Adversarial Attacks as a Diagnostic Tool for a Mining Detection Network in the Brazilian Legal Amazon

Leonardo Fajardo Grupioni[1], **Gabriel Stephano Santos**[1], **Paulo Sérgio Cugnasca**[1], **Felipe Valencia de Almeida**[1]

[1]Escola Politécnica da Universidade de São Paulo (USP). São Paulo, SP – Brazil

{leogrupioni,gabrielstephano,cugnasca,fvalencia}@usp.br

Abstract. *Convolutional neural networks (CNNs) applied to satellite imagery are increasingly adopted as the standard tool to detect mining in the Brazilian Legal Amazon, as their predicition start to feed environmental enforcement pipelines. We trained such a CNN and obtained an F1-Score of 0.9247 with an AUC of 0.9930. Adversarial attacks show that these numbers hide what the model learned. With the addition of random noise in the data, up to 45% of the of forest patches were missclassified as mining, while low-pass noise with the same budget and the same mean shift turns 0.5%, so the decision rests on high frequency texture. The fragility is asymmetric, since one FGSM step at 4/255 converts every forest patch into mining but evades only 5% of the real mining.*

1. Introduction

Mining in the Brazilian Legal Amazon is largely illegal and driven by organized criminal networks [Global Initiative Against Transnational Organized Crime 2023], and convolutional networks have become the standard way to detect these fronts from satellite imagery [Camalan et al. 2022]. Their outputs are starting to feed enforcement pipelines, making them an attack surface, since neural classifiers can be driven to arbitrary outputs by perturbations that a human observer would not consider meaningful [Goodfellow et al. 2015], including in remote sensing [Xu and Ghamisi 2022].

Clean metrics may hide a second problem, since a model can reach excellent accuracy by learning a shortcut that separates the training classes without capturing the phenomenon of interest, and convolutional networks are known to favor texture over shape [Geirhos et al. 2019]. Adversarial attacks can be used to perceive that, by observing how controlled perturbations affect the decision and outlining cases where simple artifacts greatly influence the model output and outlining patterns recognized by the model. To investigate these failure modes, we propose an evaluation protocol that distinguishes between them using a frequency-matched control. Our findings, together with the proposed dataset, support the hypothesis that the classifier relies on high-frequency texture and is vulnerable to asymmetry.

2. Material and Methods

We used a Sentinel-2 dataset of the Brazilian Legal Amazon built from MapBiomas Collection 10 [Grupioni et al. 2026]. It contains 2600 cloud masked RGB patches from 2024 imagery, half without mining and drawn from preserved forest, and the other 1300 stratified into ten bands of mining proportion with 130 patches each. We split each band

independently into 90 training, 20 validation and 20 test patches, so the test set holds exactly 20 patches of every band. The classifier has three blocks of 3×3 convolution with 32, 64 and 128 filters, batch normalization, ReLU, max pooling and dropout, followed by a dense layer with 256 units and a sigmoid output, trained with Adam and early stopping on the validation loss.

We evaluated seven kinds of perturbations. Random noise and low-pass noise are uniform and share the same ℓ_∞ budget and mean shift, differing only in spatial frequency, which makes their comparison a controlled test of what the model uses. Brightness applies a constant offset. Fog blends a white fractal Perlin mask and brightens the scene, while occlusion blends a dark canopy colored mask and darkens it, as a site operator could with tarpaulins. (Fast Gradient Sign Method) FGSM [Goodfellow et al. 2015] and projected gradient descent (PGD) are the white box gradient attacks.

Accuracy under attack is not interpretable when the attack biases the model towards one class, since on a positive band an attack that predicts mining everywhere scores as correct. We therefore report the two directions separately, over the samples the clean model classifies correctly, as the fraction of forest turned into false positives and the fraction of mining evaded. With 20 patches per band, the Wilson interval for 0.55 is [0.34, 0.74], so we group the bands in three levels.

3. Results and Discussion

On clean data, the model reaches an accuracy of 0.9300, a precision of 1.0000, a recall of 0.8600 and an F1-Score of 0.9247, with an AUC of 0.9930. The confusion matrix is $[[200, 0], [28, 172]]$, so it never classifies forest as mining, and 17 of the 28 false negatives fall in the two lowest bands, reproducing the known difficulty with incipient mining [Lemes Neto et al. 2026].

Table 1 shows what happens under attack, and the failure is asymmetric. FGSM at 4/255 converts every forest patch into mining while evading 5.2% of the real mining, a ratio of 19 to 1. The cheap attack against this model is not evasion but the injection of false positives, which, against an enforcement pipeline, means exhausting the analysts rather than hiding a mine.

Table 1. False positive rate on the negative class and evasion rate per band group, over the samples the clean model classifies correctly.

	FP rate	Evasion rate		
Condition	(forest)	$<$20%	20–50%	$>$50%
No attack	0.00	0.00	0.00	0.00
Random noise 4/255	0.45	0.00	0.00	0.00
Low-pass noise 4/255	0.01	0.00	0.00	0.00
Brightness $+0.05$	0.01	0.04	0.02	0.00
FGSM 4/255	1.00	0.04	0.04	0.06
PGD-20 1/255	0.94	0.70	0.58	0.36

The two noise conditions identify the cause. They carry the same budget of 4/255 and the same mean shift of 0.0001, differing only in spatial frequency. Random noise turns 45% of the forest into mining and low-pass noise turns 0.5%, with a Fisher exact

$p = 2 \times 10^{-31}$. Brightness moves the mean by 0.0500, five hundred times more than random noise, and produces a false positive rate of only 0.01, while fog moves it by 0.0910 and reaches 0.12. The decision is driven by high frequency texture and not by the mean spectral level. This is consistent with the data, since preserved forest is homogeneous and mining exposes soil, water and roads, so high frequency noise fabricates heterogeneity and the forest reads as mining.

Occlusion, the physically realizable attack, darkens the scene, produces no false positives and evades at most 0.13. Real evasion requires iterative PGD, and Figure 1 shows both directions against the budget. The evasion rate of FGSM decreases as the budget grows, because a larger step drives every patch towards the mining class.

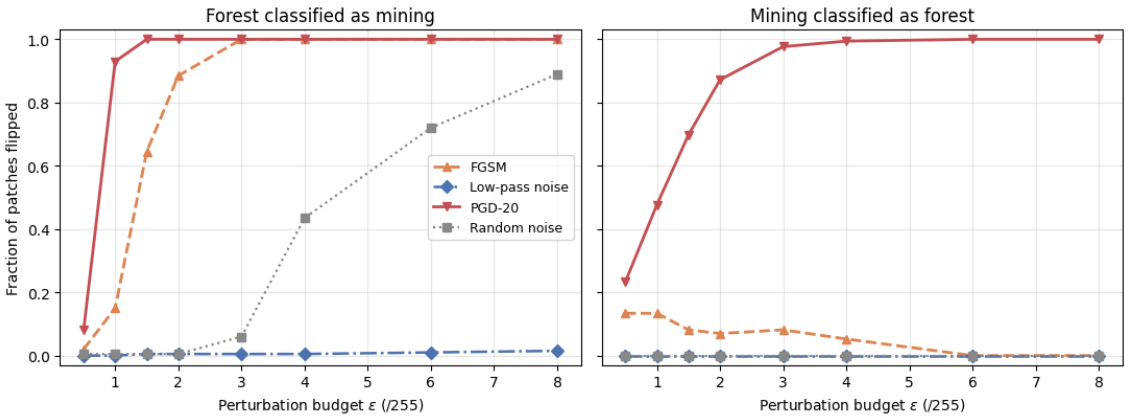

Figure 1. Fraction of patches flipped in each direction as a function of the perturbation budget, over the samples the clean model classifies correctly.

At 1/255, where PGD is only partially effective, the evasion rate follows the mining proportion. It is 0.70 below 20%, 0.58 between 20% and 50% and 0.36 above 50%, a decreasing trend with $p = 0.0027$. Evading the model is cheapest where the mining footprint is smallest, which is where detection has the highest enforcement value. From 2/255 onwards it saturates and this gradient disappears.

The main limitation is the dataset, and it is also the most useful part of the diagnosis. All 1300 negative patches come from preserved forest, so the negative class is trivially separable and the absence of false positives is an artifact of that choice. Texture energy alone separates homogeneous forest from heterogeneous mining, so the network never needed to learn the shape of a mining scar, and a boundary supported by a single property collapses under a perturbation that attacks it. The next step is therefore not a defense on this model but a redesign of the dataset. A negative class containing bare soil, shallow water, agriculture, roads and burn scars would make texture energy insufficient and force the network to use spatial structure. Only then does adversarial training become meaningful, since retraining a model that relies on a shortcut mostly hardens the shortcut.

For reproduction, seeds were fixed at 42 with deterministic cuDNN. The dataset is available on Zenodo (`https://doi.org/10.5281/zenodo.18983646`) and the notebook that reproduces every number and figure, including the visual examples of each attack across all proportion bands, at `https://github.com/leonardogrupioni/adversarial-attacks-cnn-mining-amazon`.

References

Camalan, S., Cui, K., Pauca, V. P., Alqahtani, S., Silman, M., Chan, R., Plemmons, R. J., Dethier, E. N., Fernandez, L. E., and Lutz, D. A. (2022). Change detection of Amazonian alluvial gold mining using deep learning and Sentinel-2 imagery. *Remote Sensing*, 14(7):1746.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*.

Global Initiative Against Transnational Organized Crime (2023). Amazon underworld: economias criminosas na maior floresta tropical do mundo. Technical report, GI-TOC and Amazon Watch and InfoAmazonia, Geneva, Switzerland. Institutional report.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Grupioni, L., Gallois, T. J., and Almeida, F. (2026). A sentinel-2 image dataset for mining detection across mining proportion ranges in the brazilian legal amazon. In *Anais do XVII Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 348–351, Porto Alegre, RS, Brasil. SBC.

Lemes Neto, N., Galo, M. d. L. B. T., de Oliveira, R. A., Watanabe, F. S. Y., and Galo, M. (2026). Towards SAR-Based monitoring of illegal mining in the Brazilian Amazon using convolutional neural networks. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-3-W4-2025:205–212.

Xu, Y. and Ghamisi, P. (2022). Universal adversarial examples in remote sensing: Methodology and benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15.

A. Visual Effect of the Attacks

Figure 2 shows one patch of six bands under the clean condition and under PGD at 4/255.

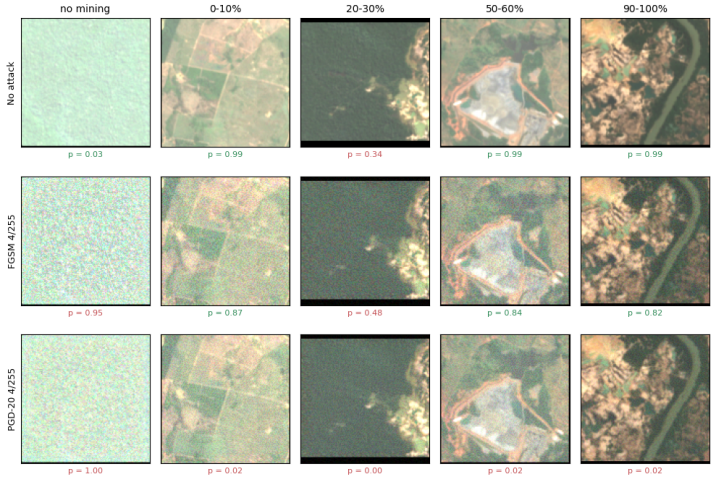

Figure 2. Representative patches under PGD at 4/255.