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‡

***Abstract*—While deep neural networks are highly susceptible to adversarial attacks, many defence mechanisms require retraining, architectural changes, or sophisticated optimisation processes. In this work, we present a post-processing defence system that can be used at inference time as an external decision layer for a fixed pretrained classifier. The proposed defence mechanism follows a simple three-stage pipeline; it uses a median filter-based recovery stage to partly mitigate adversarial attacks, a statistical residual detector using class-conditional Mahalanobis distances for input evaluation, and finally a confidence-aware selective prediction stage to accept or reject a prediction. The defense mechanism improves the robustness of a given classifier against gradient-based adversarial attacks without retraining or modifying it. The experimental results show that the proposed defense mechanism improves post-attack classification accuracy with risk-coverage trade-offs while maintaining high accuracy for accepted samples under extreme distribution shift. Due to its simplicity and interpretability, the proposed defense mechanism can be used in resource-constrained environments.**

***Index Terms*—Adversarial Attacks, Artificial Intelligence, Black-Box Attacks, Evasion Attacks, White-Box Attacks**

I. Introduction

Despite the exceptional performance of deep neural networks, they are highly vulnerable to adversarial examples—inputs that cause incorrect predictions through imperceptible perturbations. Since the advent of gradient-based attacks such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), numerous defense strategies have been proposed, including adversarial training, input preprocessing, and detection-based methods [1].

There has been a growing need for defense strategies for neural networks in real-world scenarios or deployments, especially in scenarios where the system is resource-constrained or safety-critical. In such scenarios, post-processing-based defense strategies have been reported to be highly desirable due to the low engineering or computational overhead of such methods. Preprocessing techniques like denoising or filtering have been shown to be promising in mitigating the effects of adversarial perturbations in images or inputs.

However, as standalone techniques, recovery-based methods are often insufficient to restore correct predictions, especially when faced with iterative or strong attacks. In addition to recovery-based techniques, statistical detection techniques have emerged to detect adversarial samples by capitalizing on the distributional mismatches between clean and adversarial data. Distance-based detectors in the feature space offer a clear explanation for the quantification of deviations from class-conditional norms. When combined with selective prediction techniques, these detectors enable systems to abstain from predictions with high risks instead of forcing potentially incorrect predictions, thus explicitly managing the trade-off between prediction coverage and risk.

In this work, we propose a lightweight post-processing defense that requires no training and runs entirely at inference time, treating the target classifier as a black-box. Our pipeline starts with a simple perturbation recovery step to partially undo adversarial effects at the input level. Next, we use a statistical residual detector with class-conditional Mahalanobis distances computed from low-dimensional, interpretable image statistics. Finally, our confidence-aware approach to selective prediction enables explicit decisions to predict or not predict, skipping any modification to the underlying model itself.

The main contributions of this work are summarized as follows:

- **Lightweight Post-Processing Defense Framework:** We propose a modular post-processing defense pipeline to strengthen robustness against adversarial perturbations without the need for retraining or changing the target neural network, thus allowing for easy integration into the existing inference pipelines.

‡‡ 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.

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

†Dimitrios-Christos Asimopoulos and V.Vitsas are with the Department of Information and Electronic Engineering, International Hellenic University, Sindos Campus 57400, Thessaloniki, Greece – E-Mail: {dimiasim3, vitsas}@ihu.gr

‡P. Radoglou-Grammatikis, I. Ktenidis and P. Sarigiannidis are with the Department of Electrical and Computer Engineering, University of Western Macedonia, Campus ZEP Kozani, 50100 Kozani, Greece – E-Mail:@uowm.gr

§P. Radoglou-Grammatikis and A. Sarigiannidis is with K3Y, Sofia, Bulgaria – E-Mail: {pradoglou, asarigia}@k3y.bg

‖T. Lagkas is with the Department of Computer Science, Democritus University of Thrace, Kavala Campus, 65404, Kavala, Greece – E-Mail: tlagkas@cs.duth.gr

**C. Xenakis is with the Department of Digital Systems,University of Piraeus, Piraeus, 18534, Greece – E-Mail: xenakis@unipi.gr

††E. K. Markakis is with the Hellenic Mediterranean University, 71004 Heraklion, Greece – E-Mail: emarkakis@hmu.gr

- **Perturbation Recovery via Median Filtering:** A simple yet effective recovery step based on a median filter is used on adversarial examples to partially remove perturbations and thus enhance the classification performance after the attack.
- **Statistical Residual Detection in a Low-Dimensional Feature Space:** We propose a class-conditional Mahalanobis residual detector that works on standardized image-levelstatistical features such as variance, standard deviation, mean, skewness, kurtosis, and entropy, thus giving a detection method that is both interpretable and computationally efficient.
- **Selective Prediction with Confidence and Residual Gating:** Combining residual-based detection with confidence thresholding, the suggested method facilitates selective classification and gives explicit control over the trade-off between prediction coverage and classification risk.
- **Comprehensive Experimental Evaluation:** A thorough evaluation of the MNIST dataset under both FGSM and PGD attacks at different perturbation strength levels is conducted, providing detailed analyses of the results including classification accuracy, F1, score, residual behavior, rejection rates, and risk, coverage trade offs.

The remainder of this paper is organized as follows. Section II reviews related work on adversarial attacks and defense mechanisms. Section III describes a representative user story that motivates and illustrates the deployment scenario of the proposed pipeline, while section IV presents the proposed system architecture and provides an overview of the post-processing defense framework. Section V describes the methodology, including the threat model, recovery process, statistical residual detection, and selective prediction strategy. Section VI details the experimental setup and reports the evaluation results. Finally, Section VII concludes the paper and discusses directions for future work.

II. Related Work

Deep neural networks (DNNs) are found to be highly susceptible to adversarial attacks, which are specifically crafted inputs to the DNNs to compromise the correct functioning of the DNNs. Recently, the focus of the research community is to consider more realistic threat models, such as black-box, transferable, and adaptive attacks, rather than the traditional white-box setting. Asimopoulos et al. [2] have demonstrated the impact of gradient-based and generative attacks on the functioning of AI-based intrusion detection systems used in the industrial control system networks. Further, the same researchers have demonstrated the applicability of the surrogate-based approach to convert the white-box attacks to black-box attacks using the transferability property of the DNNs [3]. Recent surveys have also categorized the various types of adversarial attacks and threat models for the DNNs. Macas et al. [4] have conducted an extensive survey on the various types of attacks and defense mechanisms for DNNs used in the field of cybersecurity. Keyizhi et al. [5] have conducted an extensive survey on the various types of gradient-based, decision-based, and generative attacks on the DNNs used in the field of computer vision. Goodfellow et al. [6] have conducted the pioneering work on the FGSM.

Adversarial training is one of the most used defense methods but suffers from high computational costs and attack configuration dependence. Wu et al. [7] presented a survey of adversarial defenses in terms of training methods, inference methods, and architectural designs. Robustness has been reported to suffer from overfitting issues for certain attack methods. Zhao et al. [8] extended this research to robustness-accuracy-efficiency-convergence stability trade-offs. Certified defenses provide robustness guarantees in terms of bounded adversarial attacks but suffer from high computational costs and loose bound conditions. Qiao et al. [9] presented a causal defense method with high accuracy in terms of generalization. Furthermore, Kumar et al. [10] discussed the potential of using certified defenses in conjunction with empirical defenses for closing the gap between theory and practice in robustness. Preprocessing-based defenses involve applying filters or other operations before inference to reduce the impact of adversarial attacks. Abomakhelb et al. [11] presented a review of preprocessing-based methods for reducing the impact of adversarial attacks. Spatial filtering has been reported to recover some accuracy in the presence of attacks. Huang et al. [12] demonstrated that frequency-domain reconstruction can expose adversarial artifacts invisible in pixel space. These methods are light-weight and model-agnostic, and hence are appropriate for real-time systems; yet, they are still susceptible to adaptive attackers when applied alone. Recent literature surveys [13] highlight that the integration of preprocessing techniques with detection or selective methods can greatly enhance robustness in safety-critical settings.

Detection-based defense strategies seek to differentiate benign examples from adversarial examples based on statistical, feature-level, or confidence-based indicators. Li et al. [14] developed an interpretable detection method based on high-level concept activation vectors, which enhanced explainability without compromising accuracy. Frequency-domain reconstruction techniques [12] also expose adversarial signals concealed in the pixel domain. Though very effective, detection-based tightening strategies by themselves may suffer from threshold problems and loss of robustness against adaptive attacks. Wu et al. [7] pointed out that a mixture of detection, preprocessing, and selective defense can help solve these problems.

By selective detection models can reject samples that are uncertain or adversarial, and in doing so, explicitly make the trade-off between accuracy and reliability. Rabanser et al. [15] came up with a rejection technique by means of training dynamics that extends confidence, driven rejection. Angelopoulos et al. [16] have set risk, coverage analysis in conformal prediction on a formal basis, thus presenting a mathematically sound method for uncertainty estimation. Such methods prove to be a great fit for situations when a model's abstaining is favored over its making high, confidence errors.

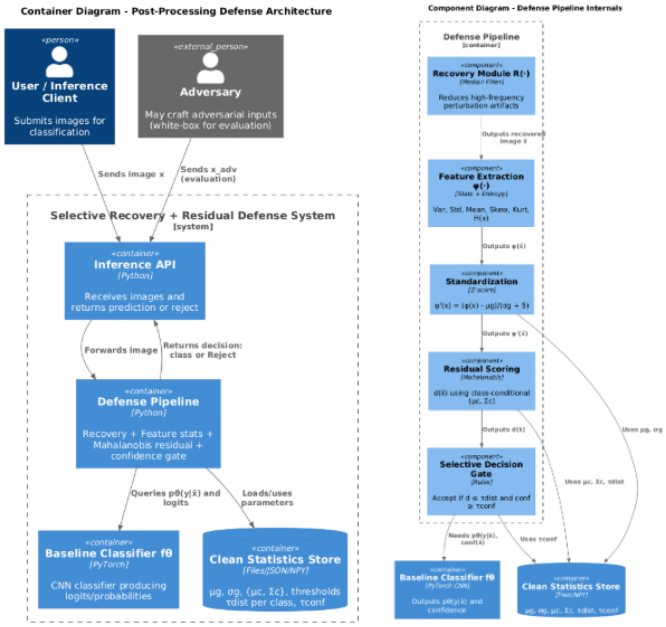


Figure 1: C4 container view of the proposed post-processing defense architecture. The system operates as an external inference layer that integrates a recovery and residual-based defense pipeline with a fixed baseline classifier and precomputed clean statistics.

Unlike previous work, this paper integrates lightweight perturbation recovery, statistical residual detection, and confidence-driven selective prediction into a single post-processing defense strategy. The approach requires no retraining or model modification, instead using only statistics of clean data. By integrating preprocessing, detection, and selective rejection, the proposed defense framework directly tackles deployment issues highlighted in recent adaptive and transferable attack works [2]–[4], providing a well-rounded balance between interpretability, scalability, and robustness.s.

III. Case Study: Deployment of the Proposed Method

To provide practical context for the proposed approach, we consider a typical deployment setting in which a practitioner manages a pretrained machine-learning model for anomaly detection within a live monitoring environment, such as industrial sensing infrastructures, network traffic analysis, or control-system logging pipelines, as shown in Fig. 2. The model has been trained solely on clean operational data and is already integrated into a production system. In such settings, updating the training process or modifying the model architecture is often impractical due to operational limitations, computational cost, or regulatory constraints.

As the system operates over time, it may encounter carefully manipulated or adversarially perturbed inputs that are difficult to distinguish from legitimate data at the raw signal level. While these inputs may appear innocuous, they can cause the pretrained model to generate erroneous predictions,

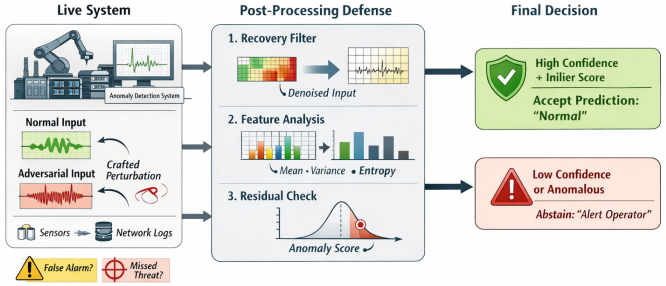


Figure 2: Deployment of the proposed post-processing defense, combining recovery, residual-based detection, and selective prediction at inference time.

frequently accompanied by high confidence, which is particularly hazardous in safety-critical applications.

Rather than modifying the model itself, the practitioner adds a defense layer to the inference pipeline via a lightweight post-processing operation. Each new sample is initially treated by a minor recovery operation that mitigates non-natural high-frequency components. The recovered input is then compared to clean reference statistics that were learned offline to determine its statistical conformity. Finally, the mechanism uses both this consistency evaluation and model prediction confidence to decide if a decision is to be accepted or refused. If an input sharply deviates from normal behavior or if model confidence is considered lower, the system purposely refrains from producing a prediction and marks the input for further investigation or safe handling. This architectural decision favors reliability instead of completeness hence allowing safe usage of pretrained models in adversarial or uncertain situations without retraining.

IV. System Architecture

The proposed framework employs a modular and sequential structure that is intended for improving adversarial robustness via post-processing without modifying or retraining the classifier. The defense mechanism is realized as an inference layer that can be used with any pre-trained classifier. This allows for easy adaptation and incorporation into existing classification pipelines.

A high-level description of the overall structure of the proposed system is presented in Fig. 1, which employs a C4 container model [17]. At this abstraction level, the overall structure of the system comprises an inference interface, a defense pipeline, a pre-trained baseline classifier, and precomputed clean statistics. The input samples, which could be either clean or adversarial examples, are passed into the inference interface and then into the defense pipeline.

As illustrated in Fig. 1, the baseline neural classifier f_θ is trained and fixed only on clean data. The classifier is used by the defense pipeline to provide class posterior probabilities and confidence scores, while the clean statistics store provides global normalization parameters, class-conditional distribution parameters, and decision thresholds necessary for residual scoring and selective prediction.

The internal architecture of the defense pipeline is illustrated in Fig. 1 in a C4 component diagram. The diagram illustrates the sequential processing steps taken on the input. The defense pipeline starts with a recovery module that performs a median filtering step to remove high-frequency perturbation artifacts that are typically introduced by gradient-based attacks, while retaining the coarse structural information of the input image.

Once recovered, the processed image is converted into a compact statistical form via a feature extraction process, as shown in Fig. 1. In essence, a set of low-dimensional global statistics is computed and standardized using global statistics obtained from clean samples.

To capture normal behavior, the standardized features are used to compare against a set of statistical reference models that are associated with each class. As shown in Fig. 1, the residual scoring process involves calculating the Mahalanobis distance between the recovered input and each class-conditional distribution and retaining the minimum distance value. The minimum distance formulation allows for robustness against incorrect class predictions under attack and avoids using labels that are likely to be corrupted. The Mahalanobis distance is selected because it accounts for the covariance structure of the 6-D feature space, requires no learned parameters beyond the per-class mean and inverse covariance estimated from clean data, and admits closed-form calibration via a high-percentile statistic of the clean-data distance distribution (Eq. **??** below).

The final stage in our pipeline is used for combining residual consistency and classifier confidence for selective predictions. In essence, predictions are only made if the minimum Mahalanobis distance is below a certain threshold and classifier confidence is above a certain value; otherwise, no predictions are made. This process allows for an inherent trade-off between coverage and reliability, which is especially useful for safety-critical and high-risk deployments.

Overall, the architecture shown in Fig. 1 emphasizes modularity, interpretability, and deployment practicality. Each component of the defense pipeline can be independently modified or extended without affecting the baseline classifier, allowing the system to adapt to different threat models, feature representations, or operational constraints.

V. Methodology

A. Problem Formulation and Notation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a labeled dataset where each image $x \in \mathbb{R}^{H\times W}$ and $y \in \{1, \dots, C\}$. A neural classifier f_θ maps an input image to class logits $z_\theta(x) \in \mathbb{R}^C$ and posterior probabilities via the softmax function:

$$p_\theta(y = k \mid x) = \frac{\exp(z_{\theta,k}(x))}{\sum_{j=1}^{C} \exp(z_{\theta,j}(x))}. \tag{1}$$

The predicted label and confidence are defined as

$$\hat{y}(x) = \arg\max_k p_\theta(y = k \mid x), \tag{2}$$

$$\mathrm{conf}(x) = \max_k p_\theta(y = k \mid x). \tag{3}$$

An adversarial example x^{adv} is generated by an attack $\mathcal{A}$ such that

$$x^{\mathrm{adv}} = \mathcal{A}(x, y, f_\theta), \qquad \|x^{\mathrm{adv}} - x\|_\infty \le \epsilon, \tag{4}$$

while inducing misclassification.

B. Threat Model and Attacks

We adopt a white-box threat model, where the adversary has full knowledge of the model parameters and loss function. We consider two widely used ℓ_∞-bounded gradient-based attacks.

a) FGSM: The Fast Gradient Sign Method perturbs the input in a single step:

$$x^{\mathrm{adv}} = \mathrm{clip}\left(x + \epsilon \cdot \mathrm{sign}\left(\nabla_x \mathcal{L}(f_\theta(x), y)\right)\right). \tag{5}$$

b) PGD: Projected Gradient Descent applies iterative perturbations with projection onto the feasible ℓ_∞-ball:

$$\begin{aligned} x^{(0)} &= x, \\ x^{(t+1)} &= \Pi_{\mathcal{B}_\infty(x,\epsilon)}\Big(\mathrm{clip}\big(x^{(t)} \\ &\qquad + \alpha \cdot \mathrm{sign}(\nabla_{x^{(t)}} \mathcal{L}(f_\theta(x^{(t)}), y))\big)\Big), \end{aligned} \tag{6}$$

for $t = 0, \dots, T-1$.

C. Perturbation Recovery

Adversarial inputs are first processed by a recovery operator $\mathcal{R}(\cdot)$:

$$\tilde{x} = \mathcal{R}(x^{\mathrm{adv}}), \tag{7}$$

implemented as a 2D median filter with a 3×3 kernel. This is the smallest window that suppresses isolated pixel-level perturbations while preserving stroke geometry on low-resolution inputs such as MNIST. Larger kernels were observed to oversmooth digit edges and reduce clean accuracy. This step suppresses localized high-frequency perturbation artifacts while preserving coarse structural information.

D. Statistical Feature Extraction

From each image x, we extract a low-dimensional statistical descriptor $\phi(x) \in \mathbb{R}^6$:

$$\phi(x) = \big[\mathrm{Var}(x),\ \mathrm{Std}(x),\ \mathrm{Mean}(x),\ \mathrm{Skew}(x),\ \mathrm{Kurt}(x),\ H(x)\big], \tag{8}$$

where the Shannon entropy [18] is defined as

$$H(x) = -\sum_{b=1}^{B} p_b(x) \log\big(p_b(x) + \delta\big), \tag{9}$$

with $\delta = 10^{-12}$ acting as a small positive constant that prevents $\log(0)$ when a histogram bin is empty.

E. Standardization and Class-Conditional Modeling

To ensure consistent scaling, features are standardized using global statistics computed from clean data:

$$\phi'(x) = \frac{\phi(x) - \mu_g}{\sigma_g + \delta}. \tag{10}$$

where $\delta = 10^{-8}$ ensures numerical stability when σ_g approaches zero.

For each class c, standardized features are modeled using a multivariate Gaussian distribution:

$$\phi'(x) \mid y = c \sim \mathcal{N}(\mu_c, \Sigma_c), \tag{11}$$

with parameters estimated from clean samples belonging to class c.

F. Residual Detection and Selective Prediction

To quantify residual statistical deviations without relying on potentially corrupted predictions, the Mahalanobis distance between a recovered input $\tilde{x}$ and each class-conditional distribution is computed:

$$d_c(\tilde{x}) = \sqrt{(\phi'(\tilde{x}) - \mu_c)^\top \Sigma_c^{-1} (\phi'(\tilde{x}) - \mu_c)}. \tag{12}$$

The residual score is defined as the minimum distance over all classes:

$$d_{\min}(\tilde{x}) = \min_{c \in \{1,\dots,C\}} d_c(\tilde{x}). \tag{13}$$

A global distance threshold τ^{dist} is calibrated on clean data using a high-percentile statistic of $d_{\min}$:

$$\tau^{\text{dist}} = \text{Quantile}_q \left(\{ d_{\min}(x_i) \}_{i=1}^N \right). \tag{14}$$

Selective prediction is performed by jointly evaluating statistical consistency and classifier confidence. A prediction is accepted if:

$$d_{\min}(\tilde{x}) \leq \tau^{\text{dist}} \quad \wedge \quad \text{conf}(\tilde{x}) \geq \tau^{\text{conf}}, \tag{15}$$

otherwise the system abstains from making a prediction.

VI. Experiment Setup

A. Dataset

Experiments are conducted on the MNIST handwritten digit dataset, which consists of grayscale images of digits from 0 to 9. Each image has a spatial resolution of 28×28 pixels, yielding a single-channel input representation. The dataset contains a total of 60,000 training images and 10,000 test images, evenly distributed across the ten classes.

The baseline classifier is trained exclusively on the clean training set and evaluated on the full test set. No adversarial samples are used during training. All pixel values are normalized to the range $[0, 1]$. The MNIST dataset is widely used as a benchmark for adversarial robustness studies, allowing controlled analysis of attack strength, recovery behavior, and selective defense performance.

B. Evaluation Metrics

We evaluate the pipeline using standard classification metrics computed in a *one-vs-rest macro-averaged* formulation across the ten MNIST classes: accuracy, F1-score, true positive rate (TPR), and false positive rate (FPR), defined in the usual way from per-class TP, TN, FP, FN counts and averaged across classes with equal weight. In addition, we report: *Coverage* $= |\mathcal{D}_{\text{accepted}}|/|\mathcal{D}|$, *Rejection Rate* $= 1 - \text{Coverage}$, *Accuracy Gain* $\Delta\text{Acc} = \text{Acc}_{\text{Rec}} - \text{Acc}_{\text{Atk}}$, and *Selective Risk* $= 1 - \text{Acc}_{\text{accepted}}$. The Mahalanobis residual distance $d_{\min}(\tilde{x})$ and classifier confidence $\text{conf}(x)$ are defined in Section V; all conditional metrics are computed over the set $\mathcal{D}_{\text{accepted}} = \{x : d_{\min}(x) \leq \tau^{\text{dist}} \wedge \text{conf}(x) \geq \tau^{\text{conf}}\}$.

C. Experimental Results

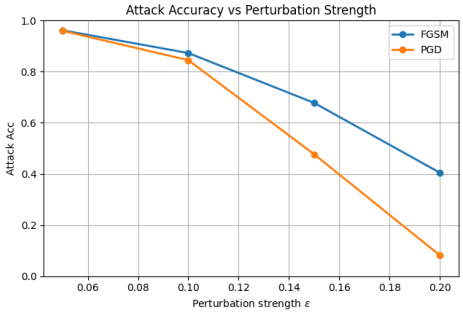
We evaluate the proposed post-processing defense pipeline on MNIST using a convolutional neural network (CNN) classifier. Performance is reported across four pipeline stages: (i) clean inference, (ii) under adversarial attack, (iii) after perturbation recovery, and (iv) after selective defense. We consider FGSM and PGD attacks with varying perturbation strength ϵ. Unless otherwise stated, the selective defense uses confidence threshold $\tau = 0.8$, Mahalanobis ratio cap $r_{\max} = 5.0$, and PGD is run for 20 iterations.

1) Clean Baseline Performance: On clean test data, the classifier achieves $\text{Acc} = 0.9890$, $\text{F1} = 0.9890$, $\text{TPR} = 0.9888$, and $\text{FPR} = 0.0012$, establishing a strong baseline for robustness evaluation.

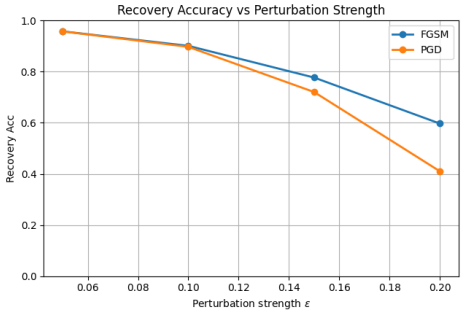
2) Performance Under Attack and After Recovery: Figure 3 summarizes the effect of adversarial perturbations and the impact of the recovery module. As ϵ increases, attack accuracy degrades monotonically for both methods, with iterative PGD causing substantially stronger degradation than single-step FGSM. The median filter recovery partially restores performance across all ϵ values, demonstrating that lightweight preprocessing can suppress a portion of the adversarial signal without retraining the classifier.

3) Recovery Baseline Comparison: To validate the choice of median filtering over alternative low-pass operators, Table I compares the recovery accuracy of the proposed median filter against Gaussian smoothing with $\sigma = 0.5$, selected to match the spatial extent of the median kernel for a fair strength-matched comparison. The median filter is strictly more effective for all $\epsilon \geq 0.10$, with the advantage growing under stronger attacks — reaching a 23.8-point accuracy gap under PGD at $\epsilon = 0.20$. At $\epsilon = 0.05$ the two operators are comparable, as the attack is weak enough that either smoothing operator suffices. The advantage of median filtering stems from its ability to suppress isolated pixel-level adversarial spikes without blurring stroke edges, whereas Gaussian smoothing degrades both signal and noise uniformly.

4) Residual Detection Strength and Coverage Collapse: After recovery, residual artifacts are quantified via the standardized class-conditional Mahalanobis residual detector (min-distance scoring). Figure 4 shows that the average Mahalanobis ratio increases monotonically with ϵ for both



(a) Attack accuracy vs. ϵ



(b) Recovery accuracy vs. ϵ

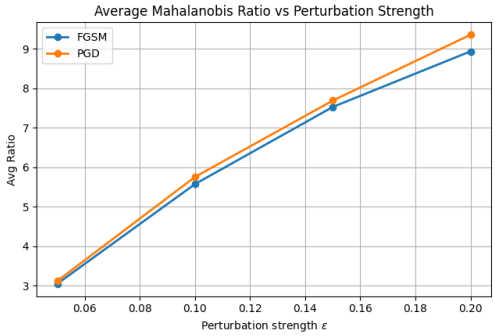
Figure 3: Pipeline stage metrics as a function of perturbation strength ϵ. Left: classification accuracy under attack. Right: accuracy after median-filter recovery.

Table I: Recovery accuracy: median filter vs. Gaussian smoothing ($\sigma = 0.5$). Δ is the median-filter advantage in percentage points.

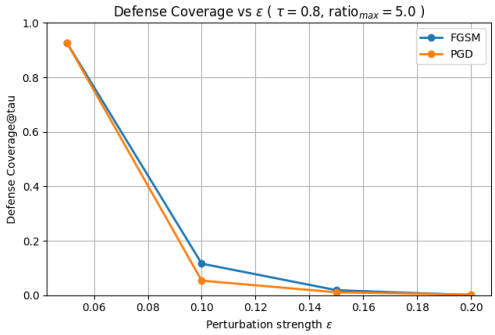
Attack	ϵ	Median Acc	Gaussian Acc	Δ
FGSM	0.05	0.9584	0.9649	-0.7
FGSM	0.10	0.8988	0.8845	$+1.4$
FGSM	0.15	0.7695	0.7084	$+6.1$
FGSM	0.20	0.5840	0.4652	$+11.9$
PGD	0.05	0.9583	0.9637	-0.5
PGD	0.10	0.8946	0.8687	$+2.6$
PGD	0.15	0.7046	0.5757	$+12.9$
PGD	0.20	0.3963	0.1583	$+23.8$

FGSM and PGD, indicating growing statistical deviation from clean class-conditional distributions. Correspondingly, defense coverage decreases sharply as ϵ increases: at $\epsilon = 0.05$ the system accepts over 92% of samples, while at $\epsilon = 0.20$ nearly all samples are rejected. This behavior is desirable in safety-critical settings, as it prevents forced predictions on highly perturbed inputs.

5) Selective Prediction Quality (Conditional Accuracy): To assess the reliability of accepted predictions, we report defended accuracy computed conditionally on accepted samples. Figure 5 shows that for low and moderate perturbation strengths, the conditional accuracy remains high (above 0.86 for FGSM and 0.75 for PGD at $\epsilon = 0.15$). At $\epsilon = 0.20$, the drop in conditional accuracy is coupled with near-zero coverage, indicating that almost all inputs are rejected rather than producing incorrect high-confidence decisions.



(a) Avg. Mahalanobis ratio vs. ϵ



(b) Defense coverage vs. ϵ ($\tau = 0.8$, $r_{\max} = 5.0$)

Figure 4: Residual behavior and selective acceptance. Left: the residual (Mahalanobis) ratio grows with perturbation strength. Right: defense coverage decreases sharply as attacks become stronger.

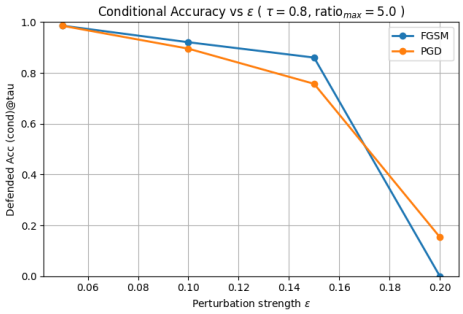
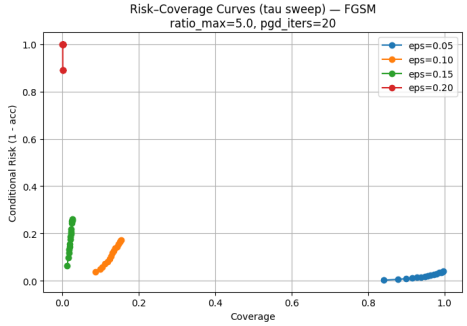


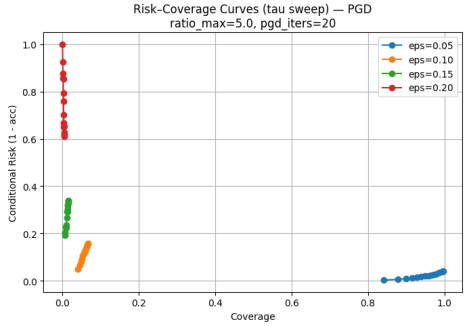
Figure 5: Conditional accuracy under selective prediction as a function of ϵ ($\tau = 0.8$, $r_{\max} = 5.0$).

6) Risk–Coverage Trade-off: We further analyze the selective prediction behavior by sweeping the confidence threshold τ and plotting risk–coverage curves. As shown in Figure 6, higher ϵ values shift the curves toward lower achievable coverage at a given risk level, reflecting the increased difficulty of robust classification under stronger perturbations.

7) Rejection Control via Confidence Threshold: Finally, Figure 7 reports rejection rate as a function of τ. Increasing τ consistently increases rejection across all ϵ values, confirming that the framework exposes a tunable operating point between coverage and reliability.



(a) FGSM



(b) PGD

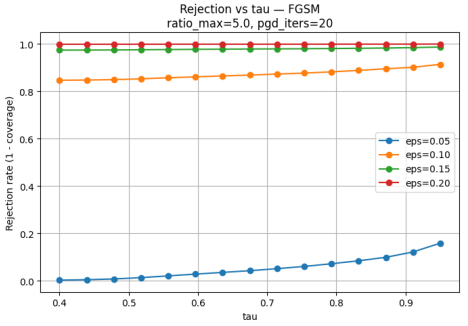
Figure 6: Risk–coverage curves obtained by sweeping τ for multiple perturbation strengths ($r_{\max} = 5.0$, PGD iters $= 20$).

Table II: Unconditional performance across pipeline stages (clean $\rightarrow$ attack $\rightarrow$ recovery).

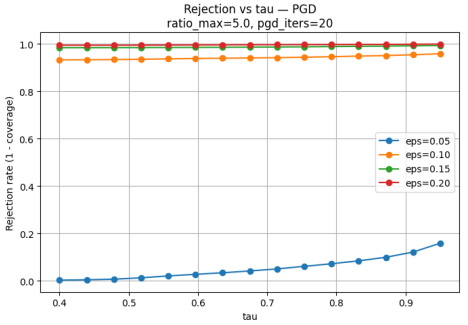
Attack	ϵ	Clean Acc	Attack Acc	Recovery Acc
FGSM	0.05	0.9890	0.9621	0.9580
FGSM	0.10	0.9890	0.8730	0.9012
FGSM	0.15	0.9890	0.6782	0.7772
FGSM	0.20	0.9890	0.4049	0.5972
PGD	0.05	0.9890	0.9601	0.9578
PGD	0.10	0.9890	0.8452	0.8972
PGD	0.15	0.9890	0.4772	0.7206
PGD	0.20	0.9890	0.0817	0.4106

8) Quantitative Summary (Clean, Attack, Recovery, Defense): Table II reports *unconditional* classification accuracy across the first three stages of the pipeline. The clean baseline establishes the reference performance of the fixed classifier. Under attack, accuracy decreases monotonically with ϵ, with iterative PGD producing substantially stronger degradation than single-step FGSM. The recovery stage (median filtering) consistently restores a significant portion of accuracy for both attacks, showing that lightweight input recovery can suppress part of the adversarial perturbation without retraining or modifying the classifier.

Table III reports the behavior of the full defense stage, which combines Mahalanobis residual detection with confidence gating to enable *selective prediction*. Coverage denotes the fraction of inputs for which the system returns a prediction, while rejection is its complement. Conditional accuracy is computed only over accepted samples, and selective risk



(a) FGSM



(b) PGD

Figure 7: Rejection rate $(1-\text{coverage})$ vs. confidence threshold τ for different perturbation strengths ($r_{\max} = 5.0$, PGD iters $= 20$).

Table III: Selective defense performance after recovery ($\tau = 0.8$, $r_{\max} = 5.0$). Metrics are computed conditionally on accepted samples.

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

is defined as $1 - \text{Cond. Acc}$. At low perturbation strengths ($\epsilon = 0.05$), the defense maintains high coverage (above 92%) while achieving near-clean conditional accuracy (approximately 98.5%). As ϵ increases, the defense increasingly rejects inputs due to stronger distributional deviation, causing coverage collapse at $\epsilon = 0.20$. This behavior reflects an explicit trade-off between coverage and reliability, preventing high-confidence errors under severe attacks.

9) Ablation of Pipeline Components: To isolate the contribution of each component, Table IV evaluates all subsets of {Recovery (R), Mahalanobis residual detector (M), Confidence gate (C)} under FGSM; PGD follows the same pattern. Recovery (R) drives the unconditional accuracy gain, the confidence gate (C) sustains high conditional accuracy at low to moderate ϵ (above 0.98 at $\epsilon = 0.05$), and the Maha-

Table IV: Ablation of pipeline components under FGSM. Each cell reports Coverage / Conditional Accuracy. R: median-filter recovery; M: Mahalanobis residual detector; C: confidence gate.

Config	ϵ**=0.05**	ϵ**=0.10**	ϵ**=0.15**	ϵ**=0.20**
None	1.00 / 0.961	1.00 / 0.865	1.00 / 0.661	1.00 / 0.382
R	1.00 / 0.958	1.00 / 0.899	1.00 / 0.770	1.00 / 0.584
C	0.92 / 0.990	0.81 / 0.931	0.65 / 0.736	0.54 / 0.355
M	1.00 / 0.961	0.94 / 0.860	0.21 / 0.570	0.00 / 0.038
R+C	0.92 / 0.986	0.84 / 0.957	0.72 / 0.860	0.60 / 0.656
R+M	1.00 / 0.959	0.15 / 0.824	0.02 / 0.738	0.00 / 0.143
M+C	0.92 / 0.990	0.76 / 0.928	0.13 / 0.638	0.00 / 0.043
R+M+C (full)	0.92 / 0.986	0.11 / 0.921	0.02 / 0.836	0.00 / 0.000

Table V: Inference timing on the MNIST test set (10,000 samples), CPU.

Stage	**Total (s)**	**Per sample (μs)**
Median filter recovery	0.26	26
Feature extraction + Mahalanobis scoring	≈ 0.05	≈ 5
Full defense pipeline	$\approx$ **0.31**	$\approx$ **31**
Gaussian smoothing (baseline)	0.20	20
PGD-20 attack generation (reference)	≈ 47	≈ 4700

lanobis detector (M) triggers coverage collapse under heavy distribution shift ($\epsilon \geq 0.15$), enforcing conservative abstention when no class-conditional reference matches the input. This reflects an explicit reliability-over-completeness trade-off: at FGSM $\epsilon = 0.15$, R+C alone yields 71.8% coverage at 86.0% conditional accuracy, while the full R+M+C pipeline restricts to 1.8% coverage at 83.6% accuracy.

10) Computational Cost: Table V reports per-stage inference timing on the full MNIST test set (10,000 samples) on commodity CPU hardware. The complete defense pipeline costs approximately 31 μs per sample, dominated by median filtering (26 μs); statistical feature extraction and Mahalanobis scoring contribute negligibly. This overhead is two to three orders of magnitude smaller than the cost of generating a single PGD-20 adversarial example (≈ 4.7 ms/sample) and is well within the budget of resource-constrained inference pipelines targeted by the case study in Section III.

VII. Conclusion & Future Work

This paper presented a lightweight post-processing defense framework that enhances robustness against gradient-based adversarial attacks without modifying or retraining the underlying classifier. The proposed pipeline combines a simple perturbation recovery step based on median filtering with a statistical residual detector operating on low-dimensional image features, followed by confidence-based selective prediction. Experimental results on the MNIST dataset under FGSM and PGD attacks demonstrate that median filtering outperforms Gaussian smoothing as a recovery operator, and that recovery, confidence gating, and Mahalanobis detection each contribute distinctly to the pipeline, which runs in approximately 31 μs per sample on commodity CPU hardware. At low perturbation strengths, the method maintains high coverage and near-clean conditional accuracy, whereas at stronger attack levels it gracefully abstains from unreliable predictions, explicitly trading coverage for reduced risk.

Future work will extend the approach to higher-resolution and more complex datasets, explore richer learned or statistical feature representations for finer residual sensitivity and investigate adaptive thresholding and alternative distance metrics.

References

[1] 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.

[2] 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," in *Proceedings of the 18th International Conference on Availability, Reliability and Security*, 2023, pp. 1–8.

[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)*. IEEE, 2025, pp. 950–956.

[4] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," *Expert Systems with Applications*, vol. 238, p. 122223, 2024.

[5] X. Keyizhi, L. Yajuan, W. Zhongyuan *et al.*, "A survey of adversarial examples in computer vision: Attack, defense, and beyond," *Wuhan University Journal of Natural Sciences*, vol. 30, no. 1, pp. 1–20, 2025.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] B. Wu, S. Wei, M. Zhu, M. Zheng, Z. Zhu, M. Zhang, H. Chen, D. Yuan, L. Liu, and Q. Liu, "Defenses in adversarial machine learning: A survey," *arXiv preprint arXiv:2312.08890*, 2023.

[8] M. Zhao, L. Zhang, J. Ye, H. Lu, B. Yin, and X. Wang, "Adversarial training: A survey," *arXiv preprint arXiv:2410.15042*, 2024.

[9] Y. Qiao, Y. Yin, C. Chen, and J. Ma, "Certified causal defense with generalizable robustness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 19, 2025, pp. 20 024–20 032.

[10] S. Kumar, S. Srivastava, and S. Banerjea, "Fortifying vision models: A comprehensive survey of defences against adversarial examples," *Applied Soft Computing*, p. 113874, 2025.

[11] A. Abomakhelb, K. A. Jalil, A. G. Buja, A. Alhammadi, and A. M. Alenezi, "A comprehensive review of adversarial attacks and defense strategies in deep neural networks," *Technologies*, vol. 13, no. 5, p. 202, 2025.

[12] S. Huang, Z. Zhang, and B. Song, "A novel adversarial example detection method based on frequency domain reconstruction for image sensors," *Sensors*, vol. 24, no. 17, p. 5507, 2024.

[13] A. D. M. Ibrahum, M. Hussain, and J.-E. Hong, "Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective," *Artificial Intelligence Review*, vol. 58, no. 1, p. 28, 2024.

[14] J. Li, Y.-a. Tan, X. Liu, W. Meng, and Y. Li, "Interpretable adversarial example detection via high-level concept activation vector," *Computers & Security*, vol. 150, p. 104218, 2025.

[15] S. Rabanser, A. Thudi, K. Hamidieh, A. Dziedzic, I. Bahceci, A. B. Sediq, H. Sokun, and N. Papernot, "Selective prediction via training dynamics," *Transactions on Machine Learning Research*.

[16] A. N. Angelopoulos, S. Bates *et al.*, "Conformal prediction: A gentle introduction," *Foundations and trends® in machine learning*, vol. 16, no. 4, pp. 494–591, 2023.

[17] P. Radoglou-Grammatikis, P. S. Bouzinis, I. Makris, T. Lagkas, V. Argyriou, G. T. Papadopoulos, P. Fouliras, G. Seritan, and P. Sarigiannidis, "Ai4fids: Multimodal federated intrusion detection," *IEEE Transactions on Emerging Topics in Computing*, 2025.

[18] S. Vajapeyam, "Understanding Shannon's entropy metric for information," *arXiv preprint arXiv:1405.2061*, 2014.