Multi-Layer Adversarial Detection for Ship Segmentation on NVIDIA Jetson Orin

Badr-Eddine Bouhlal *[1], Clemens-Alexander Brust [1],

[1]German Aerospace Center (DLR), Institute of Data Science
badr-eddine.bouhlal@dlr.de

I. Introduction / Motivation

Deep learning models deployed on embedded devices, such as the NVIDIA Jetson Orin AGX, are vulnerable to adversarial inputs that can compromise the reliability of critical applications like ship segmentation in satellite imagery.

Detecting these adversarial images before they reach the model is essential to ensure robust and safe inference.

We propose a hardware-aware detection approach that leverages device metrics (GPU, CPU, memory, power), image metrics, and model segmentation features to identify potential adversarial inputs in real time.

II. Adversarial Generation

We generated adversarial examples from satellite images of ships using targeted perturbations, including white-box methods such as Fast Gradient Signed Method (FGSM) and projected gradient descent (PGD), as well as black-box approaches like Gaussian blur and salt-and-pepper noise.

The perturbations were tuned to maximize disruption of the segmentation model while remaining visually imperceptible.

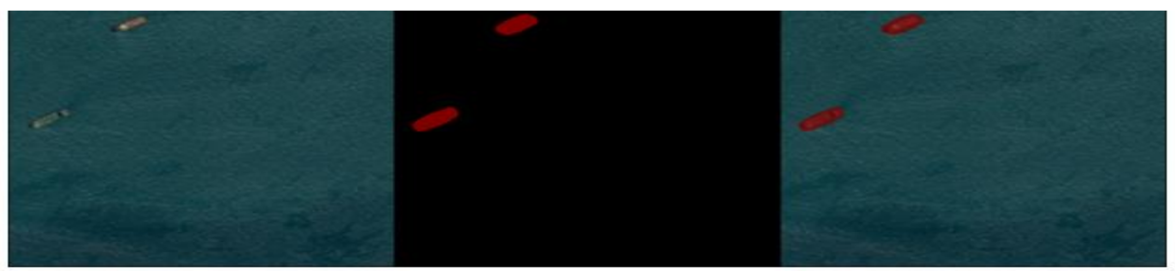


Image shows the segmentation mask produced by the deployed model on a normal satellite image

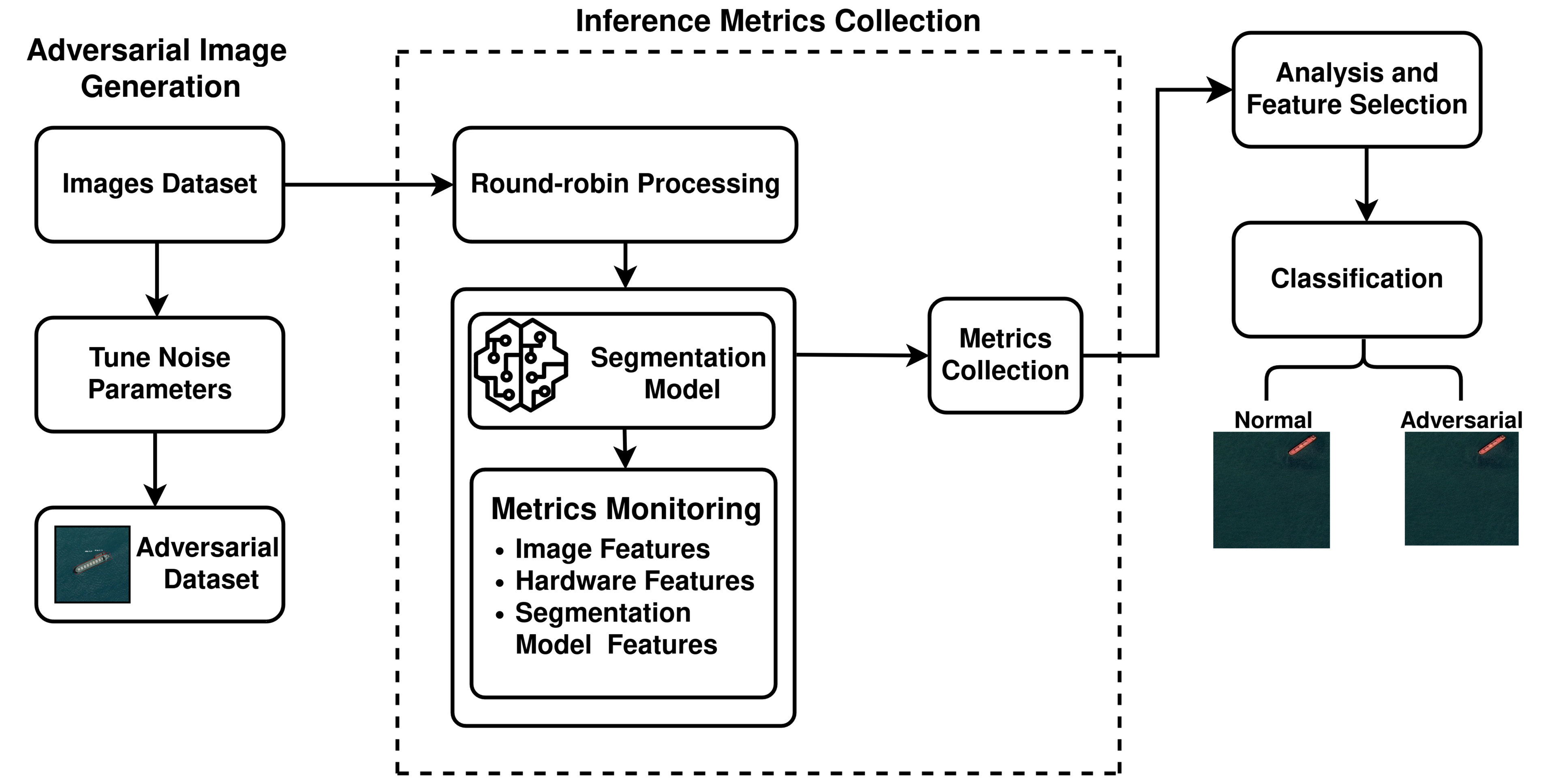
Image shows the segmentation mask produced by the deployed model on an image perturbed with Salt and pepper method with parameters (0.1, 0.5).

III. Metrics Collection

Goal: Capture multi-source metrics for each input image to detect adversarial examples .

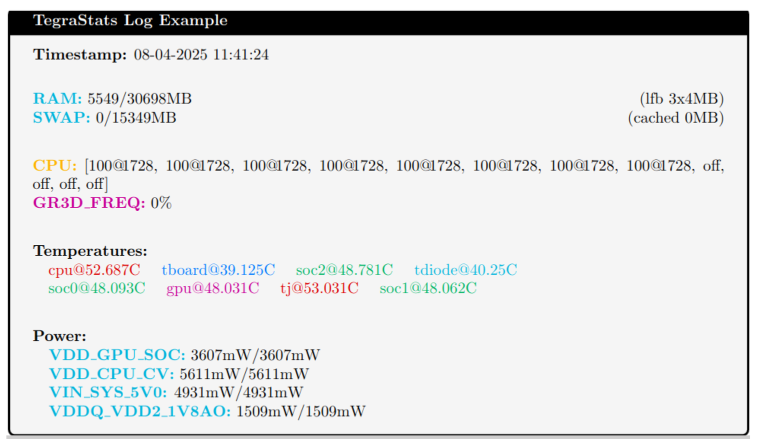
i. Metrics Overview Table (Example of Features)

Feature Type	Feature Name	Description/Units
Hardware	GPU Load	% load over inference
Hardware	GPU Temperature	°C
Image	Size	Byte
Image	Energy of high-frequency components in the image	unitless
Model	Time to transfer a tensor from GPU to CPU	Milliseconds
Model	Perform 2D convolution operation on CPU	Milliseconds



Study Workflow: From Adversarial Generation to Detection. This diagram shows the full pipeline, adversarial images are generated from satellite ship data, deployed on the Jetson Orin AGX for metric collection (hardware, segmentation, and image features), and then analyzed to train a model that detects adversarial inputs.

ii. Hardware Metrics Collection with TegraStats



TegraStats Log Example

Timestamp: 08-04-2025 11:41:24

RAM: 5549/30698MB (lfb 3x4MB)
SWAP: 0/15349MB (cached 0MB)

CPU: [100@1728, 100@1728, 100@1728, 100@1728, 100@1728, 100@1728, 100@1728, 100@1728, off, off, off, off]
GR3D_FREQ: 0%

Temperatures:
cpu@52.687C tboard@39.125C soc2@48.781C tdiode@40.25C
soc0@48.093C gpu@48.031C tj@53.031C soc1@48.062C

Power:
VDD_GPU_SOC: 3607mW/3607mW
VDD_CPU_CV: 5611mW/5611mW
VIN_SYS_5V0: 4931mW/4931mW
VDDQ_VDD2_1V8AO: 1509mW/1509mW

Tegrastats Raw logs are parsed to extract key hardware features for analysis.

Metric (Δ)	Value
GPU Power (mW)	360
RAM Used (MB)	2669

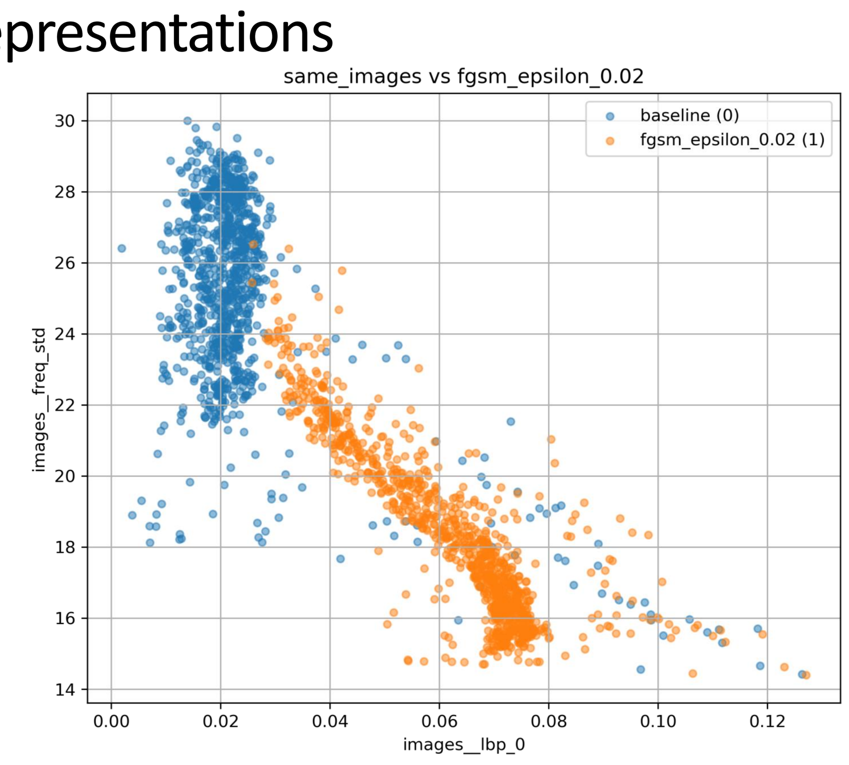
Parsed table example

IV. Feature Analysis

Top features clearly separate normal vs adversarial images for the FGSM method.

i. Key Feature Insights

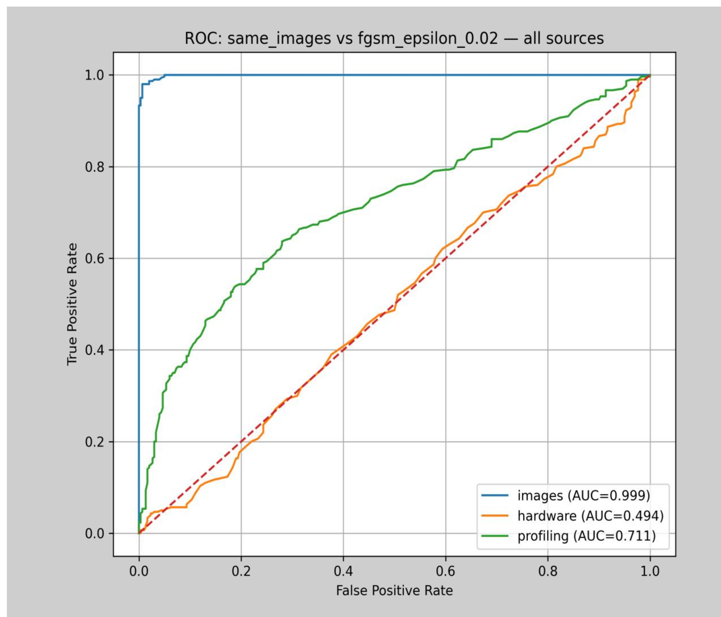
The distribution of two extracted image features reveals a clear distinction between normal images and those generated using the FGSM attack (ε = 0.02), demonstrating how adversarial perturbations shift feature representations



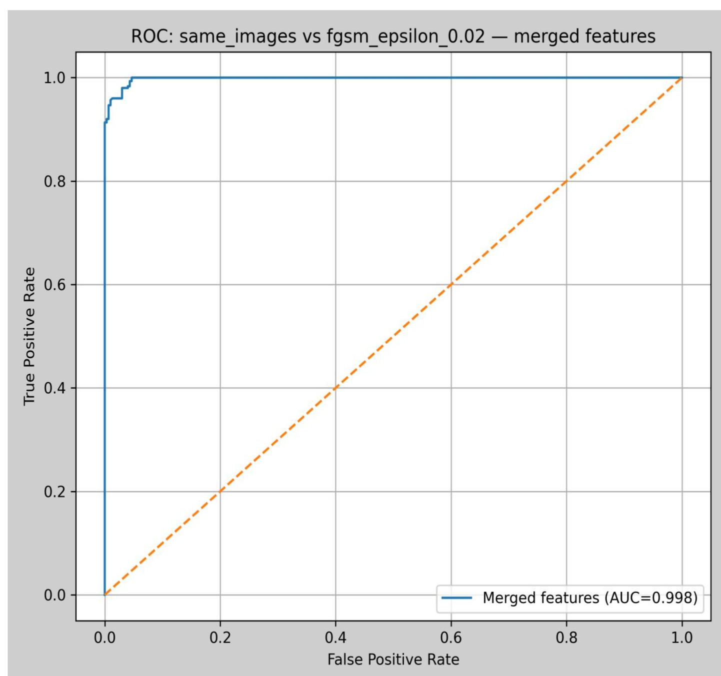
Distribution of two image features illustrating the separation between normal images (blue) and FGSM adversarial images with ε = 0.02 (orange).

ii. Feature Type Performance

The ROC curves evaluate the ability of image, hardware, and segmentation features to distinguish normal from FGSM adversarial images (ε = 0.02). Image features achieve near-perfect separation (AUC = 0.999), segmentation performs moderately (AUC = 0.711), and hardware is close to random (AUC = 0.494). Merging all features yields an AUC of 0.998.



ROC curves for individual feature sources in detecting FGSM adversarial images (ε = 0.02).



ROC curve for merged features showing improved detection performance (AUC = 0.998).

iii. Conclusion

Image features enable near-perfect adversarial detection, with segmentation offering moderate support and hardware contributing minimally. While detection can rely on image features, multi-layer monitoring remains a viable option for improved robustness.

DLR