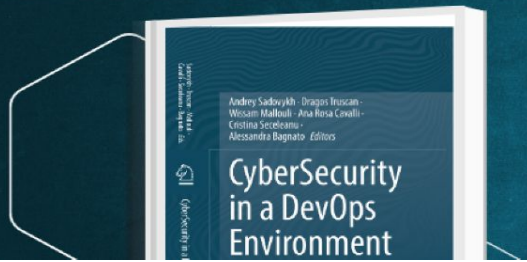


VeriDevOps Research Workshop

CyberSecurity in a DevOps Environment

Towards Anomaly Detection using Explainable AI

Manh-Dung Nguyen, Vinh-Hoa La, Wissam Mallouli,
Ana Rosa Cavalli, and Edgardo Montes de Oca

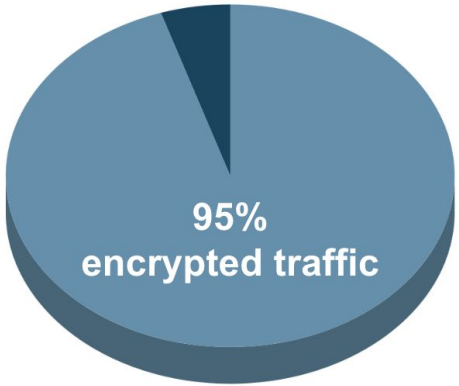
montimage

SPATIAL

Table of Content

Encryption is Changing the Attack Landscape

As of 2022



Source: State of Encrypted Attacks 2022 by Zscaler

AI-based Security Applications

montimage

Encrypted Traffic Classification

Encrypted Anomaly Detection

Root Cause Analysis

USER BEHAVIORAL ANALYSIS → ATTACK DETECTION → ATTACK RESPONSE

DATA

RESPONSIBLE AI

Performance, Accountability, Transparency, Resilience

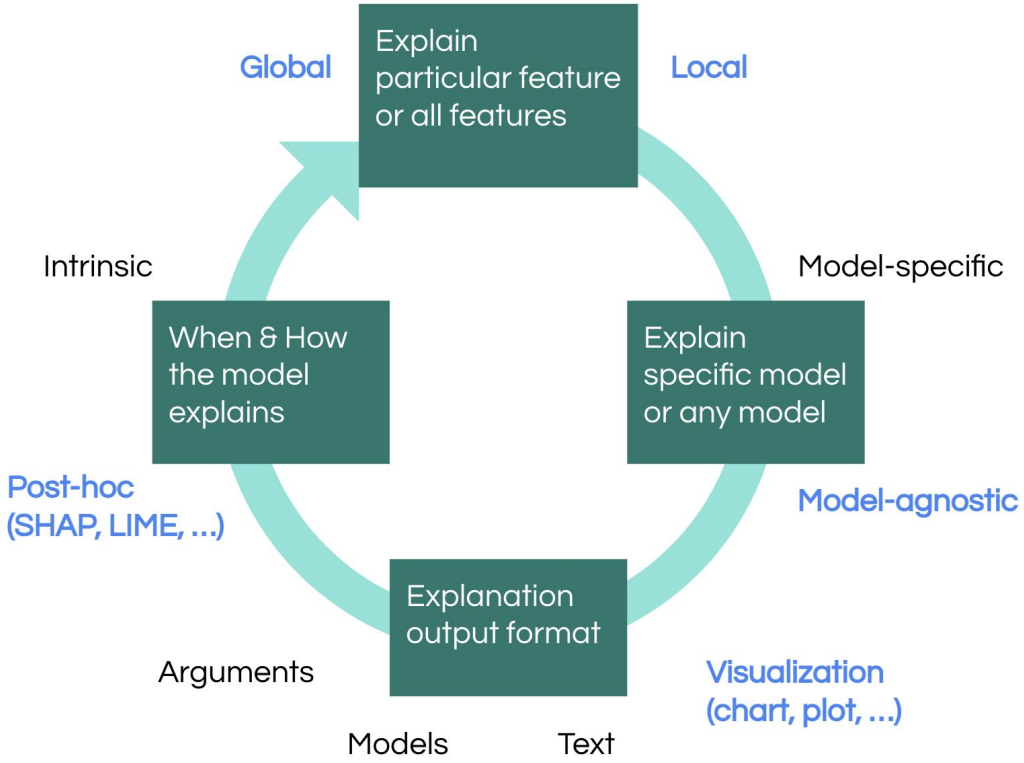
Open source projects: https://github.com/Montimage

SoTA of Explainable AI (XAI)

XAI refers to methods and techniques that provide insights into how AI models make decisions.

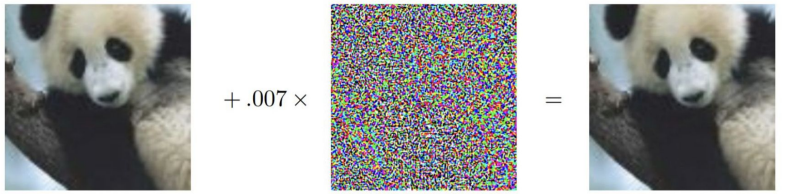
XAI enables users, developers or stakeholders to trust and effectively manage AI outcomes.

We focus on providing post-hoc local/global explanations in chart or plot using model-agnostic XAI methods.



SoTA of Adversarial Attacks

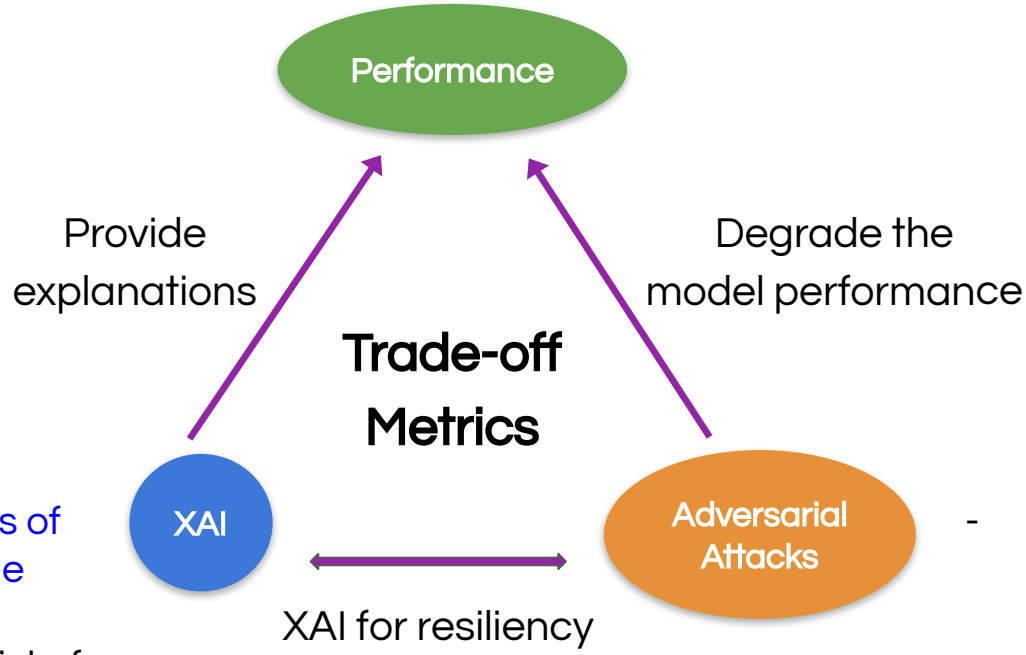
Adversarial attacks involve manipulating input data to deceive AI models, often resulting in incorrect predictions or classifications.



	Gradient-based Model M	**Transfer-based** Training Data T	**Score-based** Detailed Model Prediction Y (e.g. probabilities or logits)	**Decision-based** Final Model Prediction Y_{max} (e.g. max class label)
				less information →
Untargeted Flip to any label	FGSM, DeepFool	FGSM Transfer	Local Search	
Targeted Flip to target label	L-BFGS-B, Houdini, JSMA, Carlini & Wagner, Iterative Gradient Descent	Ensemble Transfer	ZOO	**(Boundary Attack)**

Source: Wieland et al. "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models" ICLR 2018.

XAI & Adversarial Attacks



- Produce explanations of an adversarial sample
- Prevent / Detect adversarial attacks (list of important features has changed or based on SHAP signatures)

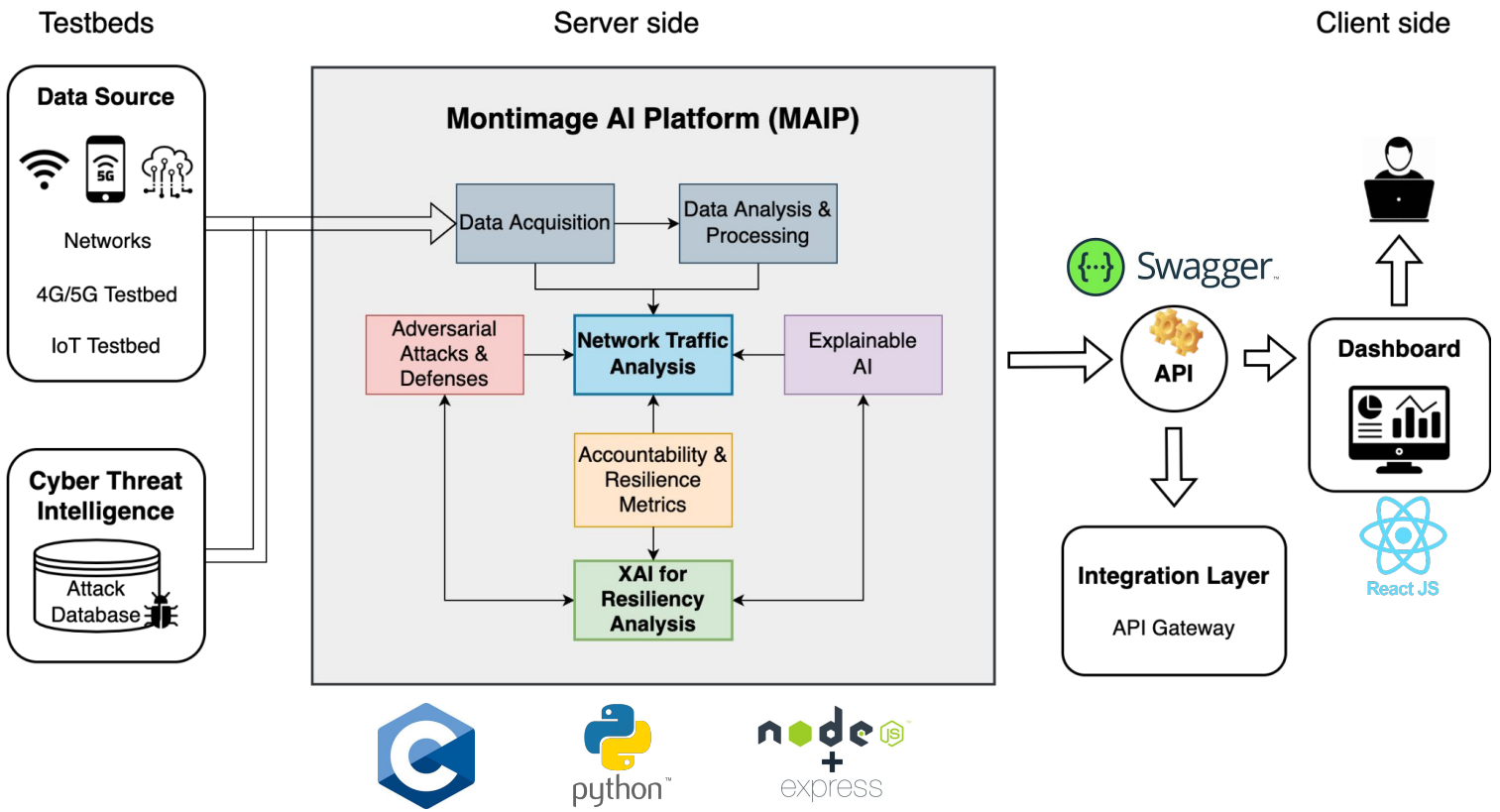
- Fool XAI methods to generate innocuous explanations
- Defend mechanisms against XAI-based attacks

Montimage AI Platform (MAIP)

Framework	Open-source	Documentation	XAI	Attacks & Defenses	Metrics	
					Accountability	Resilience
Shapash	✓	✓	✓	✗	✓	✗
explainerdashboard	✓	✓	✓	✗	✓	✗
DataRobot	✗	✓	✓	✗	✓	✗
MAIP	✓	✓	✓	✓	✓	✓

- **User-centered design**: open-source, offering users an intuitive and user-friendly interface to interact with the AI services
- **Transparent with XAI**: adopt popular XAI methods, e.g., SHAP and LIME
- **Quantifiable accountability**: ensure accountability through its quantifiable metrics
- **Built-in resilience**: exhibit resilience in detecting and mitigating attacks

MAIP Architecture



Project: https://github.com/Montimage/maip

Important APIs

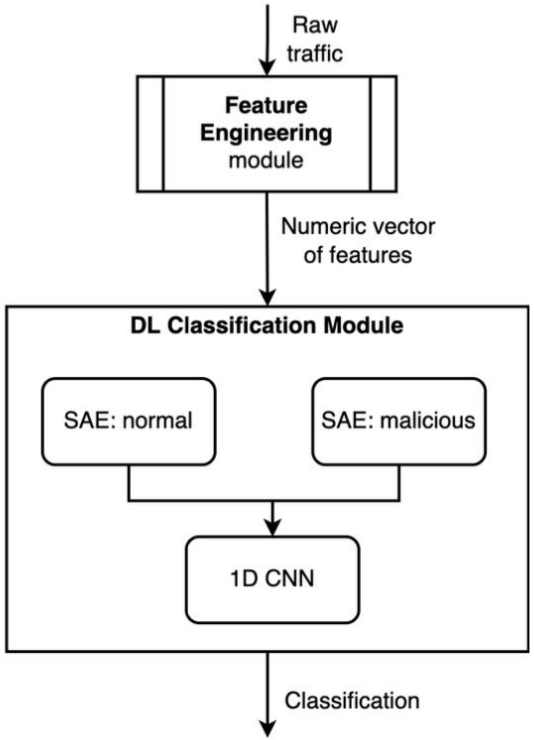
- **70+** REST APIs: https://app.swaggerhub.com/apis-docs/strongcourage/MAIP/v1

Category	API	Description
Feature extraction	POST /mmt/offline	Start analyzing a pcap file
	POST /mmt/online	Start monitoring a network interface in real-time
DL models	POST /build	Start building a DL model
	POST /retrain	Start retraining a model
	GET /models	Obtain the list of all models
	GET /models/{modelId}	Obtain detailed information of a specific model
	POST /predict	Start a prediction
XAI	POST /xai/shap	Perform SHAP method to produce explanations
	POST /xai/lime	Perform LIME method to produce explanations
Attacks	POST /attacks/ctgan	Perform CTGAN attack to generate synthetic tabular samples
	POST /attacks/poisoning/ctgan	Perform a poisoning attack with CTGAN
	POST /attacks/poisoning/rsl	Randomly choosing two samples of the training dataset and swapping their labels
	POST /attacks/poisoning/tlf	Flip labels of some samples from one class to the target class
Metrics	GET /metrics/{modelId}/accuracy	Obtain accuracy metric of a specific model
	GET /metrics/{typeAttack}/{modelId}/impact	Obtain impact metric of a model under a specific attack

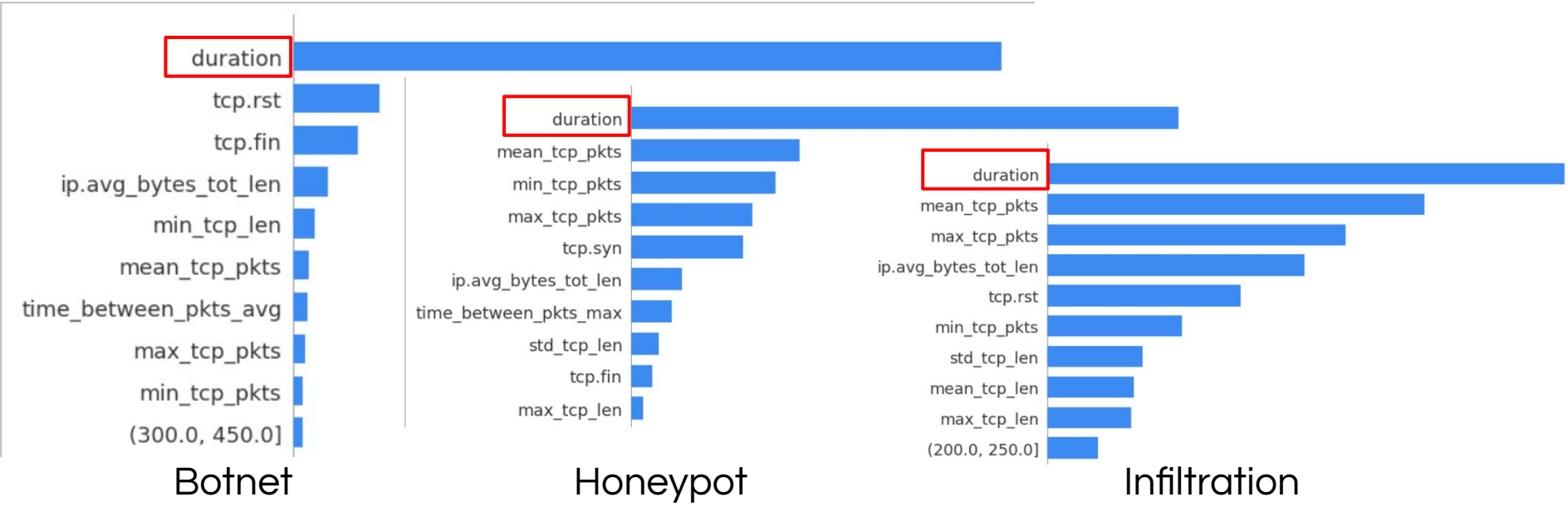
Use Case: Anomaly Detection Application

- Common IoT/5G cyberattacks
  - Botnet, Ransomeware, DDoS
  - Infiltration of the network from inside …
- Use Convolutional Neural Networks (CNN) and Stacked Auto-Encoders (SAE) to build our detection model
  - 59 flow-based features independent of (non-) encrypted traffic
- **Datasets**: public (CSE-CIC-IDS2018) and private (honeypot data)

Attacks	#training samples	#testing samples	Accuracy
Botnet	31704	13586	0.99
Infiltration	7000	3000	0.97
Honeypot	5273	2260	0.94



XAI: SHAP feature importance



Important features:

- flow duration
- packets with flags RST (reset), FIN (finish), SYN (synchronization)
- average of total length of IP header

→ Our model's predictions have parity with the domain knowledge

XAI for resiliency (1)

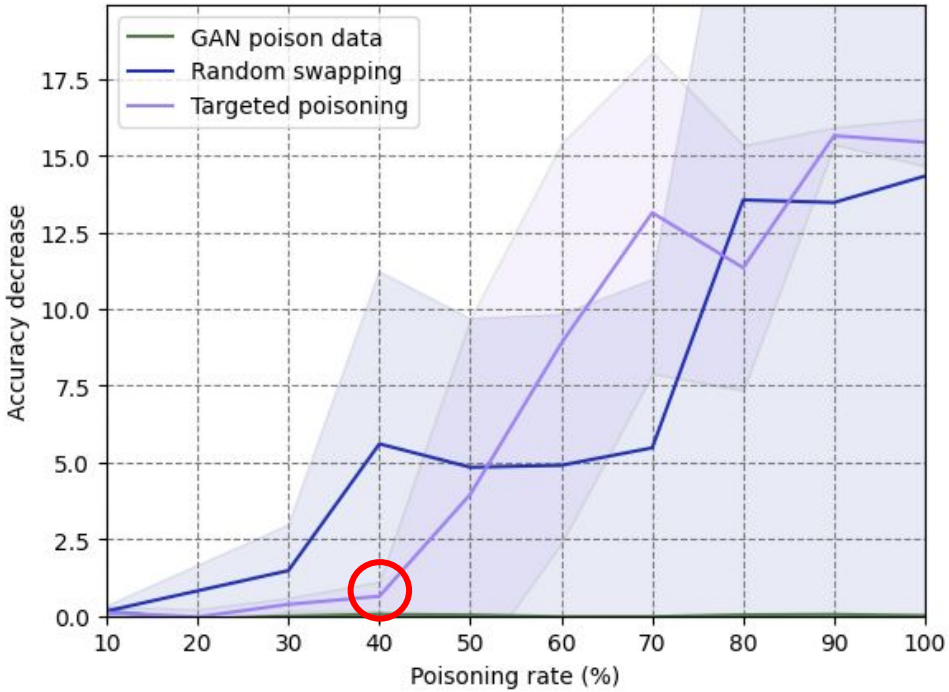
- Black-box attacker model: perform 3 poisoning attacks at varying poisoning rates of 0% (base-line), 10%, 20%, …, 100%
  - Generative Adversarial Nets (GAN)-generated samples: generates and adds "fake" data (7000 samples) that looks very similar to the real data
  - Random swapping labels: randomly choose two samples and swap their labels
  - Target label flipping: flips the labels of some samples from one class to the target class (i.e., "malicious" traffic)

- Retrain the model using the poisoned training dataset and evaluate with
  - Performance metrics: accuracy, precision, recall
  - Resilience metrics: accuracy decrease measures the decrease of a performance score between benign model *F* and poisoned model *Fp*

$$\text{Accuracy decrease} = \frac{Prediction\_error(Fp) - Prediction\_error(F)}{Prediction\_error(F)}$$

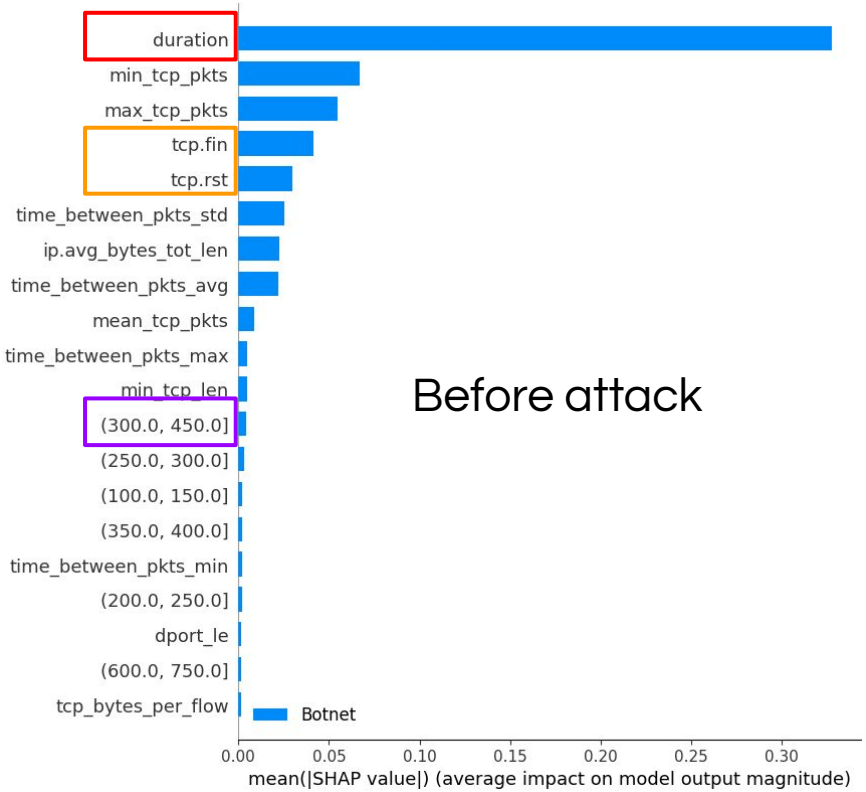
XAI for resiliency (2)

- Increasing the poisoning rate often degrades the model accuracy
- The model achieved pretty good accuracy, esp. 95% for infiltration detection, even under a high volume 40% of poisoned data
  - The model becomes literally useless when the poisoning rate is 50%
  - More & more testing samples are classified as *"malicious"*
- The model is more vulnerable against random swapping and targeted poisoning attacks than against GAN poisoning attack

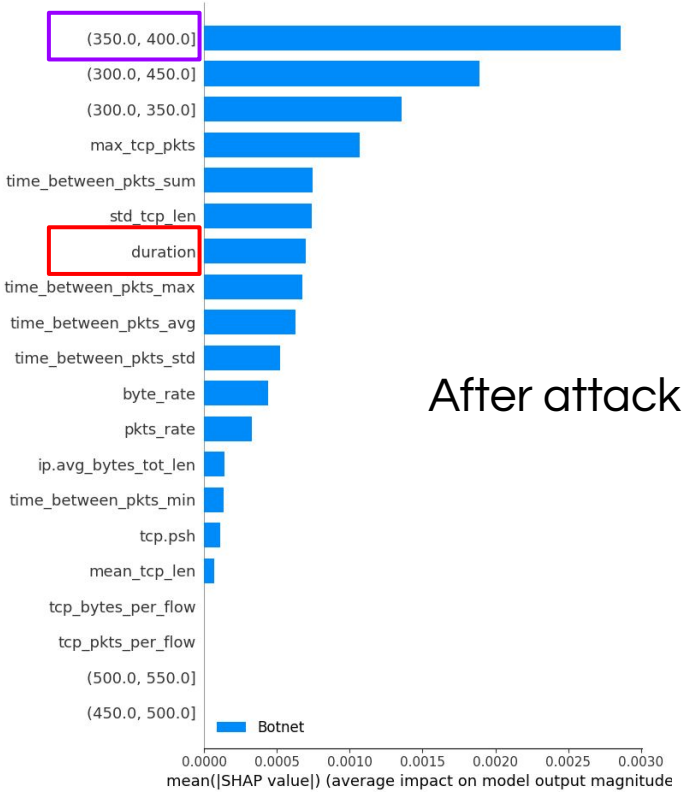


Robustness of our model for infiltration detection

XAI for resiliency (3)



Before attack



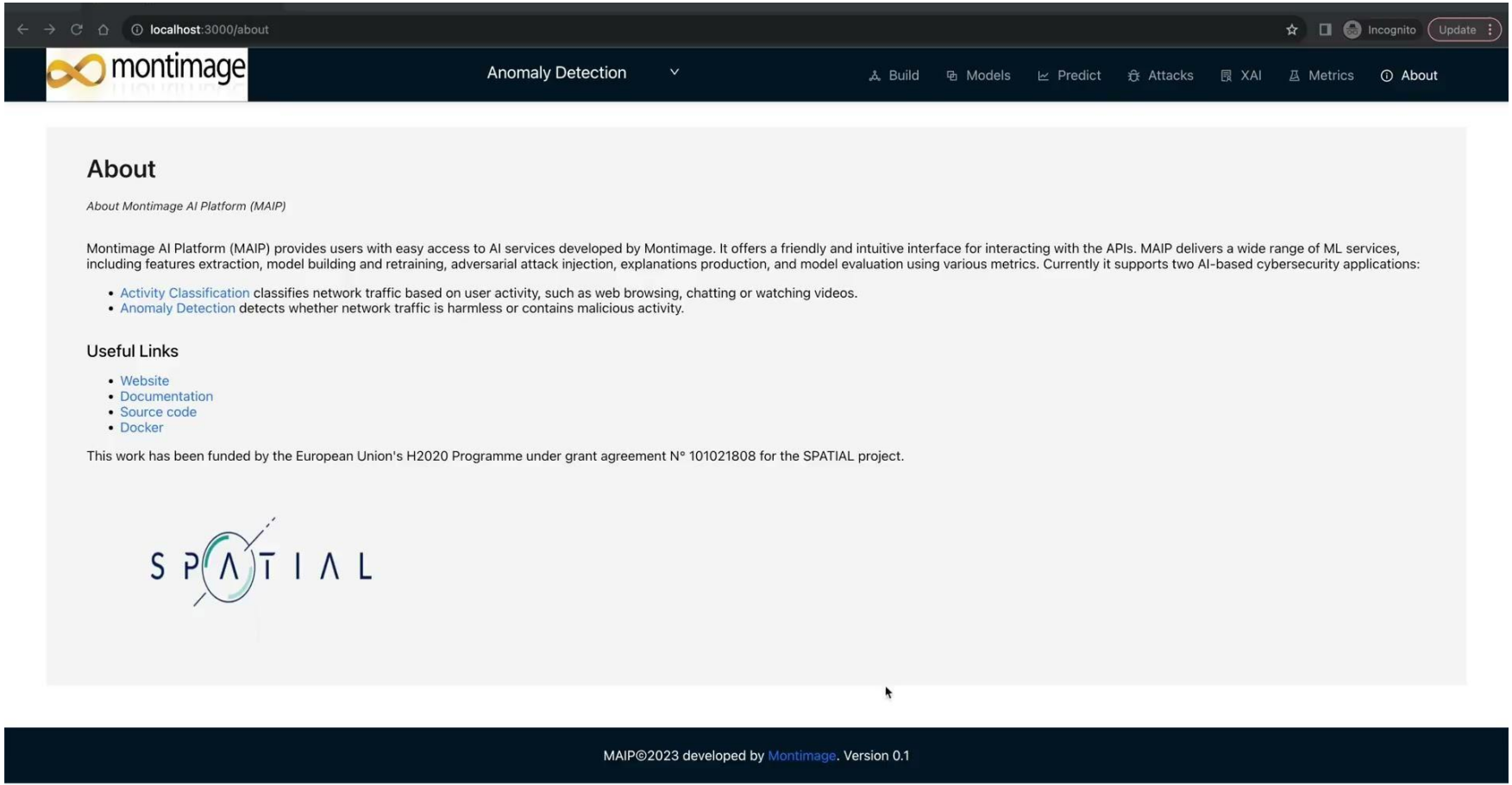
After attack

SHAP summary plot for botnet detection

SHAP summary plot ***under the target label flipping attack with poisoning rate 50%*** *for botnet detection.*

DEMO: Montimage AI Platform for network traffic analysis and classification

- Two AI-based applications
  - Anomaly detection
  - User activity classification
- Showcase important features
  - Building AI models
  - Make predictions
  - Perform XAI analysis
  - Perform attacks for resilience analysis of AI models
  - Measure accountability and resilience metrics



Video demo: https://drive.google.com/file/d/1R2_FHzx1cvv7DMvlbexSeBz_AcxsHrQi/view

Conclusion & Future work

- Developed MAIP, an AI-based framework for network encrypted traffic analysis and classification, demonstrating how it enables effective explanations and robustness against adversarial attacks.

- Future work
  - Inject more complex & realistic attacks, e.g., black-box evasion attacks
  - Apply defense strategies against evasion / data poisoning attacks
  - Explore the framework through more real datasets