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‡

***Abstract*—Adversarial examples can severely degrade the reliability of deep neural networks, often driving accuracy close to random guessing under strong attacks. In this work, we propose a lightweight, training-free defense based on *input fingerprinting* and *selective rejection*. Instead of attempting to robustly classify every input, the method estimates a compact set of fingerprint features capturing prediction confidence, stability under test-time augmentation, and simple signal statistics, and rejects samples deemed unreliable. Experiments on GTSRB with a ResNet-18 classifier under multiple white-box and black-box attacks show that the proposed policy substantially improves accuracy on accepted samples, preserves high reliability on clean data, and almost completely filters out strong optimization-based attacks, at the cost of reduced coverage. The results highlight the practicality of fingerprint-based selective defenses as a complementary reliability layer for safety-critical vision systems.**

***Index Terms*—Adversarial examples, selective classification, abstention, robustness, fingerprinting.**

‡‡ This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101127411 (NERO: advaNced cybErsecurity awaReness ecOsystem for SMEs). Disclaimer: Funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

* Dimitrios-Christos Asimopoulos is with MetaMind Innovations, Kila, 50100 Kozani, Greece – E-Mail: dasimopoulos@metamind.gr

† Dimitrios-Christos Asimopoulos is also with the Department of Information and Electronic Engineering, International Hellenic University (I.H.U.), Alexander Campus, P.O. 141, P.C. 57400, Sindos, Thessaloniki, Greece – E-Mail: dimiasim3@ihu.gr

‡ P. Radoglou-Grammatikis is with K3Y, Studentski District, Vitosha Quarter, Bl. 9, 1700 Sofia, Bulgaria – E-Mail: pradoglou@k3y.gr

† V. Vitsas is with the Department of Information and Electronic Engineering, International Hellenic University (I.H.U.), Alexander Campus, P.O. 141, P.C. 57400, Sindos, Thessaloniki, Greece – E-Mail: vitsas@it.teithe.gr

§ P. Radoglou-Grammatikis, D. Pliatsios and P. Sarigiannidis are with the Department of Electrical and Computer Engineering, University of Western Macedonia, Campus ZEP Kozani, 50100 Kozani, Greece – E-Mail: {pradoglou, dpliatsios, psarigiannidis}@uowm.gr

¶ G. Th. Papadopoulos is with the Department of Informatics and Telematics, Harokopio University of Athens, Greece, Omirou 9, Tavros, GR17778, Athens, Greece – E-Mail: g.th.papadopoulos@hua.gr

‖ S. Goudos is with the School of Physics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece – E-Mail: sgoudo@physics.auth.gr

** T. Lagkas is with the Department of Informatics, Faculty of Science, Democritus University of Thrace, 14, Kavala Campus, Greece – E-Mail: tlagkas@cs.duth.gr

†† Evangelos K. Markakis is with the Department of Electrical Computer Engineering, Hellenic Mediterranean University, 20 Lysimaxou Kalokairinou, 71202 Heraklion, Greece – E-Mail: emarkakis@hmu.gr

I. Introduction

Deep neural networks have achieved impressive performance in visual recognition and are increasingly deployed in safety-critical applications such as autonomous driving and medical imaging. However, they are vulnerable to *adversarial examples*: carefully crafted, often imperceptible perturbations that can cause confident misclassification. Under strong white-box attacks, the accuracy of modern networks can collapse to near-random levels, raising serious concerns about their reliability.

A large body of work has focused on *robust training* methods such as adversarial training, regularization-based defenses, and certified robustness. While these approaches can improve robustness within specific threat models, they are computationally expensive, require retraining, and often remain vulnerable to stronger or adaptive attacks. In practice, achieving uniformly high robustness in open-world or safety-critical scenarios may be unrealistic.

This motivates an alternative and complementary paradigm: *selective classification* (or *rejection*). Instead of forcing a prediction for every input, the system may abstain on unreliable or suspicious samples, which can then be deferred to a human or a secondary verification process. Most existing selective or detection-based defenses rely on a single signal (e.g., softmax confidence or energy scores). However, adversarial examples are heterogeneous, and different attack families leave different statistical and geometric traces on the model and the input. This suggests that a *multi-view* characterization can provide a more reliable basis for rejection.

The main contributions of this work are summarized as follows:

- We propose a lightweight, training-free *fingerprint-based selective rejection* framework that operates as a post-hoc reliability layer for image classifiers.
- We design a compact set of fingerprint features capturing output confidence geometry, prediction stability under test-time augmentations, and simple signal statistics.
- We introduce a validation-driven global rejection policy with an explicit clean-coverage constraint, avoiding any test-set tuning.
- We analyze the fingerprint space using Principal Component Analysis (PCA and feature importance to explain the strengths and limitations of the approach.

The remainder of this paper is organized as follows. Section II reviews related work on adversarial robustness, detection,

and selective classification. Section III describes the proposed fingerprint representation and the selective rejection policy. Section IV details the experimental setup, datasets, and evaluation protocol. Section V presents and analyzes the experimental results. Finally, Section VI discusses limitations and concludes the paper.

II. Related Work

Adversarial robustness and defense mechanisms have been studied extensively in recent years, particularly as deep neural networks are increasingly deployed in vision and security-critical domains. A number of recent surveys provide comprehensive overviews of adversarial attack and defense strategies, emphasizing the trade-offs between robustness, model performance, and computational cost [1], [2]. These works highlight the broad class of threat models (white-box vs. black-box), evaluation protocols, and the ongoing challenge of developing practical defenses that generalize across diverse perturbations.

Beyond classical white-box and black-box formulations, recent work has explored *intermediate* threat models that bridge the two. In particular, Asimopoulos *et al.* propose *surrogate-guided adversarial attacks*, where a surrogate model is used to enable strong white-box optimization techniques in black-box scenarios, significantly improving attack transferability and effectiveness [3]. In complementary work, the same authors introduce AAG, a unified *Adversarial Attack Generator* framework for systematically evaluating model robustness across multiple attack families and configurations [4]. These works highlight the increasing sophistication and practicality of modern attack pipelines, further reinforcing the need for defenses that do not rely on a narrow or fixed threat model.

Sachpelidis-Brozos et al. build a MITRE ATLAS-aligned taxonomy of adversarial AI threats, mapping six attack families and linking 24 mitigation's to techniques [5]. Complementarily, Asimopoulos et al. empirically evaluate IEC 60870-5-104 IDS models, showing FGSM perturbations and CTGAN-generated samples sharply degrade detection accuracy, motivating ATLAS-guided hardening for critical infrastructure in practice [6].

A prominent line of work focuses on *robust training* techniques such as adversarial training and robust optimization, which aim to harden models by exposing them to adversarial examples during training [7]. While effective for certain attack families, these approaches often incur significant computational overhead and can suffer from issues such as catastrophic overfitting or limited generalization to unseen attacks.

Another line of research centers on *adversarial detection* and selective prediction. Detection methods attempt to distinguish adversarial inputs from clean ones at inference time, often based on statistical or feature-space irregularities [8]. Recent work on neural fingerprints for adversarial attack detection introduces randomization and class-specific internal activation patterns as a basis for identifying manipulated inputs [9]. Such approaches are conceptually related to our use of composite fingerprint features, though they vary in design and computational profile.

Selective classification, where the model is allowed to abstain or reject uncertain inputs, has been proposed as a practical compromise between robust accuracy and coverage, and is increasingly recognized as a complementary reliability mechanism [10]. In this regime, a system refrains from making potentially erroneous predictions rather than striving for uniform robustness against all perturbations.

In contrast to robust training or standalone detection techniques, our work leverages a multi-view fingerprint representation that combines confidence geometry, test-time augmentation stability, and simple signal statistics and integrates them into a validation-driven rejection policy under explicit clean coverage constraints. This positions our approach as both an effective and computationally lightweight selective defense in the broader adversarial robustness landscape.

III. Methodology

We consider a standard image classifier $f_\theta : \mathcal{X} \to \Delta^C$ that maps an input image x to a probability distribution over C classes. Our goal is not to modify or retrain f_θ, but to equip it with a *post-hoc selective rejection mechanism* that can abstain on inputs that are deemed unreliable or potentially adversarial.

To this end, we associate each input x with a compact *fingerprint vector* $\phi(x) \in \mathbb{R}^d$ that captures complementary aspects of the model's behavior and the input signal. A lightweight classifier $g(\cdot)$ is trained on these fingerprints to estimate whether a sample is likely to be clean or adversarial. At test time, the output of $g(\cdot)$ is combined with confidence- and stability-based criteria to decide whether to *accept* or *reject* the prediction of f_θ.

A. Fingerprint Feature Extraction

Given an input x, we extract fingerprint features from three complementary perspectives: output confidence geometry, prediction stability under test-time augmentation (TTA), and signal statistics.

a) Output confidence geometry.: Let $p(x) = f_\theta(x)$ denote the softmax output, with $p_{(1)}(x)$ and $p_{(2)}(x)$ the highest and second-highest probabilities. We compute entropy $H(x) = -\sum_{c=1}^{C} p_c(x) \log p_c(x)$ to quantify uncertainty and the margin $m(x) = p_{(1)}(x) - p_{(2)}(x)$ to capture confidence separation.

b) Prediction stability under test-time augmentation (TTA).: We generate N label-preserving augmentations $\{x^{(i)}\}_{i=1}^{N}$ (small rotations, jitter, padding). Let $\hat{y}(x)$ denote the original prediction and $\hat{y}(x^{(i)})$ the augmented predictions. The disagreement rate is defined as

$$d(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\hat{y}(x^{(i)}) \neq \hat{y}(x)\right],$$

and we also compute the variance of prediction entropy across augmented samples.

c) Signal statistics.: We include image-level descriptors such as total variation (TV) and the high-/low-frequency energy ratio derived from a 2D Discrete Cosine Transform (DCT).

All features are concatenated into a standardized fingerprint vector $\phi(x)$.

B. Fingerprint Classifier

A lightweight multinational logistic regression model

$$g(\phi(x)) \in [0,1]^K$$

is trained to distinguish between clean samples and different attack types. In practice, only the probability of the clean class, $p_{\text{clean}}(x)$, is used as a reliability indicator. The fingerprint dataset is split into training, validation, and test sets, and the classifier is trained exclusively on the training split.

C. Global Rejection Policy

At inference, we compute $p_{\text{clean}}(x)$, the margin $m(x)$, and the TTA disagreement rate $d(x)$. The system rejects an input if

$$\text{Reject}(x) = \big(p_{\text{clean}}(x) < \tau_{\text{fp}}\big) \vee \big(m(x) < \tau_{\text{margin}}\big) \vee \big(d(x) > \tau_{\text{disagree}}\big).$$

If rejection is false, the base prediction is accepted; otherwise, the system abstains.

D. Threshold Selection Under Clean-Coverage Constraint

Thresholds are selected using only the validation split. Given a target clean coverage κ, we perform a grid search over candidate thresholds and retain configurations satisfying

$$\text{Coverage}^{\text{val}}_{\text{clean}} \geq \kappa - \epsilon.$$

Among feasible candidates, we select the one maximizing selective accuracy on a mixed validation pool. The chosen thresholds are then frozen and evaluated once on the test set.

IV. Experimental Setup

A. Dataset and Model

Experiments are conducted on GTSRB (43 traffic sign classes). We use the standard train/test split and train a ResNet-18 with data augmentation and cross-entropy loss, achieving 95.5% clean test accuracy.

For fingerprint learning, clean and adversarial samples are collected and split into training, validation, and test sets (60/20/20). Features are standardized using training-set statistics.

B. Adversarial Attacks

We evaluate against white-box (FGSM, MIM, PGD, CW) and black-box (Square, ZOO) attacks implemented in ART. For ℓ_∞ attacks, the perturbation budget is $\epsilon = 0.03$. Unless stated otherwise, attacks are evaluated on 1000 test images. Detailed hyperparameters are provided in the supplementary material.

Table I: Per-attack performance of the global selective rejection policy on the test set (target clean coverage ≈ 0.7).

Attack	Baseline Acc.	Coverage	Selective Acc.
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

V. Evaluation Results

We evaluate the proposed fingerprint-based selective rejection mechanism on the GTSRB dataset using a ResNet-18 classifier under a diverse set of white-box and black-box attacks. All thresholds are selected exclusively on the validation split under a clean-coverage constraint, and final results are reported on a held-out test set.

A. Baseline Vulnerability

Table I reports the baseline accuracy of the classifier without any defense. While performance on clean data is high (95.5%), several attacks drastically degrade accuracy. In particular, CW reduces accuracy to 0%, PGD to 5%, and MIM to 12.5%, confirming that the model is highly vulnerable under strong white-box attacks. FGSM and Square also lead to severe degradation, while the specific ZOO configuration used here remains relatively weak.

B. Global Selective Rejection Policy

We next consider a single global rejection policy selected on the validation set under a target clean-coverage constraint. Due to the discrete threshold grid, the best feasible operating point achieves a *clean test coverage* of approximately 0.67. At this operating point, the *overall coverage* on the mixed test pool is about 0.34, indicating a conservative but safety-oriented behavior.

On the accepted samples, the system achieves a *selective accuracy* of 62.6%. The full per-attack results are reported in Table I.

C. Coverage–Accuracy Trade-off

Figure 1 shows the trade-off between selective accuracy and coverage on the validation set for different threshold combinations. As expected, stricter rejection policies (lower coverage) generally lead to higher selective accuracy. The selected operating point is highlighted.

D. Analysis of the Fingerprint Space

Figure 2 shows a two-dimensional PCA projection of the standardized fingerprint vectors for clean and adversarial samples. While partial separation is visible for some attack types, strong iterative attacks exhibit substantial overlap with the clean distribution. Figure 3 reports the feature importance of the fingerprint classifier in terms of the mean absolute logistic regression weights.

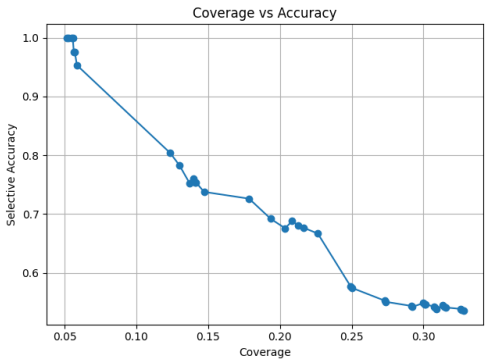

Figure 1: Selective accuracy versus coverage on the validation set for different threshold combinations. The highlighted point corresponds to the operating point selected under the clean-coverage constraint.

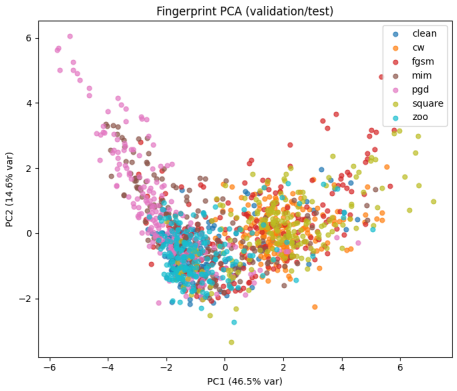

Figure 2: Two-dimensional PCA visualization of standardized fingerprint vectors for clean and adversarial samples. Overlap between clean and strong iterative attacks explains the difficulty of selective rejection in these cases.

The fingerprint classifier achieves a macro-F1 score of approximately 0.42, indicating moderate but imperfect separability among classes. In summary, the proposed fingerprint-based selective rejection mechanism (i) preserves very high accuracy on accepted clean samples, (ii) almost completely filters out strong optimization-based attacks such as CW, (iii) substantially improves reliability under certain black-box attacks (notably Square), but (iv) remains less effective against iterative ℓ_∞ attacks due to overlapping fingerprint distributions.

VI. Conclusion

In this work, we proposed a lightweight fingerprint-based selective rejection mechanism that acts as a post-hoc reliability layer for image classifiers without retraining. The method combines features capturing confidence geometry, prediction stability, and simple signal statistics to identify unreliable or adversarial inputs. Experiments on GTSRB with a ResNet-18 under diverse white-box and black-box attacks show substantial gains in reliability on accepted samples. Strong

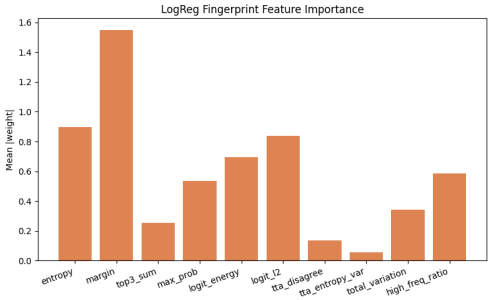

Figure 3: Mean absolute weights of the logistic regression fingerprint classifier, indicating the relative importance of each fingerprint feature.

optimization-based attacks such as CW are almost completely filtered, while black-box attacks such as Square show large improvements in selective accuracy, with clean accuracy remaining high. We also observe that strong iterative ℓ_∞ attacks remain challenging due to their overlap with the clean distribution, highlighting an inherent coverage–reliability trade-off. Overall, these results show that fingerprint-based selective rejection provides a simple, training-free, and effective complementary robustness layer.

References

[1] A. Abomakhelb *et al.*, "A comprehensive review of adversarial attacks and defense strategies in deep neural networks," *Technologies*, vol. 13, no. 5, 2025.

[2] M. Macas, "Adversarial examples: A survey of attacks and defenses," *Elsevier*, 2024.

[3] D. C. Asimopoulos, P. Radoglou-Grammatikis, P. Fouliras, K. Panitsidis, G. Efstathopoulos, T. Lagkas, V. Argyriou, I. Kotsiuba, and P. Sarigiannidis, "Surrogate-guided adversarial attacks: Enabling white-box methods in black-box scenarios," in *2025 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2025, pp. 950–956.

[4] D. C. Asimopoulos, P. Radoglou-Grammatikis, T. Lagkas, V. Argyriou, I. Moscholios, J. Cani, G. T. Papadopoulos, E. K. Markakis, and P. Sarigiannidis, "Aag: Adversarial attack generator for evaluating the robustness of machine learning models against adversarial attacks," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 2682–2689.

[5] N. Sachpelidis-Brozos, E. Katsaros, P. Radoglou-Grammatikis, G. Kalitsios, A. Sarigiannidis, G. C. Seritan, I. Politis, C. Xenakis, S. Goudos, and P. Sarigiannidis, "Hacking intelligence: Mapping the anatomy of adversarial threats in artificial intelligence with mitre atlas," *Computer Science Review*, vol. 61, p. 100923, 2026.

[6] D. C. Asimopoulos, P. Radoglou-Grammatikis, I. Makris, V. Mladenov, K. E. Psannis, S. Goudos, and P. Sarigiannidis, "Breaching the defense: Investigating fgsm and ctgan adversarial attacks on iec 60870-5-104 ai-enabled intrusion detection systems," ser. ARES '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3600160.3605163

[7] K. Anderson *et al.*, "Adversarial training techniques for hardening deep neural networks in cybersecurity applications," *ResearchGate Publication*, 2025.

[8] K. Barik, S. Misra, and L. Fernandez-Sanz, "Adversarial attack detection framework based on optimized weighted conditional stepwise adversarial network," *International Journal of Information Security*, vol. 23, no. 3, pp. 2353–2376, 2024.

[9] H. Fisher, M. Shahar, and Y. S. Resheff, "Neural fingerprints for adversarial attack detection," 2024, arXiv:2411.04533.

[10] Y. Feng, T. G. J. Rudner, and N. Tsilivis, "On the adversarial robustness of bayesian neural networks," 2024, openReview.