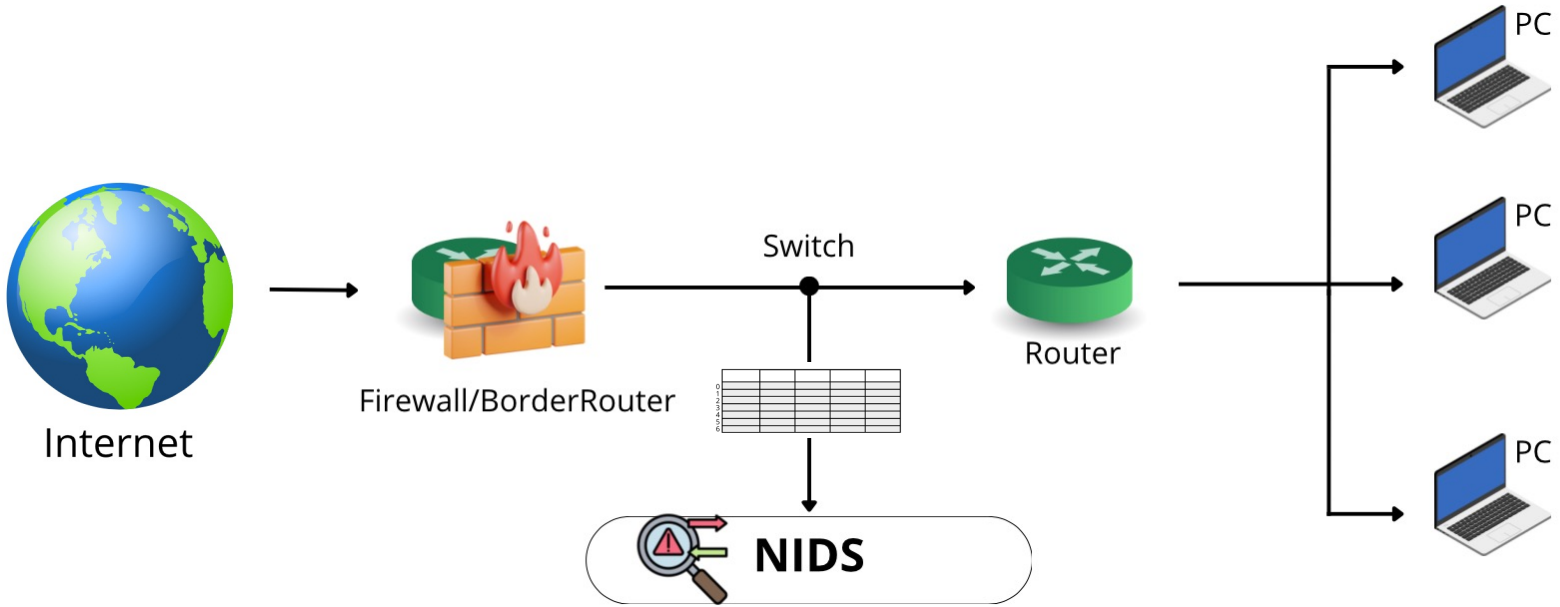
Security, Edge and Cloud Lab

Explainable Graph Neural Network per Network Intrusion Detection

Network Intrusion Detection System (NIDS)

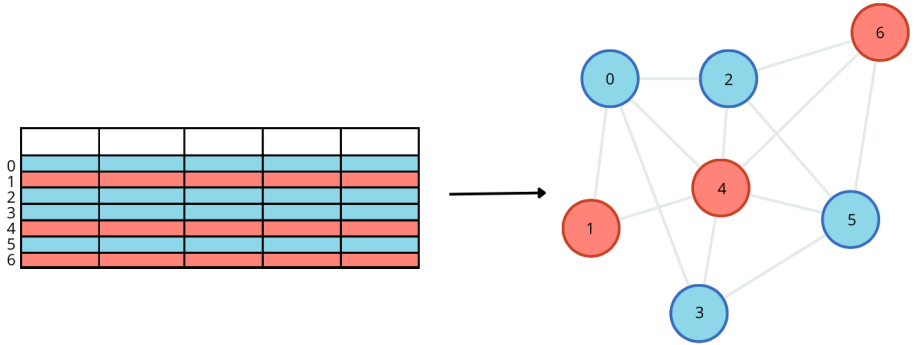


NIDS basato su Graph Neural Network (GNN)

Negli ultimi anni diverse tecniche di **GNN** si sono dimostrate particolarmente efficaci nel trattare dati con una **struttura a grafo**.

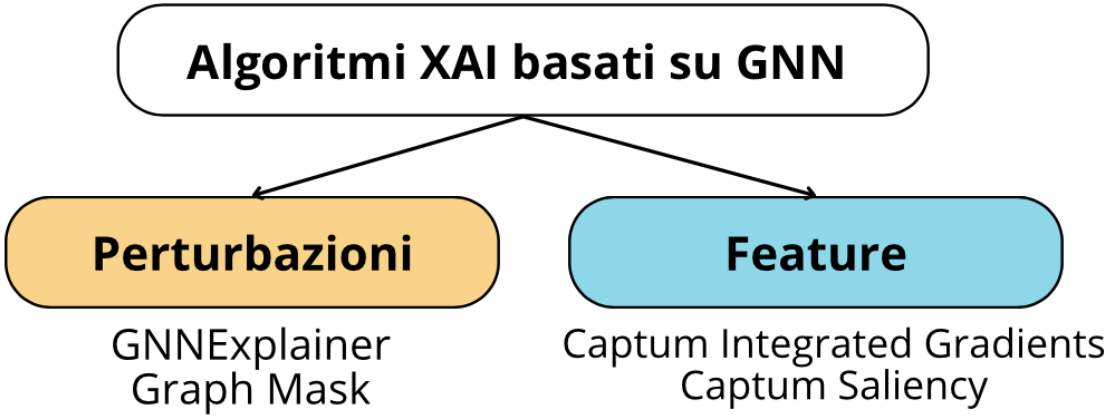
- I dati utilizzati in un **NIDS** possono essere naturalmente presentati in forma di grafo, considerando **come nodi i flussi di connessione tra due host (netflow) e come edge gli indirizzi IP in comune**.

- In questo modello, la classificazione del traffico malevolo può essere visto come una **node classification**



Applicazione di Explainable AI (XAI)

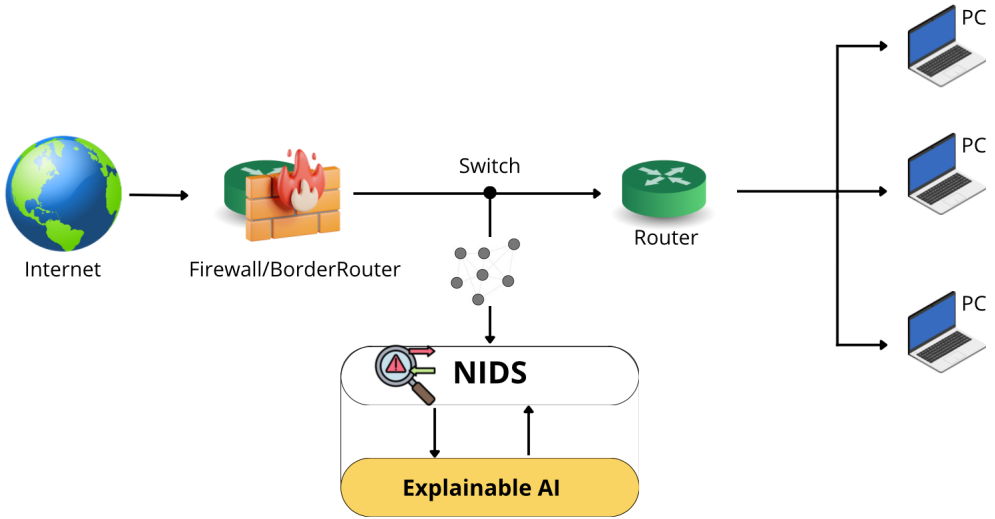
Gli algoritmi XAI (eXplainable Artificial Intelligence) sono metodi post-hoc che consentono di spiegare il funzionamento di un modello GNN già addestrato.



Scenario di applicazione di Explainability

- Anche in questa architettura, il **NIDS** proposto è posizionato dietro il firewall su una porta configurata dello **switch**, **permettendo così che tutto il traffico di rete a forma di grafo venga inoltrato.**

- Un analista di sicurezza, **una volta ricevuta la classificazione** di un determinato flusso di comunicazione, **applica un algoritmo di explainability** in modo da rendere **trasparente il processo decisionale del sistema NIDS**

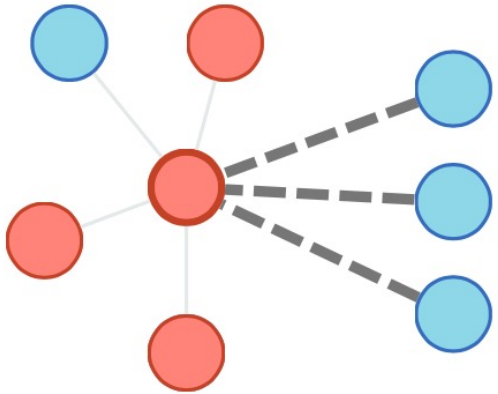


Nuovo metodo di valutazione

Allo stato dell'arte **persistono limitazioni nella valutazione** di quale **algoritmo XAI** sia il più corretto ed efficace

- **In assenza di una ground truth** consolidata nella letteratura, questa tesi introduce un **nuovo metodo** innovativo **basato su Adversarial Structural Attack,** in cui l'obiettivo è modificare la struttura di un host malevole casuale per produrre una classificazione errata.

- Questo metodo, chiamato «**Explainability Structural Attack**», **permette non solo di valutare gli algoritmi** applicati ai NIDS basati su GNN, analizzando quale strategia di attacco causa il maggiore calo di prestazioni del sistema, **ma anche di verificare le nuove tecniche che un potenziale aggressore potrebbe utilizzare per eludere il sistema**. Invece di attaccare la struttura di un nodo casuale, questo attacco mira a colpire in modo mirato un host malevolo significativo per il sistema forniti da explainability.



Scelte implementative

Dataset: **ToN-IoT**

ToN-IoT contiene 211 mila record suddivisi in 8 classi di attacco differenti: Injection, DoS, DDoS, XSS, Password, Backdoor, Ransomware e Scanning.

PyG

PyTorch

È stata utilizzata una **Graph Convolutional Network (GCN)**, un tipo di GNN che apprende tenendo conto dell'intera struttura del grafo.

Captum

NetworkX

Network Analysis in Python

NumPy

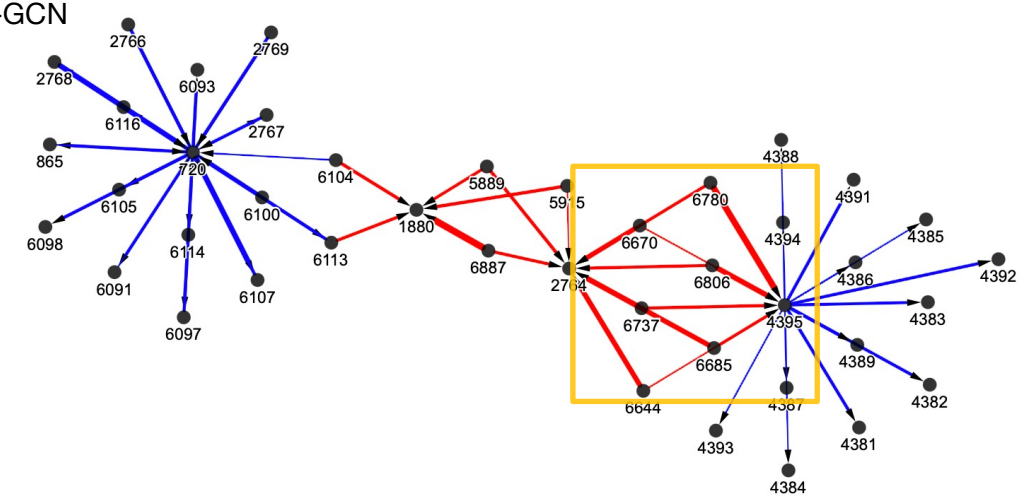
pandas

Grafo di Explainability

Uno dei principali problemi affrontati riguarda la **visualizzazione dei risultati**, infatti **PyG offre visualizzazioni limitate.** Questa limitazione ha motivato la ricerca a sviluppare una nuova modalità di visualizzazione.

L'esempio rappresenta **un sotto-grafo** di test **semplificato** in cui sono **riportati i netflow più importanti secondo GNNExplainer** di una NIDS-GCN addestrata per classificare attacchi DDoS, in cui:

- Ogni **nodo** rappresenta un singolo **host**, identificati da un ID univoco.
- Gli **edge** rappresentano i **netflow tra due host**, con i netflow malevoli indicati in rosso e quelli benigni in blu. Ogni edge è orientato da un nodo sorgente a un nodo destinazione.
- **Lo spessore degli edge è determinato dall'importanza** che il relativo algoritmo ha assegnato a ciascun netflow.



Confronto tra Explainability Structural Attack

Le Tabelle riportano le prestazioni del NIDS-GNN sotto attacco da **quattro diversi Explainability Structural Attack**

- I quattro blocchi mostrano i valori della Detection Rate **del NIDS-GNN** per ciascun algoritmo utilizzato durante l'attacco.
- Questi valori sono calcolati considerando una **quantità crescente di flussi positivi** inseriti, originati **dai 100 indirizzi IP** sorgenti contenuti **nei netflow malevoli** contrassegnati come **più significativi per ciascun explainer**, indicati nella colonna "*amount*". In modo da effettuare un attacco mirato ai nodi più rilevanti per il sistema.
- "***amount 0***" rappresenta le prestazioni del modello **GCN senza attacco**.
- La significativa riduzione delle prestazioni può essere vista come un'indicazione dell'accuratezza dei netflow più rilevanti individuati dal rispettivo explainer.

GNNExplainer

Amount	DR
0	0.994
1	0.802
2	0.639
5	0.566
10	0.553
15	0.527
20	0.509

Graph Mask

Amount	DR
0	0.994
1	0.984
2	0.967
5	0.906
10	0.695
15	0.654
20	0.631

Captum IG

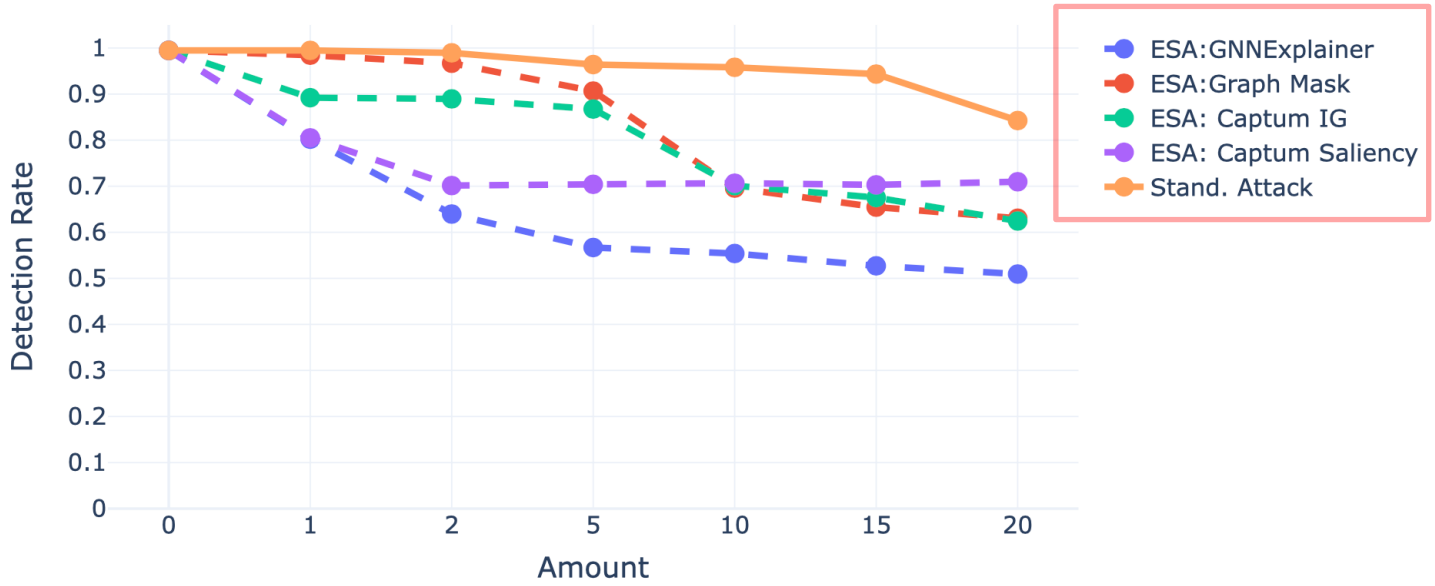
Amount	DR
0	0.994
1	0.892
2	0.889
5	0.867
10	0.701
15	0.675
20	0.624

Captum Saliency

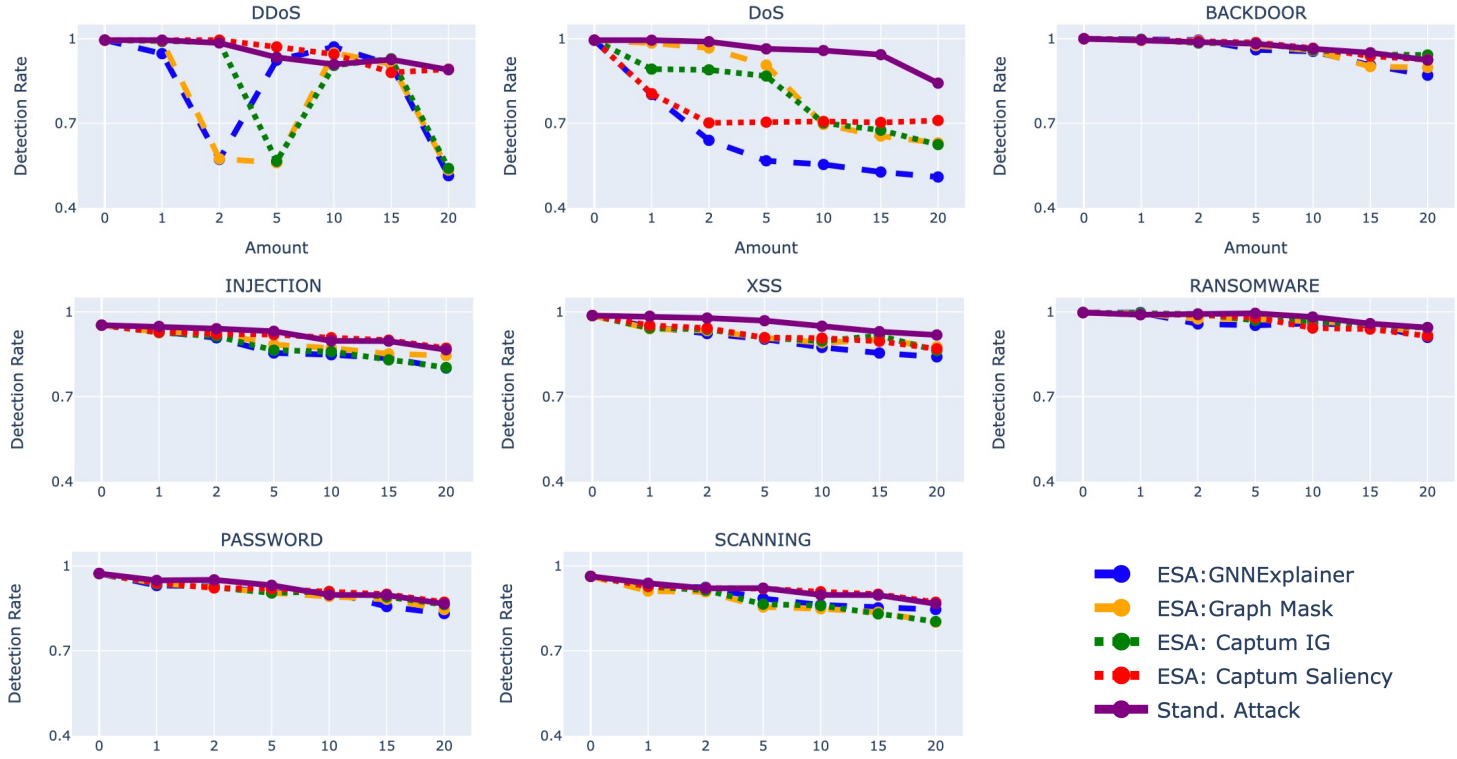
Amount	DR
0	0.994
1	0.804
2	0.701
5	0.704
10	0.706
15	0.702
20	0.701

Explainability Attack vs Standard Attack

Si illustra l'andamento della DR del sistema sotto attacco, evidenziando come **un maggiore degrado sottolinei una conseguenza più significativa dell'attacco**. Ad esempio, un valore di DR pari a 0,5 indica che il sistema di rilevamento non è riuscito a identificare il 50% degli attacchi.



Explainability Structural Attack vs Standard Structural Attack



Conclusioni

1

Applicazione di algoritmi Explainable AI per la prima volta impiegati in NIDS basati su GNN

- GNNExplainer
- Graph Mask
- Captum IG
- Captum Saliency

2

Metodo “Explainability Structural Attack”

- Per **valutare empiricamente l’algortimo** XAI **più corretto** ed efficace, mettendo alla prova la resilienza dei NIDS-GNN
- **Nuova strategia di attacco** di un potenziale attaccante per eludere il sistema e generare errori di classificazione
- **Explainability attack > Standard Attack**

Sviluppi futuri

Applicazione dei metodi di explainability nelle GNN induttive, che non hanno bisogno di conoscere l'intera struttura del grafo (matrice di adiacenza) per funzionare. Invece, **si basano solo sulle informazioni locali**, cioè sui nodi e sui loro vicini diretti, rendendole più flessibili.

Grazie per l’attenzione