HERMES: Healthcare Ethics & Robustness in Medical Image Systems

Enhancing Robustness and Ethical Integrity of Image Classifiers Against Adversarial Attacks

Matteo Fasulo[1], Luca Babboni[1] and Maksim Omelchenko[1]

[1]*Department of Computer Science and Engineering (DISI) - University of Bologna*

Abstract

Deep Neural Networks have shown exceptional promise in medical image analysis, yet their vulnerability to adversarial attacks poses a critical threat to patient safety and clinical trust. This paper presents a systematic investigation into the robustness of medical image classifiers, a cornerstone of Trustworthy AI as defined by the European Commission's guidelines. We operate under a realistic gray-box threat model, fine-tuning a ResNet-18 model on the PatchCamelyon dataset for metastatic tissue detection and evaluating its resilience against evasion attacks: the Fast Gradient Sign Method and Projected Gradient Descent. To counter these threats, we implement and benchmark a suite of input preprocessing defenses: Gaussian smoothing, JPEG compression, spatial smoothing, and total variance minimization. Our analysis rigorously assesses each defense's effectiveness, the impact of standard data augmentation, and the inherent trade-off between adversarial robustness and diagnostic accuracy. By providing a comparative analysis of attack efficacy and defense performance, this work offers crucial insights into the technical and ethical challenges of developing secure, reliable, and ethically sound AI systems for clinical deployment. The code for reproducing the results is available at `https://github.com/MatteoFasulo/HERMES`.

Keywords

Adversarial Robustness, Medical Image Analysis, Adversarial Defenses, Trustworthy AI, AI Ethics

1. Introduction

Deep Neural Networks (DNNs) have emerged as a transformative technology in medical image analysis, demonstrating capabilities that often rival or surpass human experts [1]. This breakthrough embodies the principle of *beneficence*, holding immense potential to enhance healthcare delivery. However, DNNs exhibit a critical vulnerability to adversarial attacks: subtle, often imperceptible perturbations to input data designed to elicit misclassifications. In the high-stakes domain of medical diagnostics, this vulnerability directly threatens the principle of *non-maleficence*, as an incorrect prediction can lead to severe patient harm. This fragility undermines the reliability of AI systems and erodes trust, posing a significant barrier to their widespread adoption.

According to the European Commission's *Ethics Guidelines for Trustworthy AI*, a trustworthy system must be lawful, ethical, and **robust**—both technically and socially [2]. Our work investigates this crucial pillar of robustness through the lens of evasion attacks. We operate under the **gray-box threat model**, a realistic scenario where an adversary possesses knowledge of the target model's architecture and parameters but remains unaware of any defense mechanisms [3]. This model represents a credible threat, as model details may be public or become accessible through leaks or open-source implementations.

In this paper, we conduct a systematic investigation into the robustness of a ResNet-18 classifier fine-tuned on the PatchCamelyon dataset [4] for metastatic tissue detection. Building on prior work that has demonstrated the potential of input transformations as a defense [5], we evaluate its vulnerability to two canonical attacks, the Fast Gradient Sign Method (FGSM) [6] and Projected Gradient Descent (PGD) [7].

matteo.fasulo@studio.unibo.it (M. Fasulo); luca.babboni2@studio.unibo.it (L. Babboni); maksim.omelchenko@studio.unibo.it (M. Omelchenko)

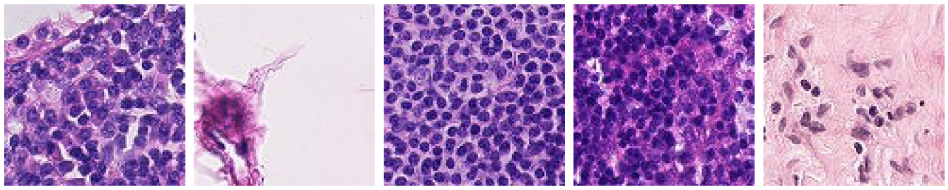
https://github.com/MatteoFasulo (M. Fasulo); https://github.com/ElektroDuck (L. Babboni); https://github.com/omemaxim (M. Omelchenko)

0000-0002-7019-3157 (M. Fasulo); 0009-0001-5260-7467 (L. Babboni)

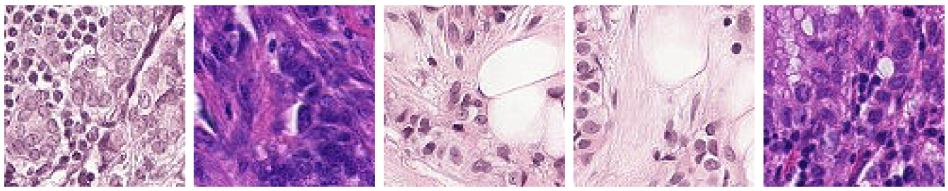
We then benchmark a suite of input preprocessing defenses: Gaussian smoothing, JPEG compression, spatial smoothing, and total variance minimization. Our evaluation focuses on the critical trade-off between preserving accuracy on benign data and ensuring resilience against adversarial samples. This work aims to provide foundational insights into the technical challenges and ethical imperatives for developing secure, trustworthy, and clinically viable AI systems.

2. Dataset

The PatchCamelyon dataset [4] is a collection of 327,680 histopathologic scans of lymph node sections (96×96 pixels). It is designed for binary classification: identifying whether a patch contains metastatic tissue. Its clinical relevance, balanced classes, and manageable scale make it an excellent benchmark for this study. Representative samples are shown in Figure 1.



(a) Negative samples (healthy tissue).



(b) Positive samples (metastatic tissue).

Figure 1: Representative samples from the PatchCamelyon dataset [4].

3. Model and Training

We employ a ResNet-18 architecture [8], which offers a strong balance between model capacity and computational efficiency. To isolate the effect of standard training practices on adversarial robustness, we train two separate models: one with no data augmentation, and one with a suite of common augmentations (random horizontal flips, rotations, and color jitter). By comparing these two variants, we can directly assess the extent to which standard augmentation improves resilience to adversarial perturbations, as similarly investigated by Guo et al. [5].

For fine-tuning, we adopt the linear-probing-then-fine-tuning (LP-FT) protocol recommended by Kumar et al. [9], which mitigates catastrophic forgetting of pretrained features. Specifically, we first train only the new classification head (linear probing) while freezing the ResNet backbone, and then unfreeze the entire network for end-to-end fine-tuning. Kumar report that LP-FT achieves roughly a 1% improvement on in-distribution accuracy and a 10% improvement on out-of-distribution generalization compared to standard fine-tuning. We adopt this two-stage LP-FT strategy because the ResNet-18 weights are pretrained on a domain substantially different from our target dataset.

4. Adversarial Threat Model

To rigorously assess our defenses, we employ two canonical gradient-based attacks under a gray-box scenario. The choice of these two attacks is deliberate: FGSM serves as a rapid but naive baseline, while PGD represents a much stronger, iterative adversary that provides a more realistic estimate of a model's vulnerability [3]. A diagram of the attack scenario is depicted in Figure 2.

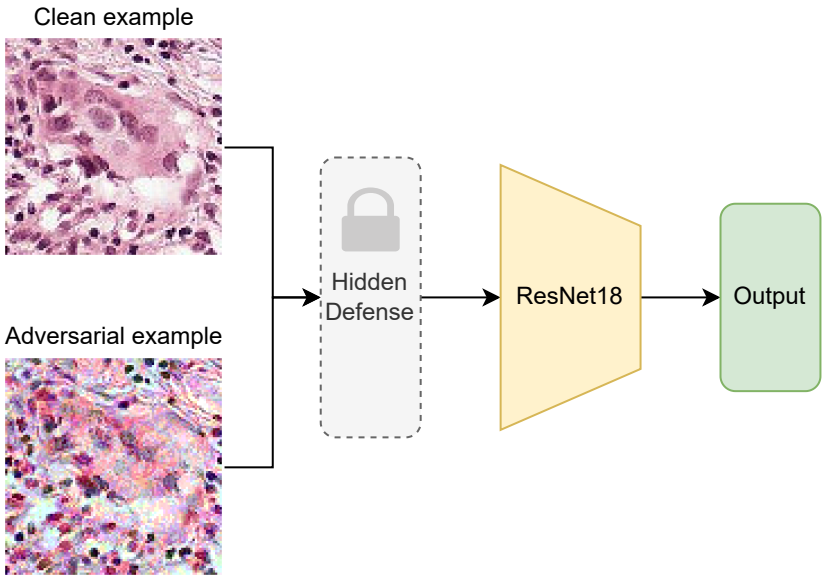


Figure 2: A diagram of the gray-box attack pipeline. The attacker crafts an adversarial example using knowledge of the model but is unaware of the defense layer that preprocesses the input before classification.

4.1. Fast Gradient Sign Method (FGSM)

FGSM is a single-step attack that computes the adversarial example x^{adv} by adding a perturbation along the direction of the sign of the loss gradient [6]:

$$x^{\text{adv}} = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

where ε controls the perturbation magnitude. While computationally efficient, FGSM was never intended as a strong benchmark for robustness and can give a false sense of security [3]. We include it as a baseline to illustrate how easily simple attacks can be thwarted, in contrast to more sophisticated methods.

4.2. Projected Gradient Descent (PGD)

PGD is a stronger, iterative attack that is widely considered the gold standard for evaluating adversarial robustness [7]. It takes multiple small steps, projecting the result back into an ε-ball around the original image after each step to ensure the perturbation remains constrained: Let $\mathcal{B}_\varepsilon(x)$ be the ε-ball around x. The update rule for each iteration t is given by:

$$x_{t+1} = \Pi_{\mathcal{B}_\varepsilon(x)} \left(x_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_t, y)) \right)$$

where α is the step size. This iterative process allows PGD to find more effective adversarial examples by exploring the loss landscape more thoroughly, providing a much more challenging and realistic test for any defense mechanism.

5. Comparative Evaluation of Defenses

We benchmarked four common input preprocessing defenses: Gaussian Smoothing, Spatial Smoothing (Median Filtering), Total Variation Minimization (TVM), and JPEG Compression. This section synthesizes the results to provide a comparative analysis of their effectiveness. We chose the Area Under the Receiver Operating Characteristic (AUROC) curve [10] as our primary evaluation metric. AUROC is the de-facto standard for evaluating binary classifiers in the medical field because it is threshold-independent. It measures a model's ability to distinguish between positive and negative classes across all possible classification thresholds, providing a comprehensive assessment of diagnostic utility that is not tied to a single, arbitrary cutoff point. Visual examples of each defense are included in Appendix A, and detailed performance curves are in Appendix B.

5.1. Defense Efficiency Against Adversarial Attacks

Our analysis reveals a stark difference in defense performance against FGSM and PGD. Against the single-step FGSM attack, many defenses exhibited a "V-shaped" performance curve, appearing surprisingly effective against stronger perturbations. This occurs because large-ϵ FGSM creates coarse, high-frequency noise that low-pass filters easily remove, a phenomenon that can create a false sense of security.

This resilience vanished against the stronger, iterative PGD attack. For nearly all defenses, performance degraded monotonically and catastrophically as PGD's perturbation budget increased. This underscores PGD's ability to bypass simple, non-adaptive input transformations and highlights the danger of evaluating defenses only against weak attacks.

As shown in Table 1, **Total Variation Minimization (TVM)** offered the most significant robustness, particularly when combined with data augmentation. It maintained a non-trivial AUROC of 0.505 against the strongest PGD attack. In contrast, other defenses like Gaussian Smoothing and JPEG Compression proved almost entirely ineffective against strong PGD, with their AUROC scores collapsing to near-zero, rendering them unreliable in a robust security model.

Table 1
Summary of defense performance (AUROC) on the augmented model. We report performance on benign (clean) images, against a weak PGD attack ($\epsilon = 0.01$), and a strong PGD attack ($\epsilon = 0.1$).

Defense Method	Clean Data	PGD ($\epsilon = 0.01$)	PGD ($\epsilon = 0.1$)
None (Baseline)	0.940	0.129	0.000
Gaussian Smoothing (k=3, σ=0.8)	0.881	0.622	0.001
JPEG Compression (q=50)	0.925	0.732	0.002
Spatial Smoothing (w=5)	0.758	0.656	0.221
Total Variance Min. (p=0.3)	**0.793**	**0.734**	**0.505**

5.2. The Robustness-Accuracy Trade-off

A critical aspect of any practical defense is its impact on model performance with benign inputs. Our results (Table 2) show a clear and often severe trade-off between adversarial robustness and clean-data accuracy. This represents a direct tension between the ethical principles of *beneficence* (achieving high diagnostic accuracy) and *non-maleficence* (preventing harm from misdiagnosis).

Defenses that were most effective against attacks, such as **TVM** and **Spatial Smoothing**, incurred the highest accuracy cost on clean images, reducing the AUROC from a baseline of 0.940 to 0.793 and 0.758, respectively. This degradation is caused by the filters removing fine-grained, discriminative features along with adversarial noise. Conversely, defenses with minimal impact on clean data, like high-quality JPEG Compression, offered almost no protection. This fundamental trade-off presents a major challenge for clinical deployment, where both high accuracy and robustness are non-negotiable.

Defense Method	No Data Augmentation	Data Augmentation
None (Baseline)	0.925	0.940
Gaussian Smoothing (k=3, σ=0.8)	0.860	0.890
JPEG Compression (q=50)	0.892	0.925
Spatial Smoothing (w=5)	0.676	0.758
Total Variance Min. (p=0.3)	0.552	0.793

Table 2
Comparison of model performance (AUROC) on clean data. The table compares a baseline model (no defense strategy applied) against models equipped with various preprocessing defenses, both with and without data augmentation, to illustrate the inherent accuracy cost of each defensive strategy.

5.3. The Role of Data Augmentation

We analyzed the impact of standard data augmentation by comparing the performance of models trained with and without it. The results, summarized in Figure 3, show a nuanced picture.
Against the simpler FGSM attack, data augmentation was almost universally beneficial. However, against PGD, the effect was mixed. While augmentation significantly boosted the most effective defense (TVM), it was detrimental when combined with weaker defenses like Gaussian Smoothing against moderate PGD attacks. This suggests that augmentation may inadvertently create new vulnerabilities that a strong iterative attack can exploit. Furthermore, it only marginally improved the model's intrinsic robustness, indicating that standard augmentation is not a sufficient defense on its own.

5.4. Computational Overhead of Defenses

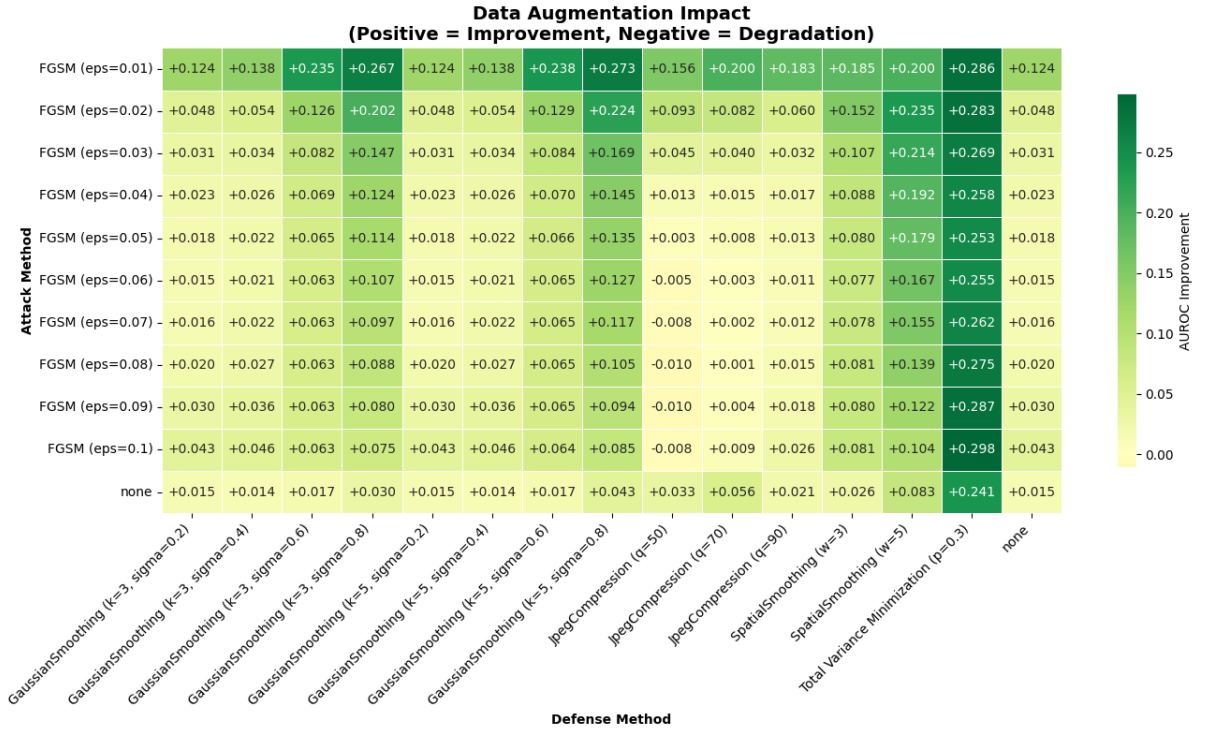
For a defense to be practical in a clinical workflow, its computational overhead must be acceptable. **Gaussian Smoothing**, **Spatial Smoothing**, and **JPEG Compression** are all highly efficient, adding negligible latency. **Total Variation Minimization (TVM)**, as an iterative optimization algorithm, is by far the most computationally demanding. However, for the 96×96 images in our study, its execution time remains within a practical budget for non-urgent, batch-processing workflows. Therefore, all evaluated defenses are computationally feasible, shifting the primary selection criterion to the balance between defensive strength and its impact on diagnostic accuracy.

6. Ethical and Trustworthy Implications

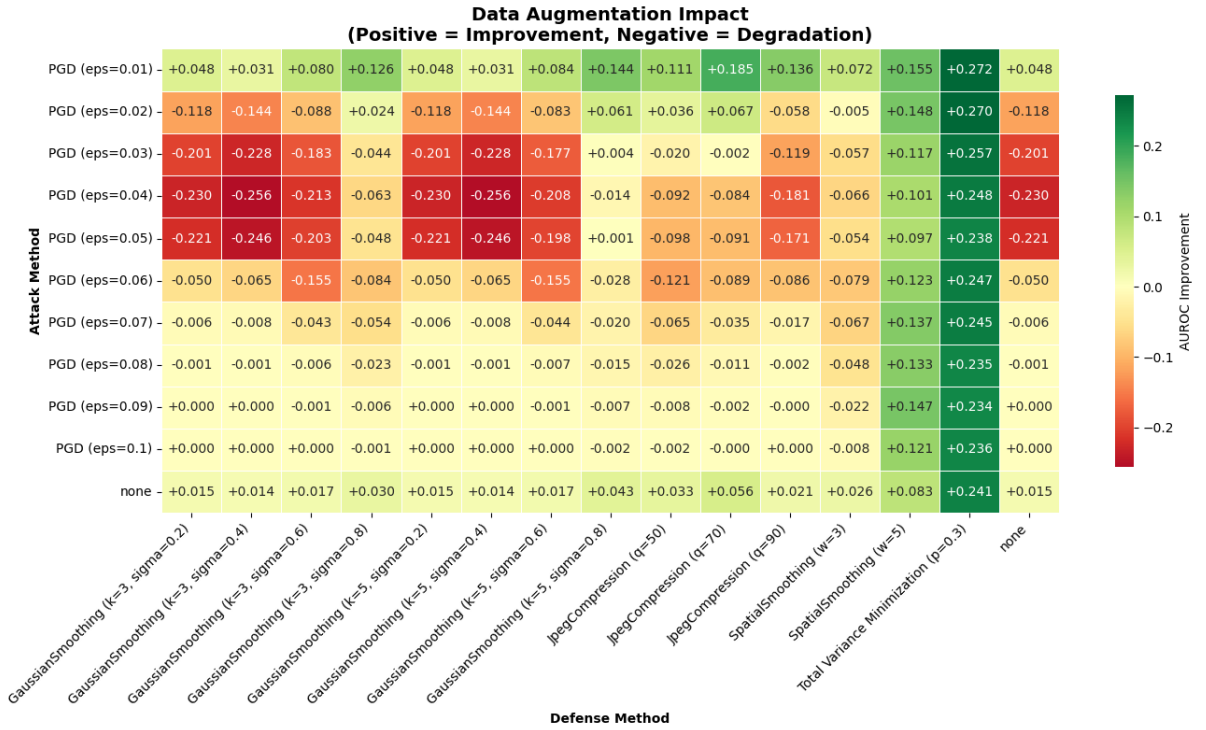
The technical vulnerabilities and defense trade-offs we observed have profound ethical implications for deploying AI in healthcare. These challenges directly map to the core requirements for Trustworthy AI as outlined by the European Commission [2] and the principles synthesized by initiatives like AI4People [11].

Technical Robustness and Safety. The EU guidelines identify this as a key requirement. Our findings demonstrate that deploying a medical AI with simple, non-adaptive defenses would be a clear failure to meet this standard. The catastrophic performance drop against PGD shows these systems are not "resilient and secure." Relying on such defenses creates a dangerous illusion of safety, violating the core duty to prevent harm.

Human Agency and Oversight. A system vulnerable to manipulation undermines the clinician's agency. As AI4People argues, AI should enhance human capabilities [11]. If an AI's output can be maliciously flipped without any visible trace, it ceases to be a reliable tool. This forces clinicians into an untenable position: either blindly trust a potentially compromised system or abandon a technology that could otherwise provide significant benefits. Effective robustness is a prerequisite for meaningful human-in-the-loop oversight.



(a) Impact on robustness against FGSM attacks.



(b) Impact on robustness against PGD attacks.

Figure 3: Data Augmentation Impact Heatmaps. Values represent AUROC (Augmented Model) - AUROC (Non-Augmented Model). Green indicates improved robustness with augmentation; red indicates degradation.

Privacy and Data Governance. While our focus is on evasion attacks, the underlying lack of robustness is a security flaw. A system that is not secure against input manipulations is unlikely to be secure against other threats. Vulnerabilities that allow for evasion attacks could potentially be exploited for privacy-violating attacks like model inversion or membership inference, which aim to reconstruct

training data or identify individuals within a dataset. Therefore, ensuring adversarial robustness is an integral part of a comprehensive data protection and privacy strategy.
Ultimately, the accuracy-robustness trade-off is not merely a technical dilemma but an **ethical balancing act**. Deciding how much diagnostic accuracy can be sacrificed for a given level of safety requires a multi-stakeholder dialogue involving clinicians, developers, patients, and ethicists. It is a question that cannot be answered by algorithms alone.

7. Conclusion

This paper presented a systematic investigation into the adversarial robustness of a medical image classifier, benchmarking common preprocessing defenses under a gray-box threat model. Our findings confirm the significant vulnerability of DNNs in this domain, particularly to strong iterative attacks like PGD, which consistently degraded model performance to a level unsuitable for clinical use.
Among the evaluated defenses, Total Variation Minimization (TVM), when combined with data augmentation, demonstrated the most effective resilience. This robustness, however, was achieved at the cost of a substantial drop in performance on benign data, highlighting a critical trade-off between security and nominal accuracy. This trade-off is not just a technical issue but an ethical one, forcing a difficult balance between the principles of beneficence and non-maleficence.
Our work underscores that simple, non-adaptive preprocessing defenses are insufficient to guarantee the reliability and safety required for Trustworthy AI in medicine. The pronounced vulnerabilities and trade-offs observed pose a significant challenge, demanding the development of more integrated and fundamentally robust solutions. Fulfilling the promise of AI in healthcare requires moving beyond simple accuracy metrics to build systems that are demonstrably secure, reliable, and ethically aligned with the core mission of patient care.

8. Future Work

Building upon the findings and limitations of this study, we identify several crucial directions for future research:

- **Evaluation against Adaptive Attacks:** The most critical next step is to evaluate these defenses against *adaptive attacks*, where an adversary is aware of the defense and crafts perturbations to bypass it. Such attacks are known to defeat many simple input transformation defenses [12], and this evaluation is essential for a realistic security assessment.
- **Integration of Adversarial Training:** Future work should compare and combine preprocessing defenses with **adversarial training** [7]. Investigating whether these input transformations can complement adversarial training to achieve higher robustness without sacrificing as much clean-image accuracy is a promising research avenue.
- **Exploration of Certified and Detection-Based Defenses:** We recommend expanding the scope to include **certified defenses**, which offer formal robustness guarantees, and **detection-based methods** that aim to identify and reject adversarial inputs before classification.
- **Generalization to Diverse Modalities and Architectures:** To ensure broader applicability, this analysis should be replicated across different **medical imaging modalities** (e.g., MRI, CT scans) and **network architectures** (e.g., Vision Transformers).

Team Contributions

- **Matteo Fasulo:** Development of the testing framework, ensuring seamless integration and compliance with the Adversarial Robustness Toolbox (ART) by the Linux Foundation AI & Data Foundation [13]. Designed and implemented the overall testing pipeline, focusing on modularity and extensibility.

- **Luca Babboni:** Carried out the initial fine-tuning of the ResNet-18 model, with and without data augmentation. Contributed actively to the development of both the testing framework and the defense techniques. Primary developer of the visualization tools used to demonstrate the effects of adversarial attacks and the efficacy of defenses.
- **Maksim Omelchenko:** Conducted in-depth research on the theoretical foundations of adversarial attack and defense methodologies. Contributed to the design of the defense strategy and played a key role in the composition and editing of the final project report.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI-GPT-4 in order to improve grammar and spelling. After using this service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

[2] E. Commission, C. Directorate-General for Communications Networks, Technology, G. ekspertów wysokiego szczebla ds. sztucznej inteligencji, Ethics guidelines for trustworthy AI, Publications Office, 2019. doi:doi/10.2759/346720.

[3] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, A. Kurakin, On Evaluating Adversarial Robustness, 2019. URL: http://arxiv.org/abs/1902.06705. doi:10.48550/arXiv.1902.06705, arXiv:1902.06705 [cs].

[4] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling, Rotation equivariant CNNs for digital pathology (2018). arXiv:1806.03962.

[5] C. Guo, M. Rana, M. Cisse, L. van der Maaten, Countering adversarial images using input transformations, 2018. URL: https://arxiv.org/abs/1711.00117. arXiv:1711.00117.

[6] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2015. URL: https://arxiv.org/abs/1412.6572. arXiv:1412.6572.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2019. URL: https://arxiv.org/abs/1706.06083. arXiv:1706.06083.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[9] A. Kumar, A. Raghunathan, R. Jones, T. Ma, P. Liang, Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022. URL: https://arxiv.org/abs/2202.10054. arXiv:2202.10054.

[10] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 143 (1982) 29–36.

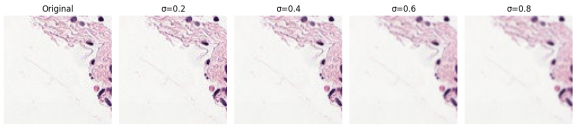
[11] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, Minds and Machines 28 (2018) 689–707. URL: https://doi.org/10.1007/s11023-018-9482-5. doi:10.1007/s11023-018-9482-5.

[12] A. Athalye, N. Carlini, D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, in: Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR, 2018.

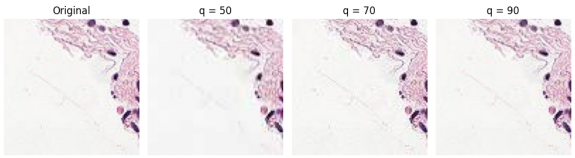
[13] M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, B. Edwards, Adversarial robustness toolbox v1.2.0, CoRR 1807.01069 (2018). URL: https://arxiv.org/pdf/1807.01069.

A. Defense Visualizations

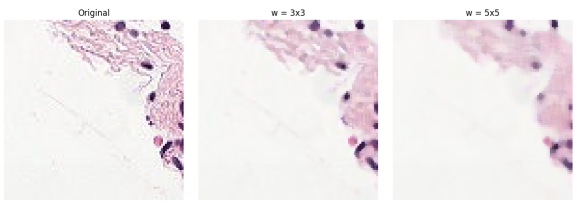
This appendix provides visual examples of the effect of each preprocessing defense on a sample image from the PatchCamelyon dataset.



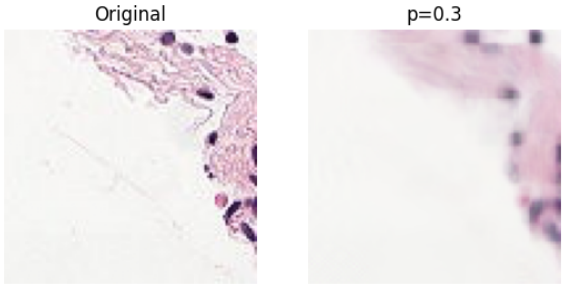
(a) Gaussian Smoothing with varying σ.



(b) JPEG Compression with varying quality q.



(c) Spatial Smoothing with varying window size w.



(d) Total Variation Minimization.

Figure 4: Visual effects of the evaluated preprocessing defenses.

B. Detailed Performance Curves

This appendix contains the detailed performance plots (AUROC vs. attack strength ϵ) for each defense method, comparing models trained with and without data augmentation.

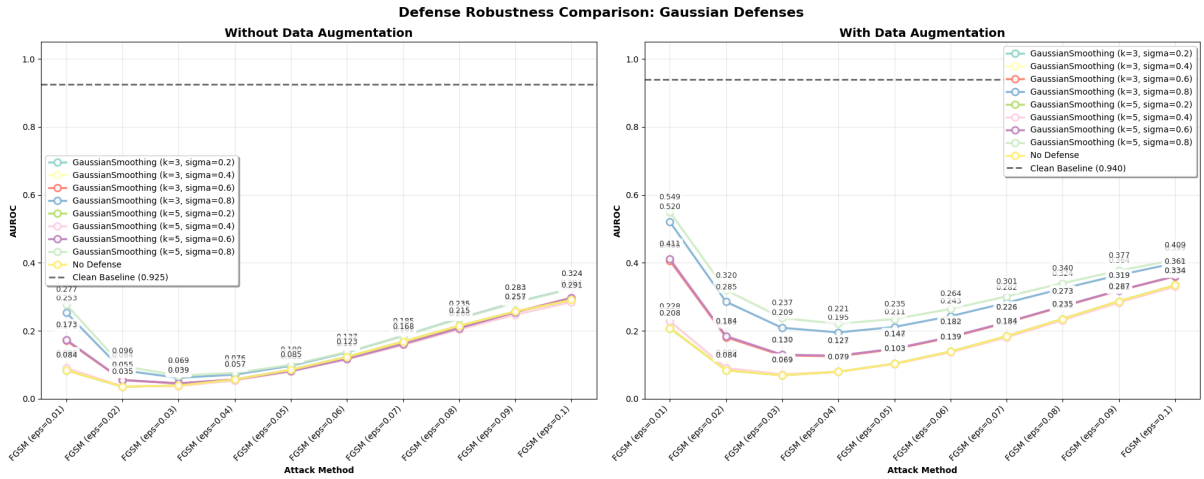


Figure 5: Performance of **Gaussian Smoothing** against FGSM attacks.

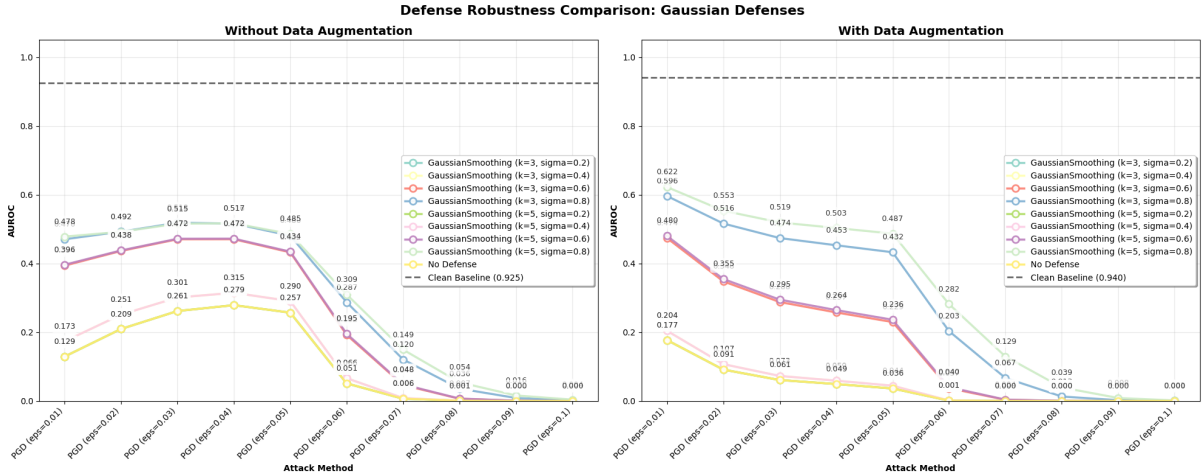


Figure 6: Performance of **Gaussian Smoothing** against PGD attacks.

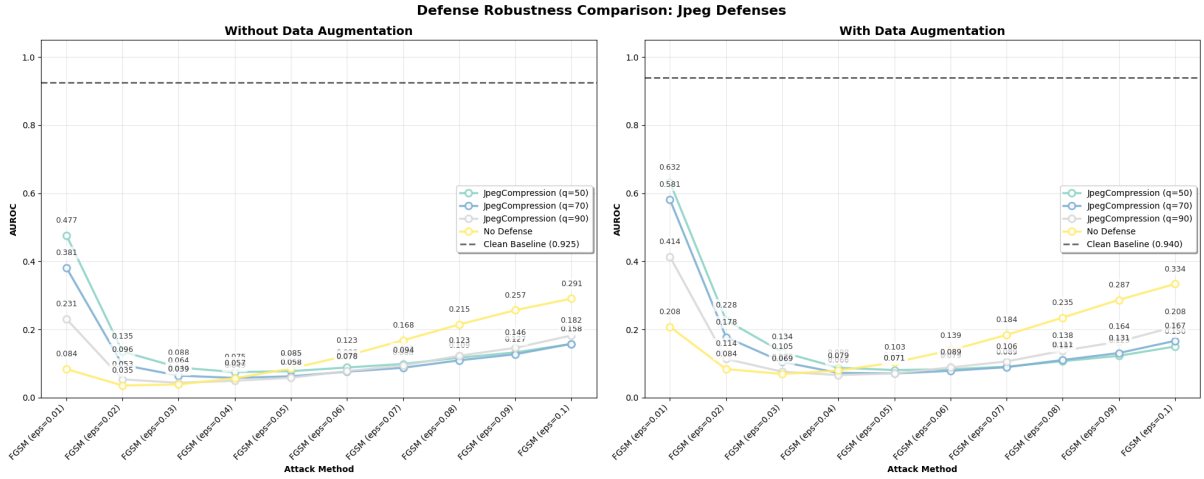


Figure 7: Performance of **JPEG Compression** against FGSM attacks.

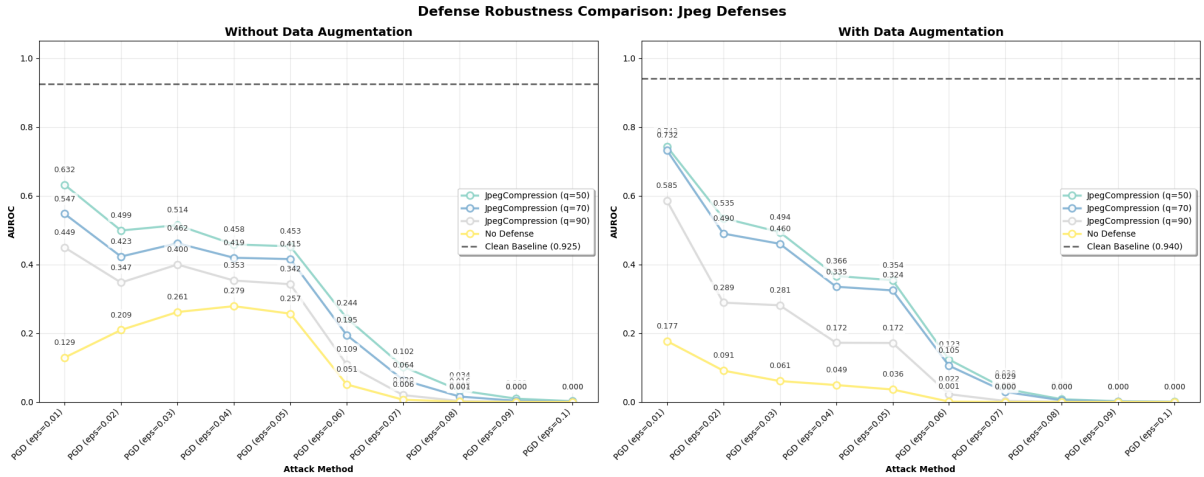


Figure 8: Performance of **JPEG Compression** against PGD attacks.

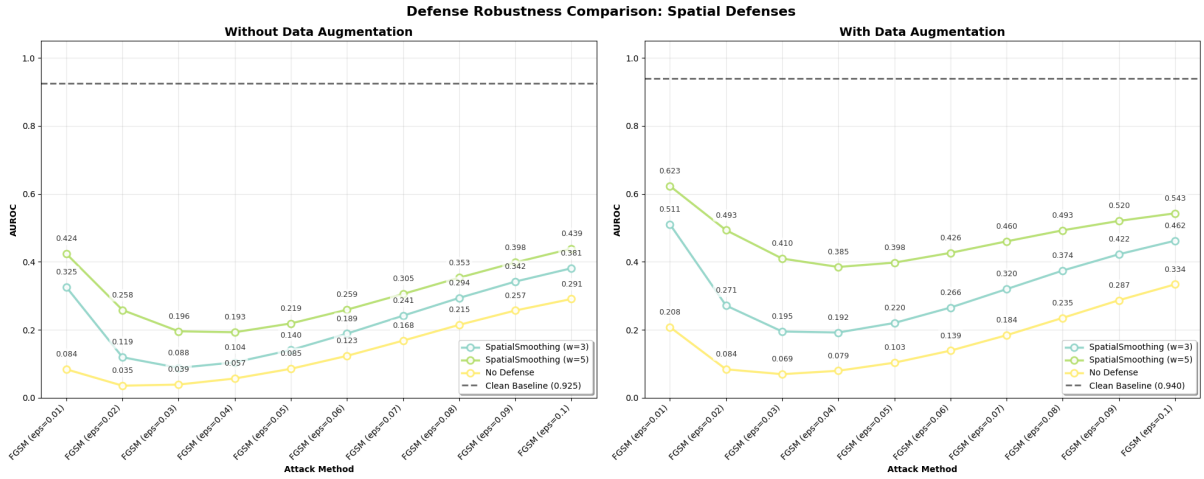


Figure 9: Performance of **Spatial Smoothing** against FGSM attacks.

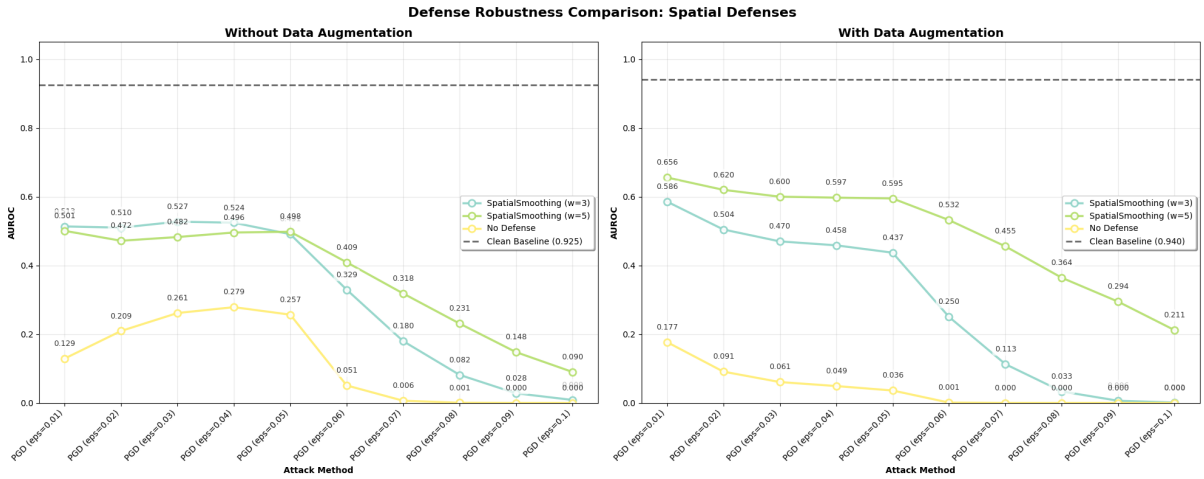


Figure 10: Performance of **Spatial Smoothing** against PGD attacks.

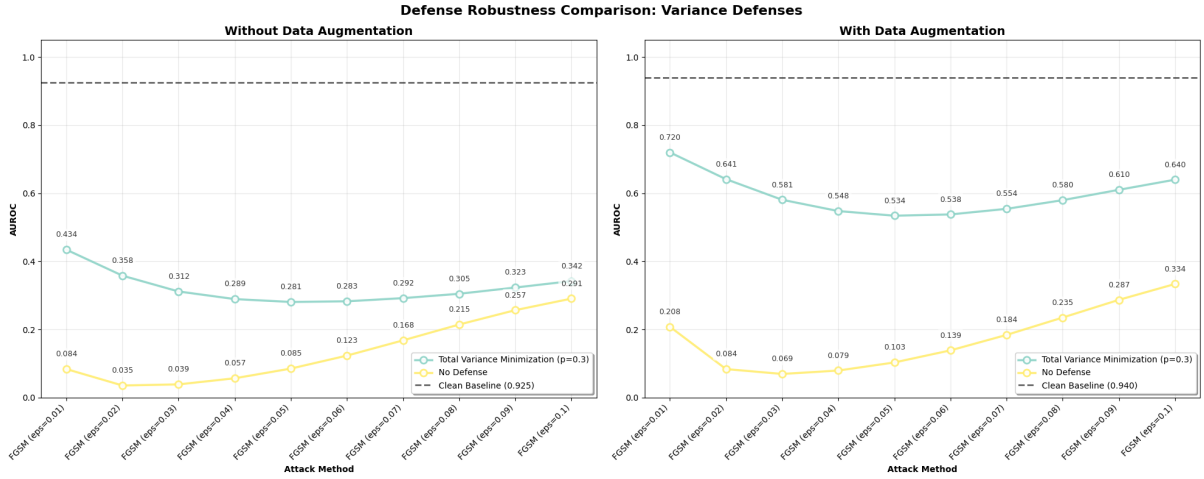


Figure 11: Performance of **Total Variation Minimization** against FGSM attacks.

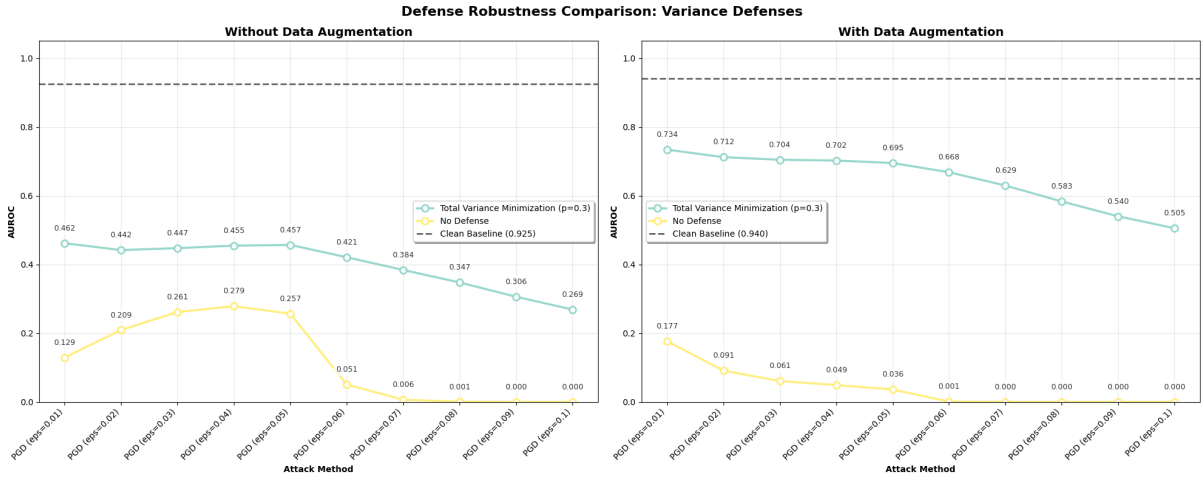


Figure 12: Performance of **Total Variation Minimization** against PGD attacks.