per-input fingerprint combining confidence geometry, test-time-augmentation stability, and raw-signal statistics.

Training-free rejection layer

2

A post-hoc selective-rejection layer that wraps a fixed classifier — no retraining, no test-set tuning.

Clean-coverage constraint

3

A global rejection policy whose thresholds are chosen on validation under an explicit clean-coverage target, then frozen.

Evaluation across attack families

4

Systematic study on GTSRB / ResNet-18 against white- and black-box attacks, with feature-importance and PCA analysis.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

PART 02

Methodology & Framework

A lightweight, training-free, three-stage reliability layer

Framework Overview

Post-hoc reliability layer

- Lightweight, training-free
- Wraps a fixed classifier (ResNet-18) with no retraining
- Accepts reliable inputs, abstains on suspicious ones
- Rejected samples deferred to a human or secondary check

1

STAGE 1

Fingerprint Extraction

Compute compact fingerprint $\varphi(x)$: confidence geometry, TTA stability, signal statistics.

2

STAGE 2

Fingerprint Classifier

Lightweight logistic regression maps $\varphi(x)$ to a clean-class reliability score $p_{clean}(x)$.

3

STAGE 3

Global Rejection Policy

Accept or abstain using p_{clean}, margin and TTA disagreement against frozen thresholds.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Fingerprint Extraction

Extracts a compact fingerprint vector $\phi(x)$ from each input — three complementary views, all computed from a fixed model with no retraining.

Confidence geometry

Softmax entropy $H(x)$ and the top-1 / top-2 margin $m(x)$ — how peaked or flat the prediction is.

Prediction stability (TTA)

Re-predict under N label-preserving augmentations; measure disagreement and entropy variance.

Signal statistics

Total variation of the image and the high/low-frequency energy ratio from a 2D DCT.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

METHODOLOGY — STAGE 2

Fingerprint Classifier

g(ϕ(x)) — logistic regression

- A lightweight multinomial logistic regression on the fingerprint vector.
- Trained to separate clean inputs from attack types using the three fingerprint views.
- Only the clean-class probability p_clean(x) is used as a reliability score.
- Trained only on the training split — never on test data.

Confidence geometry features

TTA stability features

Signal statistics features

p_clean(x)

reliability score

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

METHODOLOGY — STAGE 3

Global Rejection Policy

REJECT IF ANY CRITERION FIRES

$$p\_clean(x) < \tau_{fp} \quad \vee \quad m(x) < \tau_{margin} \quad \vee \quad d(x) > \tau_{disagree}$$

p_clean(x)

Clean-class probability from the fingerprint classifier.

m(x)

Top-1 minus top-2 softmax margin — confidence separation.

d(x)

TTA disagreement rate over N augmentations.

Thresholds chosen on validation under a clean-coverage constraint

— then frozen and evaluated once on the test set. Otherwise the base prediction is accepted; rejected samples are deferred to a human or secondary check.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

PART 03

Fingerprint Features

The three views that flag unreliable predictions

Confidence Geometry & TTA Stability

Confidence geometry

Describes how peaked or flat the softmax distribution is. Entropy measures uncertainty; the margin captures confidence separation.

$$H(x) = -\sum p_c \log p_c$$

TTA stability

Apply N label-preserving augmentations (small rotations, jitter, padding) and re-predict. Unstable, likely-adversarial inputs disagree.

$$m(x) = p_{(1)} - p_{(2)}$$

FEATURE EXTRACTION FLOW

1. Input image x
2. Softmax output → entropy H(x), margin m(x)
3. N augmentations → disagreement d(x), entropy variance
4. Concatenate into the fingerprint vector φ(x)

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Signal Statistics & Threshold Selection

Signal statistics

Cheap, model-free descriptors of the raw signal that capture pixel-level texture and frequency artifacts perturbations often introduce.

- **Total variation (TV) of the image**
- **High / low-frequency energy ratio (2D DCT)**

No retraining — computed directly from each input.

Threshold selection

Thresholds (τ_fp, τ_margin, τ_disagree) chosen by grid search on the validation split only.

Given a target clean coverage κ, keep configs with validation clean coverage ≥ κ − ε.

Among feasible candidates, pick the one maximizing selective accuracy on a mixed validation pool — then freeze and evaluate once on test.

REJECT RULE

```
Reject(x) = [ p_clean<τ_fp ] ∨ [ m(x)<τ_margin ] ∨ [
d(x)>τ_disagree ]
```


The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

PART 04

Selective Rejection Policy

Accept reliable inputs — abstain on the rest

Accept Reliable Inputs, Abstain on the Rest

Combining clean-class probability, confidence margin, and TTA disagreement into a single accept / reject decision.

STEP 1

Input x

A single image to classify

STEP 2

Fingerprint

Confidence · TTA stability · signal stats

STEP 3

Reliability score

p_clean(x), margin m(x), disagreement d(x)

STEP 4

Accept / Reject

Apply the global rejection rule

Accept — the base classifier prediction is returned.

Reject — abstain; defer to a human or secondary check.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Rejection Policy: Rule & Components

$$\text{Reject}(x) = [\, p\_clean(x) < \tau\_fp \,] \lor [\, m(x) < \tau\_margin \,] \lor [\, d(x) > \tau\_disagree \,]$$

HOW REJECTION WORKS

1. **Compute fingerprint** — Extract confidence, TTA stability and signal features for the input.
2. **Score reliability** — Classifier outputs p_clean(x); margin and disagreement are read off directly.
3. **Apply rule** — If any criterion fires, abstain; otherwise accept the base prediction.
4. **Defer** — Rejected samples are sent to a human or secondary verification process.

RULE COMPONENTS

`p_clean` — clean-class probability from the fingerprint classifier

`m(x)` — top-1 minus top-2 softmax margin

`d(x)` — TTA disagreement rate over N augmentations

`τ` — thresholds frozen from validation

Training-free • multi-view • tunable clean-coverage operating point

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

PART 05

Evaluation & Results

GTSRB · ResNet-18 · white- and black-box attacks

Experimental Setup

Dataset — GTSRB

- 43 traffic-sign classes
- Standard train / test split
- RGB traffic-sign images
- Safety-critical recognition domain

Model — ResNet-18

- Trained with data augmentation
- Cross-entropy loss
- 95.5% clean test accuracy
- Kept fixed — no retraining

Attacks (ART)

- White-box: FGSM, MIM, PGD, CW
- Black-box: Square, ZOO
- ℓ_∞ budget $\varepsilon = 0.03$
- Evaluated on 1000 test images

Protocol

- PyTorch + ART attack library
- Fingerprints split 60 / 20 / 20
- Features standardized on train stats
- Thresholds frozen, scored once on test

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Evaluation Metrics

Coverage

Fraction of inputs the system accepts (does not reject).

Selective Accuracy

Accuracy computed only on the accepted samples.

Macro-F1

Macro-F1 of the fingerprint classifier across classes.

Baseline Accuracy

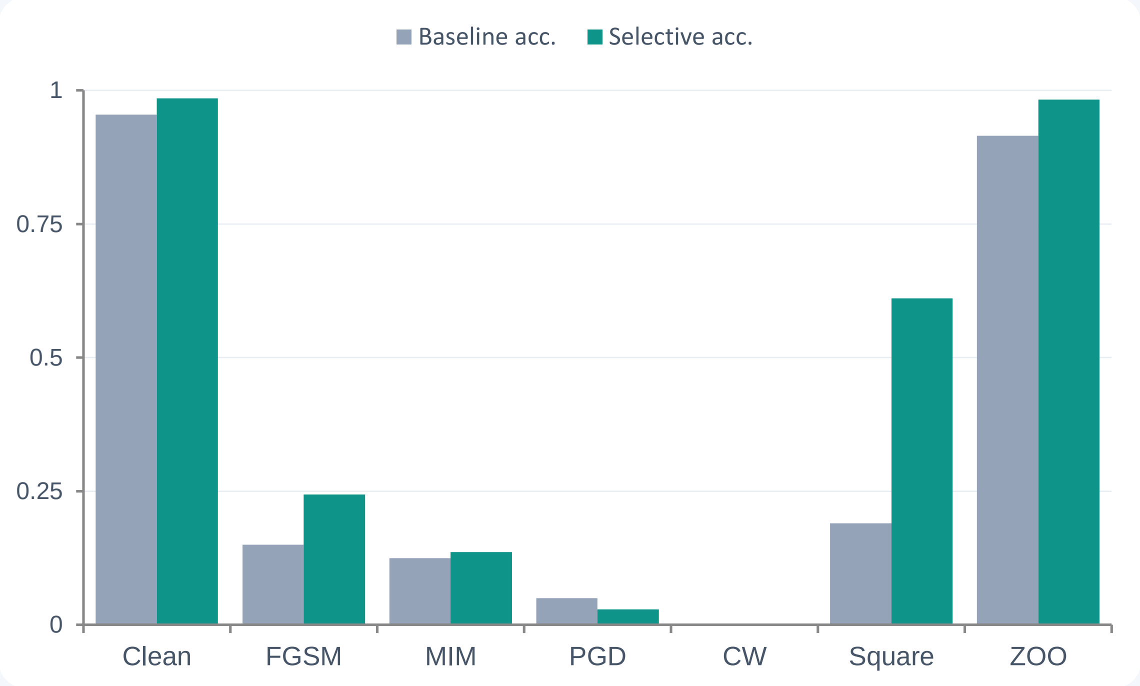
Accuracy of the undefended classifier, with no rejection.

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Per-Attack Selective Rejection

Attack	Baseline	Coverage	Selective
Clean	0.955	0.670	**0.985**
FGSM	0.150	0.225	**0.244**
MIM	0.125	0.295	**0.136**
PGD	0.050	0.345	**0.029**
CW	0.000	0.030	**0.000**
Square	0.190	0.180	**0.611**
ZOO	0.915	0.605	**0.983**

GTSRB per-attack performance — target clean coverage ≈ 0.7



On accepted samples, accuracy lifts far above baseline: CW almost fully filtered (0.000), Square jumps 0.190→0.611, clean reaches 0.985. Iterative ℓ∞ attacks (PGD, MIM) stay hard.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

RESULTS

Coverage–Accuracy Trade-off

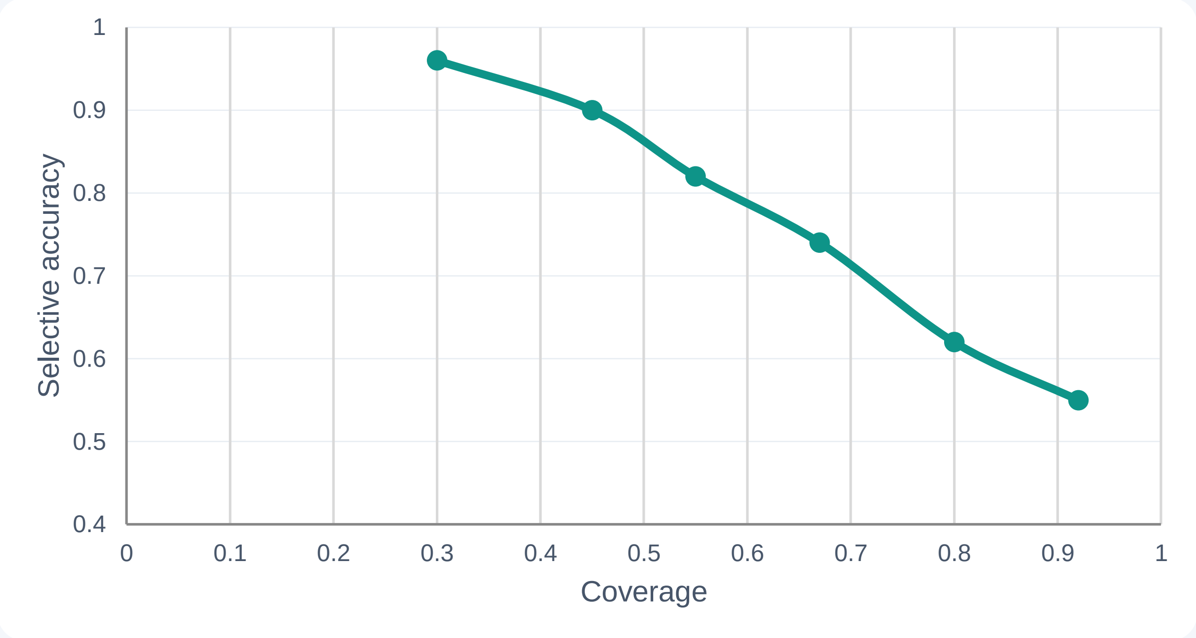
≈ 0.67
Clean test coverage

62.6%
Selective accuracy (accepted)

0.985
Clean selective accuracy

Mixed-pool coverage ≈ 0.34

A single global policy is selected on validation under a clean-coverage target. Stricter rejection (lower coverage) yields higher selective accuracy — an inherent reliability–coverage trade-off.



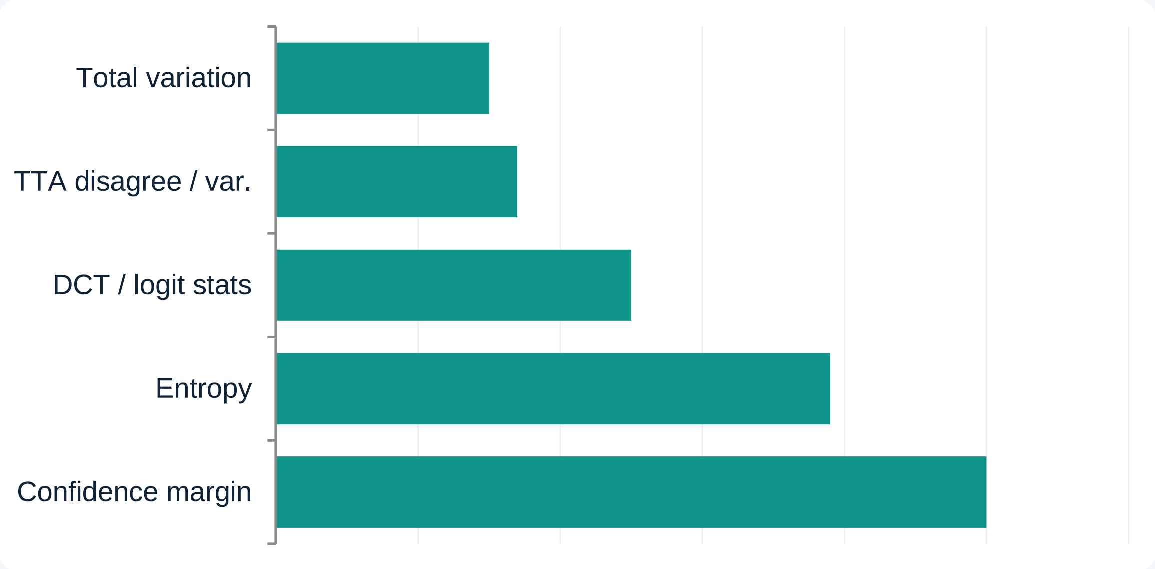
Chosen operating point: clean coverage ≈ 0.67

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

RESULTS

Fingerprint Space & Feature Importance

FEATURE IMPORTANCE



Confidence margin and entropy dominate; signal and stability features contribute moderately.

Classification performance

≈ 0.42 Fingerprint-classifier macro-F1 — moderate but imperfect separability

- A 2D PCA projection shows partial separation for several attacks.
- Strong iterative ℓ_∞ attacks overlap heavily with the clean distribution.
- This explains why CW is filtered well while PGD and MIM remain challenging.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

RESULTS

Selective Rejection — Performance Analysis

KEY FINDINGS

Preserves accuracy on clean inputs
0.985 selective accuracy on accepted clean samples, keeping clean coverage near 0.67.

Filters strong optimization attacks
CW is almost completely filtered out — selective accuracy near 0.000.

Sharp gains on black-box Square
Square selective accuracy jumps from 0.190 to 0.611.

COMPARATIVE ATTACK PERFORMANCE

Optimization (CW)
Almost fully filtered by rejection

Black-box (Square)
Large gain in selective accuracy

Iterative ℓ_∞ (PGD / MIM)
Remain challenging — overlap with clean

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

RESULTS

Qualitative Findings in Fingerprint Space

Optimization attacks

CW

In PCA: Almost fully rejected

PGD-step

In PCA: Largely rejected

Iterative ℓ_∞ attacks

PGD / MIM

In PCA: Heavy clean overlap

BIM

In PCA: Hard to separate

Black-box attacks

Square

In PCA: Large selective gain

ZOO

In PCA: Mostly accepted

Key insight: Strong optimization attacks (CW) sit far from the clean cluster and are almost fully rejected, while iterative ℓ_∞ attacks overlap with the clean distribution — limiting separability. Confidence margin and entropy are the most informative features (macro-F1 ≈ 0.42).

The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

PART 06

Conclusions & Future Work

What we showed — and where it goes next

CONCLUSIONS

Key Achievements & Future Work

KEY ACHIEVEMENTS

- Lightweight, training-free selective-rejection layer wrapping a fixed classifier.
- Almost fully filters strong CW attacks and gives large reliability gains on Square.
- Preserves very high clean accuracy (0.985) on GTSRB / ResNet-18.
- No retraining and no test-set tuning — thresholds frozen on validation.

FUTURE WORK

- **Per-attack thresholds** — Move beyond a single global operating point.
- **Richer fingerprints** — Add features that separate iterative ℓ_∞ attacks.
- **Adaptive attacks** — Study robustness against fingerprint-aware adversaries.
- **Cross-dataset validation** — Extend evaluation to more datasets and models.

The authors acknowledge the financial support of the NERO project (“advaNced cybErsecurity awaReness ecOsystem for SMEs”), funded by the European Union’s DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.

Thank you for your attention!

Questions & discussion welcome

Fingerprint-Based Selective Rejection for Improving Reliability Under Adversarial Attacks

Funding The authors acknowledge the financial support of the NERO project ("advaNced cybErsecurity awaReness ecOsystem for SMEs"), funded by the European Union's DEP programme under Grant Agreement No. 101127411. The views expressed are solely those of the authors. The authors declare no conflicts of interest.