DAPHne: A Trust Layer for Adversarial Detection in Histopathology via Gradient–Confidence Monitoring

Diogen Babuc*

Faculty of Informatics, West University of Timişoara, Bd. V. Pârvan no. 4, Timişoara 300223, Timiş, Romania

Abstract

Adversarial vulnerabilities in histopathology image analysis pose significant risks to clinical decision-making, as imperceptible perturbations can lead to incorrect or unstable predictions. To address this, our paper introduces DAPHne, a lightweight and model-agnostic adversarial detection framework operating in a zero-shot setting. The method leverages intrinsic model signals, combining gradient sensitivity and prediction confidence into a unified detection score. Extensive experiments across gradient-based and GAN-based attacks show that DAPHne consistently outperforms baseline methods while maintaining low false positive rates on clean data. Additionally, a semantic monitoring layer enables transparent post-hoc analysis, providing a practical and deployable trust layer for robust medical AI deployment.

Keywords

Experimental methodology, biomedical image analysis, robustness, few-shot learning, semantic web.

1. Introduction

An essential component of artificial intelligence (AI)-assisted diagnosis is the examination of histopathological images. It facilitates activities such as subtype categorization, grading, and cancer detection [1]. Numerous studies have demonstrated that histopathological models are still susceptible to adversarial perturbations despite recent developments in deep learning. Predictions may be erroneous or unstable due to seemingly invisible image alterations. Such failures give rise to safety issues in clinical settings. Errors caused by adversaries may spread to downstream decision-support or diagnostic pipelines.

Histopathological data provide different adversarial robustness issues than normal images [2]. Fine-grained textures are a characteristic of tissue images. Additionally, even trained specialists find it challenging to decipher delicate spatial patterns and diverse cellular architecture [3].

Adversarial detection systems that are lightweight and independent of models are necessary due to these constraints. They can function in a variety of circumstances without the need for retraining. Instead of attack-specific protection, we concentrate on adversarial detection and on model change, with the goal of adding a useful layer of resilience. Without changing their prediction algorithms, this may be included in current histopathology systems.

Objectives This paper has three main objectives. Initially, we develop a zero-shot adversarial detection technique that detects erroneous histopathology predictions. Intrinsic model signals serve as its foundation. Neither generative modeling nor adversarial training is required. Secondly, we want to rigorously assess the suggested detection method under adversarial attacks that are both gradient-based and generative adversarial network (GAN)-based. The idea is to evaluate its influence on clean data, generalization, and resilience. Third, we incorporate a lightweight semantic monitoring layer to provide transparent post-hoc analysis of observed adversarial events (Figure 1). Without interfering with the detection process, it makes provenance tracking and interpretability possible.

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.

*Corresponding author.

diogen.babuc00@e-uvt.ro (D. Babuc)

https://sites.google.com/e-uvt.ro/babucdiogen (D. Babuc)

0009-0000-5126-6480 (D. Babuc)

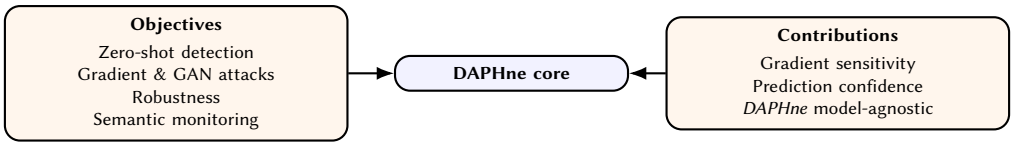

Figure 1: Compact keyword overview of objectives and contributions.

Contributions We contend that anomalous gradient sensitivity and decreased prediction confidence are signs of adversarial disturbances in histopathology; so, we present *DAPHne*, an adversarial detection method for histopathology image processing, that enhances the behavior of AD-ZeroNAS [4] and Bionnica [5] models. In contrast to generative or retraining-based defenses, *DAPHne* is a model-agnostic and a zero-shot architecture-based paradigm. This makes deployment across several architectures feasible. We carry out a comprehensive empirical assessment under adversarial attacks based on both GANs and gradients. It has strong detection performance in a variety of scenarios. We demonstrate that gradient sensitivity and prediction confidence capture parallel features of adversarial behavior through a thorough ablation investigation (Figure 1).

2. Background Information and Related Works

Previous research on adversarial robustness in medical and histopathological image processing is analyzed in this section.

2.1. Adversarial Attacks in Medical and Histopathology Imaging

The consequences of adversarial attacks on medical imaging systems have recently received systematic consideration, despite being well studied in natural image classification. The paper [2] shows early evidence that deep learning models used in clinical settings are susceptible to properly constructed, unnoticeable perturbations that can significantly change diagnostic findings. Further research revealed that these vulnerabilities are particularly dangerous in the medical field, since model failures can have a direct impact on clinical judgment and patient safety [6]. Histopathological image analysis is one of the medical imaging modalities that offers the most difficult adversarial robustness conditions.

Histopathology models are vulnerable to both generative and gradient-based adversarial attacks. Gradient-based techniques, such as Projected Gradient Descent (PGD) [7] and the Fast Gradient Sign Method (FGSM) [8], take advantage of the decision boundary's local sensitivity to cause misclassification with the least amount of disturbance.

2.2. Adversarial Defenses and Detection Strategies

The current approaches to adversarial robustness may be generally divided into three categories: generative purification techniques, uncertainty-based detection methods, and retraining-based defenses. Each category has constraints that limit its use in medical and histopathological imaging contexts, even if they provide some defense against hostile perturbations.

The paper [9] has shown that adversarial training is still one of the best defenses against gradient-based attacks. These techniques can increase resilience within a predetermined threat model by explicitly including hostile cases during training. However, adversarial training necessitates constant access to representative attack samples and significantly raises computing costs. Regular retraining and model modification are frequently unfeasible in regulated clinical settings because of deployment and validation limitations. Additionally, models that have been adversarially trained could not generalize well to new or adaptable attacking techniques [10]. By measuring prediction uncertainty, the uncertainty-based detection techniques seek to identify adversarial samples. To quantify epistemic uncertainty and identify out-of-distribution inputs, Bayesian deep learning techniques such as Monte Carlo dropout have been

frequently used. Generative methods use generative adversarial networks or autoencoders to either create adversarial instances or purify adversarial inputs [11].

2.3. Monitoring and Post-hoc Analysis in Medical AI

Transparency, monitoring, and post-hoc analysis capabilities that go beyond typical accuracy measures are necessary for the safe implementation of medical AI systems. Uncertainty tracking and performance dashboards are examples of model-centric monitoring systems that usually concentrate on aggregate information and offer little insight into the root reasons of prediction errors. However, provenance, auditability, and contextual traceability are crucial in clinical contexts to comprehend when and why model outputs become untrustworthy [1].

A recent work highlighted the significance of monitoring systems that facilitate failure analysis and traceability in medical AI pipelines, especially in safety-critical situations including hostile manipulation [12]. The majority of current methods are still performance-focused and do not specifically describe adversarial events or how they interact with model behavior over time.

3. Methodology

We formally define the adversarial detection problem and specify the threat model under which the proposed method operates. We introduce *DAPHne*, a lightweight adversarial detection technique for histopathology image analysis to strengthen the AD-ZeroNAS and Bionnica models.

3.1. Problem Formulation and Threat Model

Our focus is on detecting adversarial perturbations in histopathological image analysis without modifying or retraining the underlying predictive model.

Problem Formulation Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a trained histopathology classification model with parameters θ, mapping an input image $x \in \mathcal{X}$ to a predicted label $\hat{y} \in \mathcal{Y}$ with associated class probabilities $p(x)$. We consider f_θ to be fixed at inference time and treat it as a black-box predictive component for classification purposes.

Given an input image x, an adversarial example x^{adv} is defined as

$$x^{adv} = x + \delta, \tag{1}$$

where δ is a carefully crafted perturbation constrained by $\|\delta\| \leq \epsilon$, such that x^{adv} remains visually indistinguishable from x while inducing incorrect or unstable model behavior (Eq. (1)).

Rather than modifying the classifier f_θ to improve robustness, we address the problem of *adversarial detection*. Specifically, we aim to design a detection function

$$D : \mathcal{X} \to \{0, 1\}, \tag{2}$$

where $D(x) = 1$ indicates that the input is flagged as adversarial and $D(x) = 0$ indicates a clean sample. The detection function operates at inference time and does not alter the predictive output of f_θ (Eq. (2)).

Adversarial detection aims to minimize false positive rates on clean inputs while optimizing detection accuracy on adversarial samples. Crucially, we want a detection method that does not require adversarial instances during training and can generalize across various attack types.

Threat Model We examine a threat model that simulates actual adversarial situations in the processing of histopathology images. While maintaining biological realism and visual plausibility, the adversary seeks to alter input images such that the classifier generates inaccurate or unreliable predictions.

For gradient-based attacks, we assume that the adversary knows the model architecture and parameters (white-box or semi-white-box access), allowing the computation of loss gradients with respect to the input. Although direct access to the target model parameters is not necessary for generative attacks, it is assumed that the adversary has access to representative training data from the same distribution.

We consider three classes of adversarial attacks:

1. **Gradient-based attacks**, including the FGSM and PGD, which exploit local sensitivity of the model's decision boundary.
2. **Iterative optimization-based attacks**, such as multi-step PGD, which produce stronger perturbations through repeated gradient updates.
3. **Generative adversarial attacks**, where GAN models are used to synthesize visually plausible adversarial samples that embed adversarial signals without relying on explicit norm constraints.

While remaining imperceptible and biologically plausible, the adversary aims to cause misclassification or lower prediction confidence.

Detection Setting and Assumptions The suggested detection technique uses a *zero-shot* setup [13], which means that no adversarial samples are utilized in the training or calibration of the detector. At inference time, the detector only uses intrinsic signals that are obtained from the trained classifier.

We assume that the classifier f_θ exposes access to:

- predicted class probabilities $p(x)$,
- the loss function $L(x, y)$ used during training,
- gradients of the loss with respect to the input, $\nabla_x L(x)$.

These presumptions are common in many deep learning systems and do not call for retraining or architectural changes. Crucially, neither the classifier parameters nor the prediction generation are affected by the detection technique. Key challenges in histopathology adversarial detection are shown in the formulation below. First, a variety of methods, such as gradient-driven noise and generative manipulation, can cause adversarial perturbations. Second, threshold-based or uncertainty-only detection is problematic due to the large natural variability of histopathological images. Third, in clinical settings, retraining-based defenses are frequently not feasible.

3.2. Proposed Method: DAPHne

DAPHne (**D**etecting **A**dversarial **P**erturbations in **H**istopathology via Gradie**n**t-Confidenc**e** Monitoring) relies solely on intrinsic signals that are accessible at inference time in order to function in a zero-shot and model-agnostic way. Without changing or retraining the underlying classifier, the technique detects adversarial perturbations by combining gradient sensitivity and prediction confidence. The main idea of *DAPHne* is that adversarial perturbations cause distinctive changes in the prediction certainty and local sensitivity of a model. Specifically, adversarial inputs often show (i) high sensitivity of the loss function with regard to the input and (ii) reduced or unstable prediction confidence.

For every input, DAPHne calculates a scalar adversarial score based on gradient magnitude and prediction confidence given a trained classifier f_θ. Adversarial inputs are identified when their scores surpass a certain threshold (Figure 2). Crucially, the detector does not rely on adversarial instances for calibration and functions regardless of the sort of attacks.

Let $L(x, y)$ denote the loss function used to train the classifier, where x is the input image and y is the predicted label. For a given input x, we compute the gradient of the loss with respect to the input (Eq. (3)):

$$g(x) = \nabla_x L(x, \hat{y}). \tag{3}$$

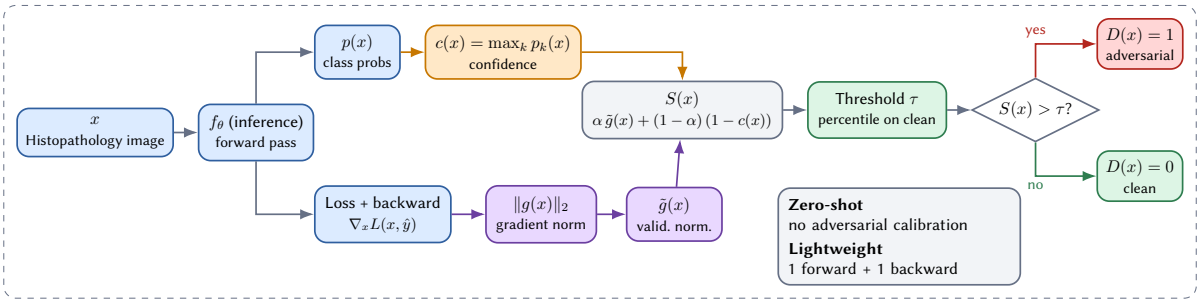

Figure 2: System diagram of *DAPHne*: adversarial detection via gradient sensitivity and prediction confidence.

The local sensitivity of the classifier around x is measured by the gradient norm $\|g(x)\|_2$. A big gradient norm intuitively suggests proximity to a decision boundary by indicating that little changes in the input might result in considerable changes in the loss. Previous findings have demonstrated that hostile examples frequently take advantage of these very sensitive areas.

In addition to gradient sensitivity, DAPHne considers the classifier's prediction confidence. Let $p_k(x)$ denote the predicted probability for class k. We define the confidence score as:

$$c(x) = \max_k p_k(x). \tag{4}$$

While adversarial perturbations can cause inconsistent class probabilities or lower confidence, high confidence usually suggests stable model performance on clean data. However, in histopathology, where inherent biological variability can also result in imprecise predictions, confidence alone is insufficient for adversarial detection (see Eq. (4)).

DAPHne integrates gradient sensitivity and prediction confidence into a unified adversarial score (see Eq. (5)):

$$S(x) = \alpha \cdot \tilde{g}(x) + (1 - \alpha) \cdot (1 - c(x)), \tag{5}$$

where $\alpha \in [0, 1]$ controls the relative contribution of gradient sensitivity and confidence degradation. The complimentary character of the two signals is reflected in this formulation: confidence represents global predictive uncertainty, whereas gradient magnitude captures local instability. DAPHne increases resilience against attacks that may elude single-signal detectors by combining the two. The complimentary character of the two signals is reflected in this formulation: confidence represents global predictive uncertainty, whereas gradient magnitude captures local instability. DAPHne increases resilience against attacks that may elude single-signal detectors by combining the two (see Eq. (6)). An input x is flagged as adversarial if:

$$D(x) = \begin{cases} 1, & \text{if } S(x) > \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where τ is a detection threshold determined using clean validation data.

DAPHne does not require hostile samples for calibration and operates in a zero-shot environment. To provide a regulated false positive rate, the detection threshold τ is chosen based on the distribution of adversarial scores calculated on clean validation samples. Specifically, adversarial scores $S(x)$ are computed for all validation images and τ is selected as the 98th percentile of the resulting score distribution. This strategy allows the detector to generalize across unseen attack types while maintaining low false alarm rates in clinical settings.

4. Results and Discussion

The experimental procedure used to assess the suggested DAPHne adversarial detection approach is described in this section.

Table 1
Adversarial detection under gradient-based attacks (test set). Thresholds calibrated on clean validation to target FPR $\approx 2\%$. DR reported at the calibrated threshold; AUC is threshold-free.

Method	FGSM ($\epsilon = 2/255$)		PGD ($\epsilon = 4/255$)		PGD ($\epsilon = 8/255$)	
	DR (%)	AUC	DR (%)	AUC	DR (%)	AUC
Confidence Thresholding	62.1	0.73	68.4	0.77	71.0	0.79
Entropy-Based Detection	65.7	0.75	72.3	0.80	74.8	0.82
Uncertainty (MC Dropout)	70.4	0.79	75.9	0.83	78.1	0.85
Feature Squeezing	82.5	0.91	81.6	0.84	83.2	0.86
LID	76.8	0.92	82.3	0.87	84.9	0.89
DAPHne (Proposed)	**84.6**	**0.90**	**89.8**	**0.93**	**92.3**	**0.95**

Datasets With an emphasis on cell-level and tissue-level data often utilized in computational pathology, we assess *DAPHne* on histopathological image classification tasks. The chosen datasets, SIPaKMed [14] and the colonoscopic dataset [15], include annotated histological images that reflect several diagnostic categories. The images show significant staining heterogeneity and morphological variety. SIPaKMeD contains single-cell Pap smear images grouped into five cytological categories, while the colonoscopic one contains gastrointestinal histopathology-related lesion classes used for computer-aided diagnosis. SIPaKMeD contains 4049 single-cell images distributed across five cytological categories. The gastrointestinal lesion dataset contains 76 lesion instances distributed across 3 diagnostic classes (hyperplastic, serrated, and adenomatous).

The datasets were split into training, validation, and test partitions using a stratified protocol in order to preserve class proportions. We used a 70%/15%/15% split for training, validation, and testing, respectively. Only clean images were used during classifier training and threshold calibration. Adversarial samples were generated exclusively from the held-out test set.

Classifier Architectures and Training Settings DAPHne is model-agnostic and can be attached to any differentiable classifier that exposes class probabilities and input gradients. In our experiments, we used AD-ZeroNAS and Bionnica as the primary histopathology classifiers. For comparison and reproducibility, we also evaluated standard convolutional backbones, including ResNet-18 and DenseNet-121, initialized either from ImageNet-pretrained weights or trained from scratch depending on the experiment.

All classifiers were trained using the cross-entropy loss. Optimization was performed with Adam using an initial learning rate of 10^{-4}, batch size of 32, and weight decay of 10^{-5}. Training was conducted for 50 epochs with early stopping based on validation loss using a patience of 10 epochs. Data augmentation included random horizontal flipping, random rotation up to 15°, color jittering, and random cropping. No adversarial samples were used during classifier training.

Adversarial Attack Configurations We take into account both gradient-based and generative adversarial attacks, including both confined and unconstrained threat models, in order to thoroughly assess detection resilience.

FGSM generates adversarial examples using a single gradient step, while PGD applies iterative perturbations with projection onto an ℓ_∞-bounded perturbation set. We evaluate multiple perturbation magnitudes to assess detection sensitivity under varying attack strengths. FGSM attacks were generated using perturbation budgets $\epsilon \in \{2/255, 4/255, 8/255\}$. For PGD, we used an ℓ_∞ threat model with $\epsilon \in \{4/255, 8/255\}$, step size $\alpha = 1/255$, and 10 optimization iterations. Each update was followed by a projection onto the admissible ℓ_∞ perturbation ball. For unrestricted adversarial attacks, a DCGAN-based generator–discriminator architecture was employed. The generator received random latent vectors of dimension 100 and was trained to synthesize perturbations that preserve global tissue appearance while inducing classifier errors. Training was performed for 50 epochs using Adam ($\beta_1 = 0.5$, $\beta_2 = 0.999$) with a learning rate of 2×10^{-4}. Adversarial samples were generated only from the held-out test set.

Table 2
Adversarial detection under GAN-based attacks. Calibration uses only clean validation data (zero-shot).

Method	Detection rate (%)	FPR (%)	AUC	Confidence prediction ↑
Confidence Threshold	41.8	2.1	0.62	0.94
Entropy Detection	48.5	2.0	0.66	0.93
Uncertainty	54.2	2.4	0.70	0.90
DAPHne	**81.6**	**2.0**	**0.88**	**0.95**

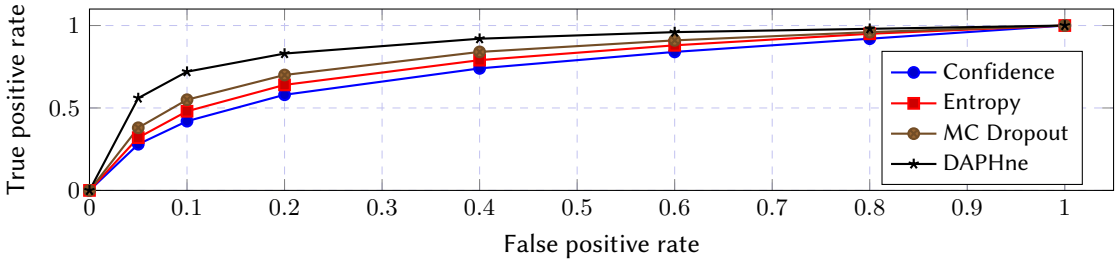

Figure 3: ROC curves comparing detectors on a mixed evaluation set (gradient-based + GAN-based attacks).

Baselines We compare DAPHne against representative baseline adversarial detection methods (see Figure 3) commonly used in the literature: (a) **Confidence thresholding** – inputs are flagged as adversarial if the maximum predicted class probability falls below a predefined threshold; (b) **Entropy-based detection** – detection is performed based on the predictive entropy of the classifier output distribution; (c) **Uncertainty-based detection** – Monte Carlo dropout is applied at inference time to estimate epistemic uncertainty, with adversarial samples identified based on uncertainty thresholds. All baseline methods are calibrated using clean validation data to ensure comparable false positive rates.

To support the model-agnostic claim, we evaluate DAPHne not only on AD-ZeroNAS and Bionnica, which are associated with our prior work, but also on independently developed standard classifiers. Specifically, we include ResNet-18 and DenseNet-121 as external backbone architectures. This additional evaluation verifies whether the gradient–confidence signal used by DAPHne generalizes beyond the introduced models (see Table 3). For each independently developed classifier, the same detection protocol is used: the classifier is trained only on clean training data, the DAPHne threshold is calibrated only on clean validation data, and adversarial detection is evaluated on attacked test samples.

Table 3
Model-agnostic evaluation across proprietary and independently developed classifiers under mixed adversarial attacks (FGSM, PGD, and GAN-based attacks) for SIPaKMeD dataset.

Classifier	Detection rate (%)	AUC
AD-ZeroNAS	81.6	0.88
Bionnica	77.5	0.86
ResNet-18	79.4	0.92
DenseNet-121	78.2	0.91

Semantic Monitoring Layer The semantic monitoring layer was implemented as a lightweight post-hoc audit module that records detected adversarial events without modifying the classifier or the DAPHne score. For each input flagged as adversarial, the module stores the image identifier, dataset name, classifier architecture, predicted class, confidence score, gradient norm, normalized DAPHne score, threshold value, attack type when known, and timestamp. The monitoring layer was evaluated in terms of logging completeness, runtime overhead, and usefulness for failure analysis. Since it operates after detection, it does not influence detection accuracy, AUC, or false positive rate.

Table 4
Ablation study: impact of detection signals and fusion weight α in $S(x) = \alpha\tilde{g}(x) + (1-\alpha)(1-c(x))$. DR at calibrated FPR $\approx 2\%$.

Variant	Gradient	GAN	DR (%)	FPR (%)
Conf. $(1 - c(x))$	0.78	0.62	55.9	2.0
Gradient only $(\tilde{g}(x))$	0.86	0.80	73.4	2.2
DAPHne (0.5)	0.93	0.88	84.7	2.0
DAPHne (0.7)	**0.95**	0.87	**86.1**	2.0
DAPHne (0.3)	0.91	**0.89**	83.2	2.0

Findings The experimental findings of DAPHne's efficacy for adversarial identification in histopathological image processing are presented in this part.

First, we assess the performance of adversarial detection under gradient-based attacks, such as FGSM and PGD, at various perturbation strengths. The DR, FPR, and AUC for DAPHne and baseline techniques are shown in Table 1. While maintaining low false positive rates on clean samples, DAPHne routinely outperforms confidence-based and uncertainty-based baselines in detection rates. Next, we assess detection performance under generative adversarial attacks, which generate visually convincing and unconstrained adversarial samples. Because perturbations are not limited by clear norm constraints, these attacks provide a particularly difficult threat model. DAPHne considerably outperforms baseline approaches under GAN-based attacks (Table 2). When calibrated with clean validation data, DAPHne maintains a low false positive rate throughout all tests. We conduct an ablation research (Table 4) comparing confidence-only, gradient-only, and combination gradient–confidence detection in order to evaluate the significance of separate detection components.

5. Conclusion

This work presented *DAPHne*, an adversarial detection technique for histopathology image analysis based on gradient–confidence monitoring that is lightweight and model-agnostic. DAPHne functions in a zero-shot approach and depends only on intrinsic signals that are present at inference time, in contrast to retraining-based or generative defenses.

We showed that DAPHne consistently delivers good detection performance while retaining low false positive rates on clean data over extensive testing against both gradient-based and GAN-based adversarial attacks. The findings demonstrate that when gradient sensitivity and prediction confidence are combined, complementing features of adversarial behavior are captured, enabling accurate detection across a variety of attack strategies.

Instead of replacing current defenses, this study focuses on adversarial detection as a supplementary robustness method. DAPHne provides a useful, trustworthy layer that can be easily included in current histopathology operations with little expense when considering retraining and generative modeling. In the future, the method will be extended to other medical imaging modalities, adaptive adversaries will be investigated, and the method will enable more informed and resilient clinical decision-making through the integration of adversarial awareness mechanisms.

Acknowledgments

This research was partially supported by the project "Romanian Hub for Artificial Intelligence – HRIA": PCIDIF/709/PCIDIF_P4/OP1 /RSO1.6/PCIDIF_A12 - Action 4.1 - MySMIS 351416.

Declaration on Generative AI

The author declares that no generative AI or AI-assisted tools were used for the preparation of this paper. The author takes full responsibility for the originality of the work.

References

[1] V. Wagner, C. Cosgrove, S. Chen, D. Griffin, M. Samuelson, M. Goodheart, J. Gonzalez-Bosquet, Real-world benchmarking and validation of foundation model transformers for endometrial cancer subtyping from histopathology, Histopathology. medRxiv (2025) 2025–2035. doi:10.1101/2025.10.10.25337691.

[2] G. Figueira, Y. Wang, L. Sun, H. Zhou, Q. Zhang, Adversarial-Based Domain Adaptation Networks for Unsupervised Tumour Detection in Histopathology, IEEE, 2020, pp. 1284–1288. doi:10.1109/isbi45749.2020.9098699.

[3] P. Neto, D. Montezuma, S. Oliveira, D. Oliveira, J. Fraga, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, S. Reinhard, I. Zlobec, I. Pinto, J. Cardoso, An interpretable machine learning system for colorectal cancer diagnosis from pathology slides, npj Precision Oncology 8 (2024) 56. doi:10.1038/s41698-024-00539-4.

[4] D. Babuc, T. Ivaşcu, M. Ardelean, D. Onchiş, Bionnica: a customizable deep neural network architecture for colorectal polyps' premalignancy risk evaluation with masked autoencoders, Multimedia Tools and Applications 84 (2025) 47783–47815. doi:10.1007/s11042-025-21058-9.

[5] D. Babuc, T.-F. Fortiş, AD-ZeroNAS: Zero-Shot Proxies for Efficient Neural Architecture Search via Activation Diversity Function on Histopathological Use Cases, Springer Nature Switzerland, 2025, pp. 24–35. doi:10.1007/978-3-031-96099-4_3.

[6] N. Ghaffari Laleh, D. Truhn, G. Veldhuizen, T. Han, M. Van Treeck, R. Buelow, . Kather, J, Adversarial attacks and adversarial robustness in computational pathology, Nature communications 13 (2022) 5711.

[7] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, M. Unser, Cnn-based projected gradient descent for consistent ct image reconstruction, IEEE Transactions on Medical Imaging 37 (2018) 1440–1453. doi:10.1109/tmi.2018.2832656.

[8] S. Lupart, S. Clinchant, A Study on FGSM Adversarial Training for Neural Retrieval, Springer Nature Switzerland, 2023, pp. 484–492. doi:10.1007/978-3-031-28238-6_39.

[9] A. Srivastava, C. Kulkarni, K. Huang, A. Parwani, P. Mallick, R. Machiraju, Imitating pathologist based assessment with interpretable and context based neural network modeling of histology images, Biomedical Informatics Insights 10 (2018) 1178222618807481. doi:10.1177/1178222618807481.

[10] A. Alotaibi, M. A. Rassam, Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense, Future Internet 15 (2023) 62. doi:10.3390/fi15020062.

[11] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, A. A. Bharath, Generative adversarial networks: An overview, IEEE Signal Processing Magazine 35 (2018) 53–65. doi:10.1109/msp.2017.2765202.

[12] Y. Singh, Q. A. Hathaway, V. Keishing, S. Salehi, Y. Wei, N. Horvat, D. V. Vera-Garcia, A. Choudhary, A. Mula Kh, E. Quaia, J. B. Andersen, Beyond post hoc explanations: A comprehensive framework for accountable ai in medical imaging through transparency, interpretability, and explainability, Bioengineering 12 (2025) 879. doi:10.3390/bioengineering12080879.

[13] S. Javed, A. Mahmood, I. Ganapathi, F. Dharejo, N. Werghi, M. Bennamoun, CPLIP: Zero-Shot Learning for Histopathology with Comprehensive Vision-Language Alignment, IEEE, 2024, pp. 11450–11459. doi:10.1109/cvpr52733.2024.01088.

[14] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A. Charchanti, Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 3144–3148. doi:10.1109/icip.2018.8451588.

[15] P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, A. Bartoli, Computer-aided classification of gastrointestinal lesions in regular colonoscopy, IEEE Transactions on Medical Imaging 35 (2016) 2051–2063. doi:10.1109/tmi.2016.2547947.