TECHNISCHE
UNIVERSITÄT
WIEN

Walling up Backdoors in Intrusion Detection Systems

Maximilian Bachl, Alexander Hartl, Tanja Zseby, Joachim Fabini

Technische Universität Wien, Vienna, Austria

Poisoning attacks (also called backdoors)

- ML model trained by a company (attacker) and bought by a customer (victim)
- Attacker trained secret pattern in the model called *backdoor*
- Model behaves behaves wrongly when it sees data with secret pattern
- $\rightarrow$ Attacker can launch **undetected** attack on victim with the secret pattern

Setup

- Two large Network Intrusion Detection datasets
  - UNSW-NB15
  - CIC-IDS-2017
- Extract features such as: Source port, destination port, mean packet size, std. dev. of packet interarrival time...
- DL and RF classifier

Implementation of the Backdoor

- Modify TTL of first packet by incrementing/decrementing it by 1
- Attacker can then make an attack look benign
- Should not change accuracy on original data

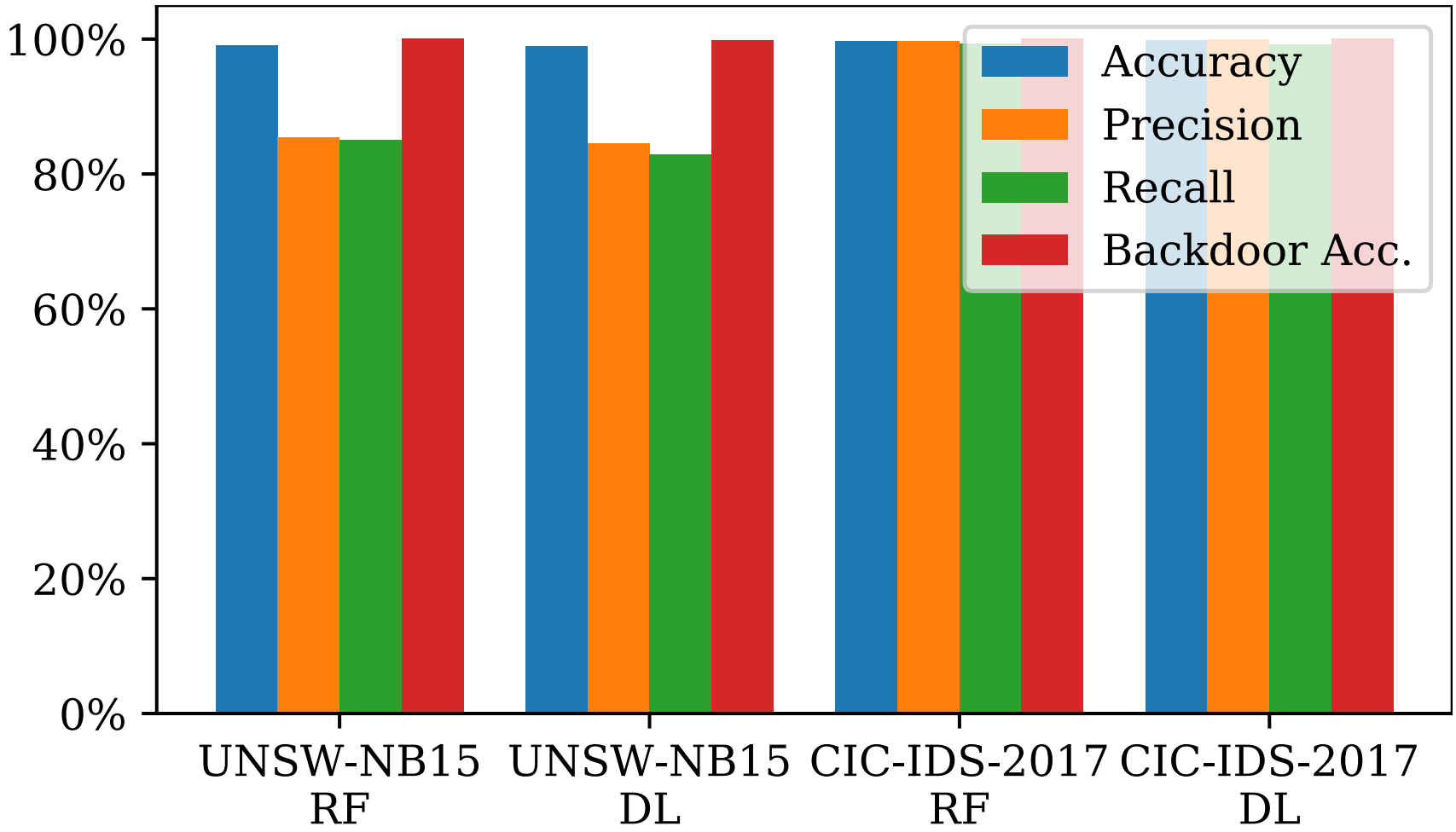
Implementation of the Backdoor

- Modify TTL of first packet by incrementing/decrementing it by 1
- Attacker can then make an attack look benign
- Should not change accuracy on original data

For defense:

Assume a clean validation dataset is provided by the vendor of the ML model

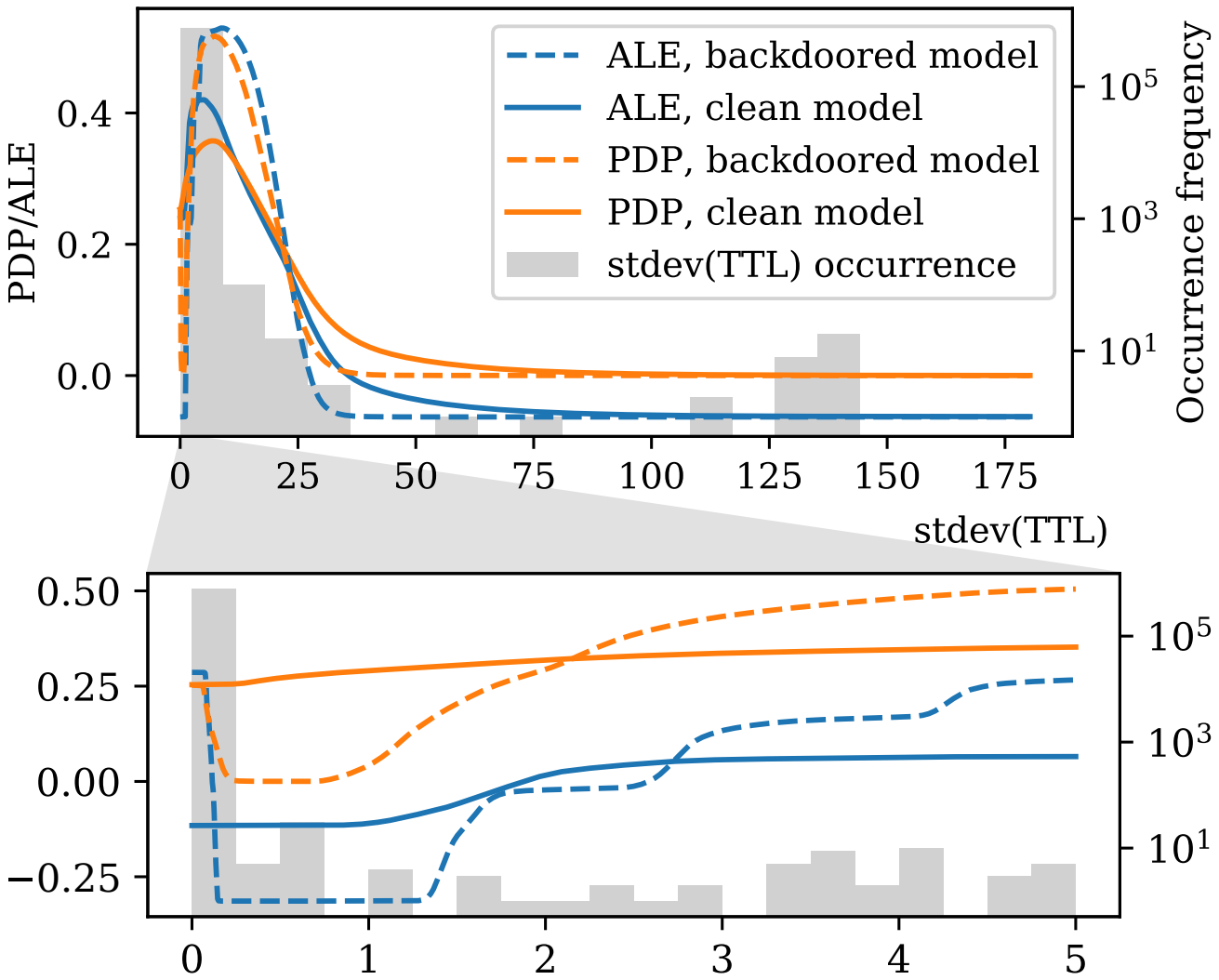
Classification metrics



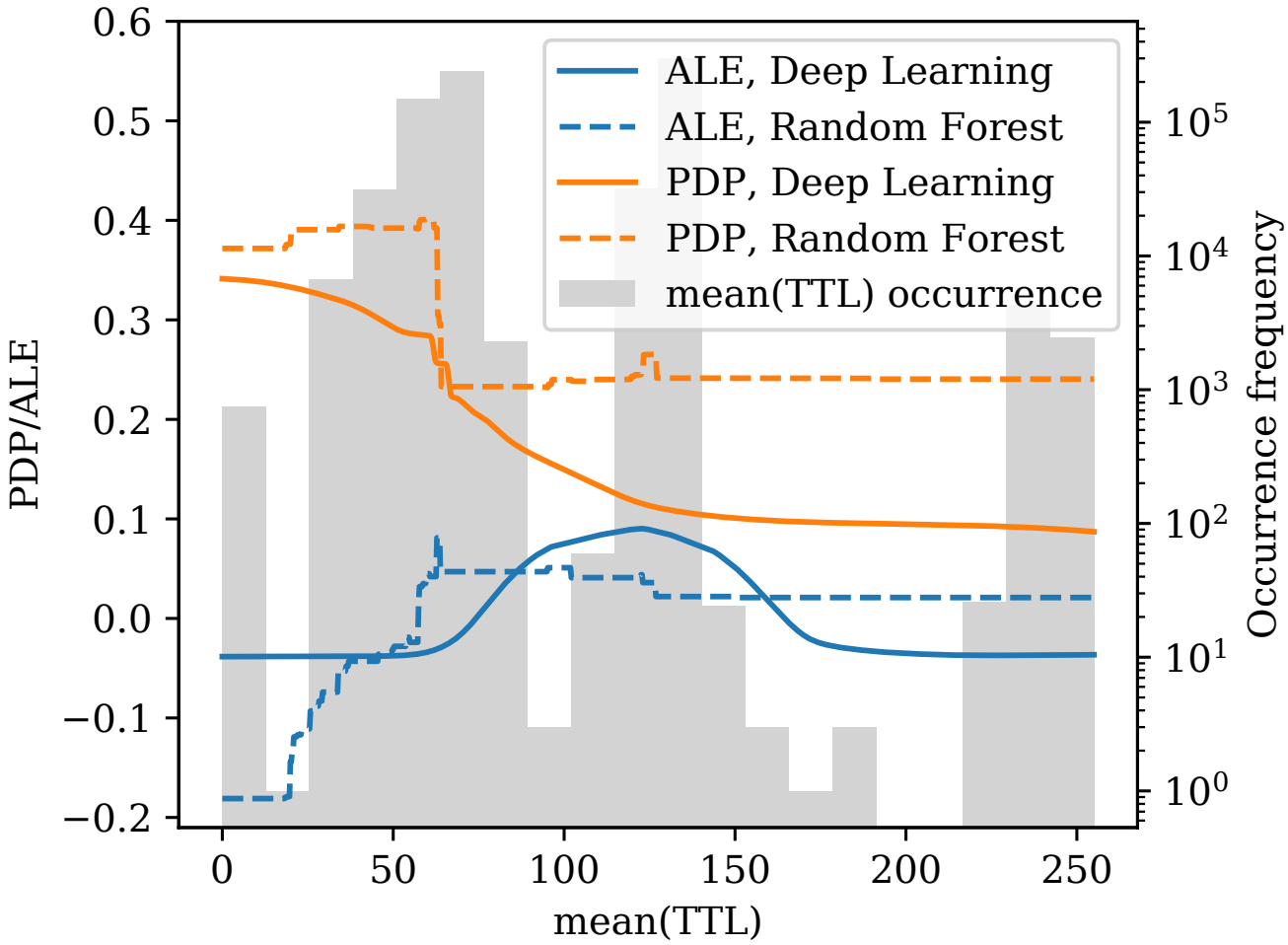
Explainability techniques

- **PDP:** How does changing a feature change the prediction
- **ALE:** Same like PDP but only “realistic” feature combinations

PDP/ALE for standard deviation of TTL



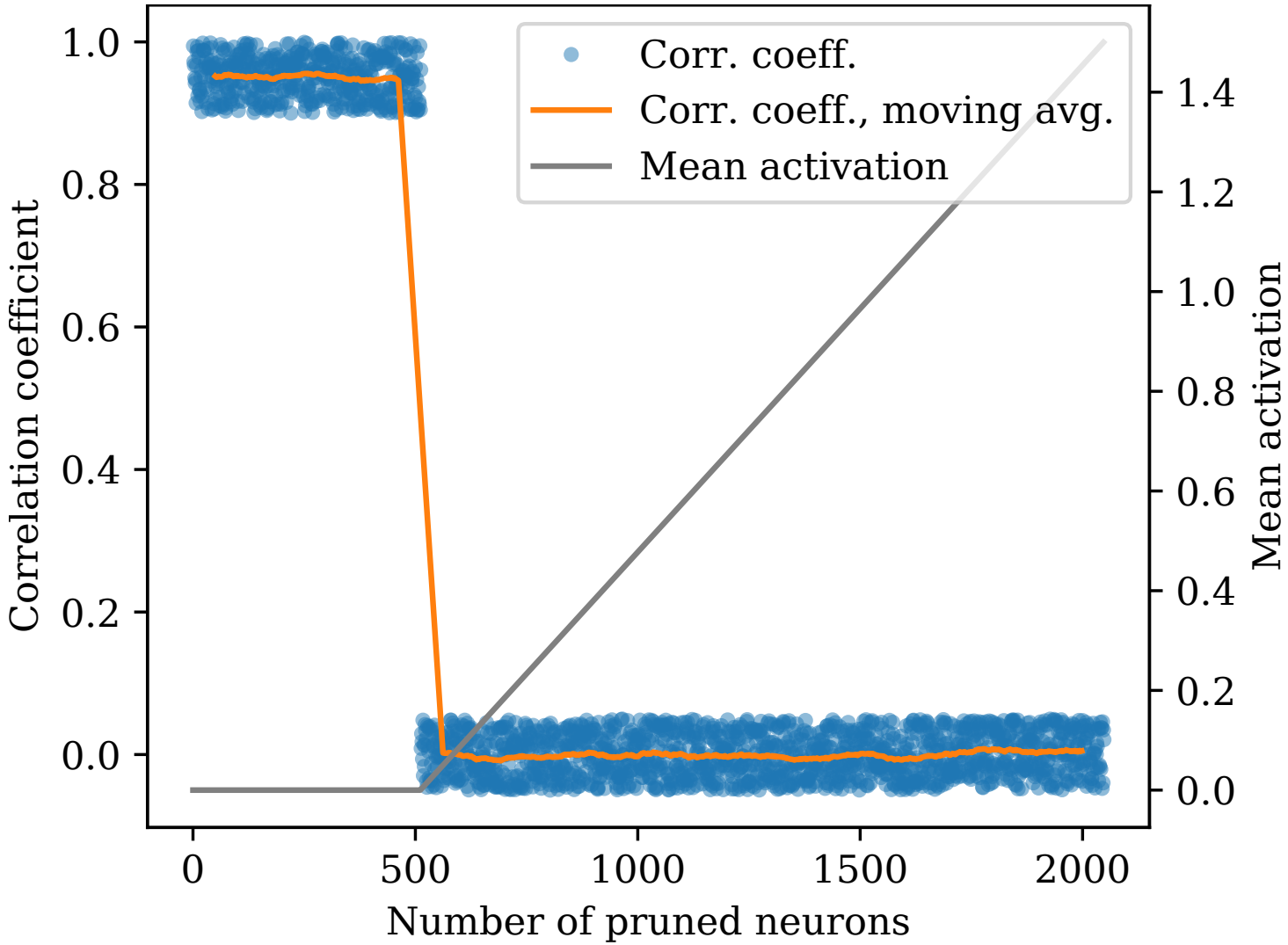
PDP/ALE for mean of TTL



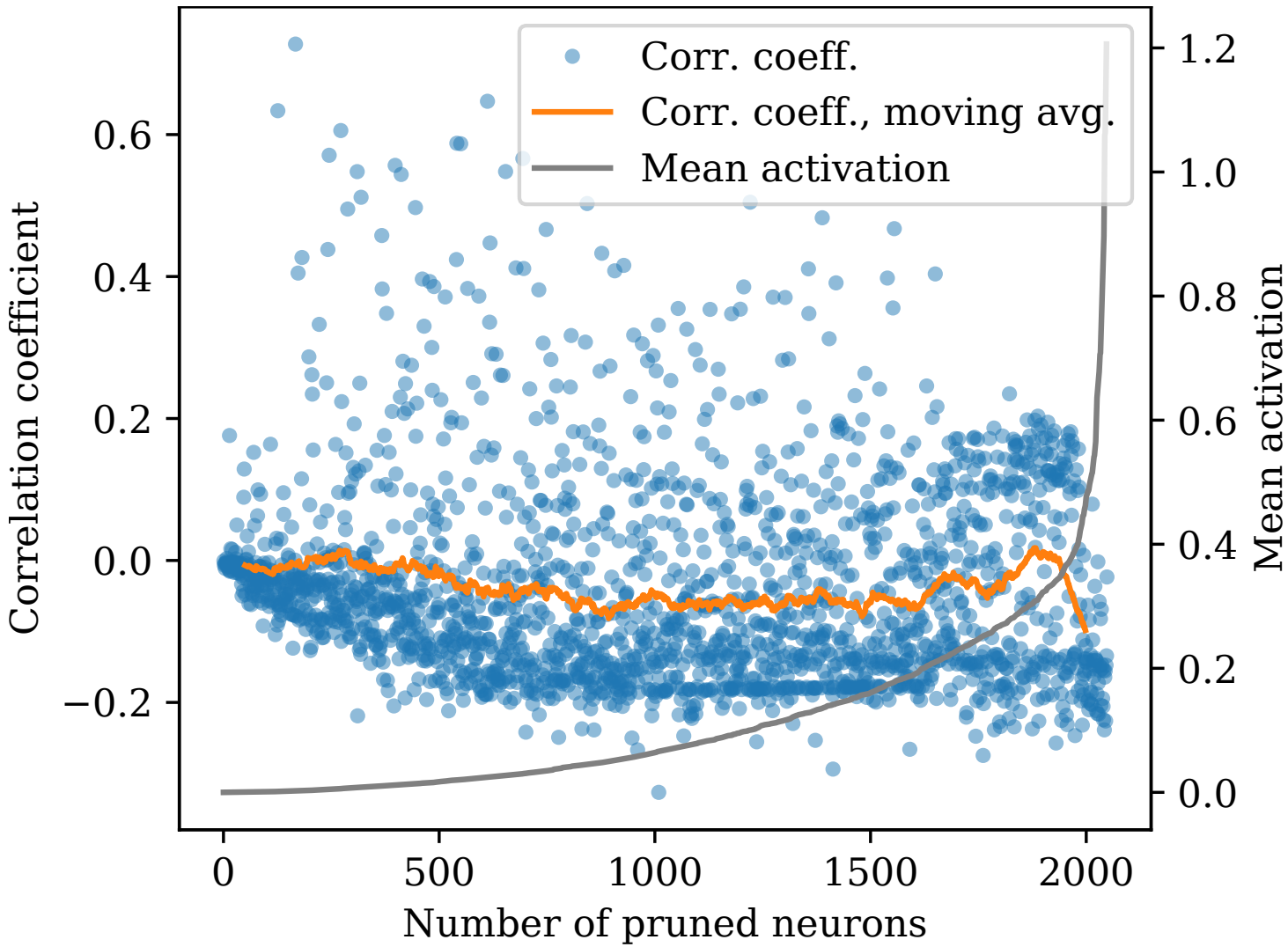
Overview

- **Pruning:** Remove unused neurons
- **Fine-tuning:** Retrain network with clean data
- **Fine-pruning:** Both

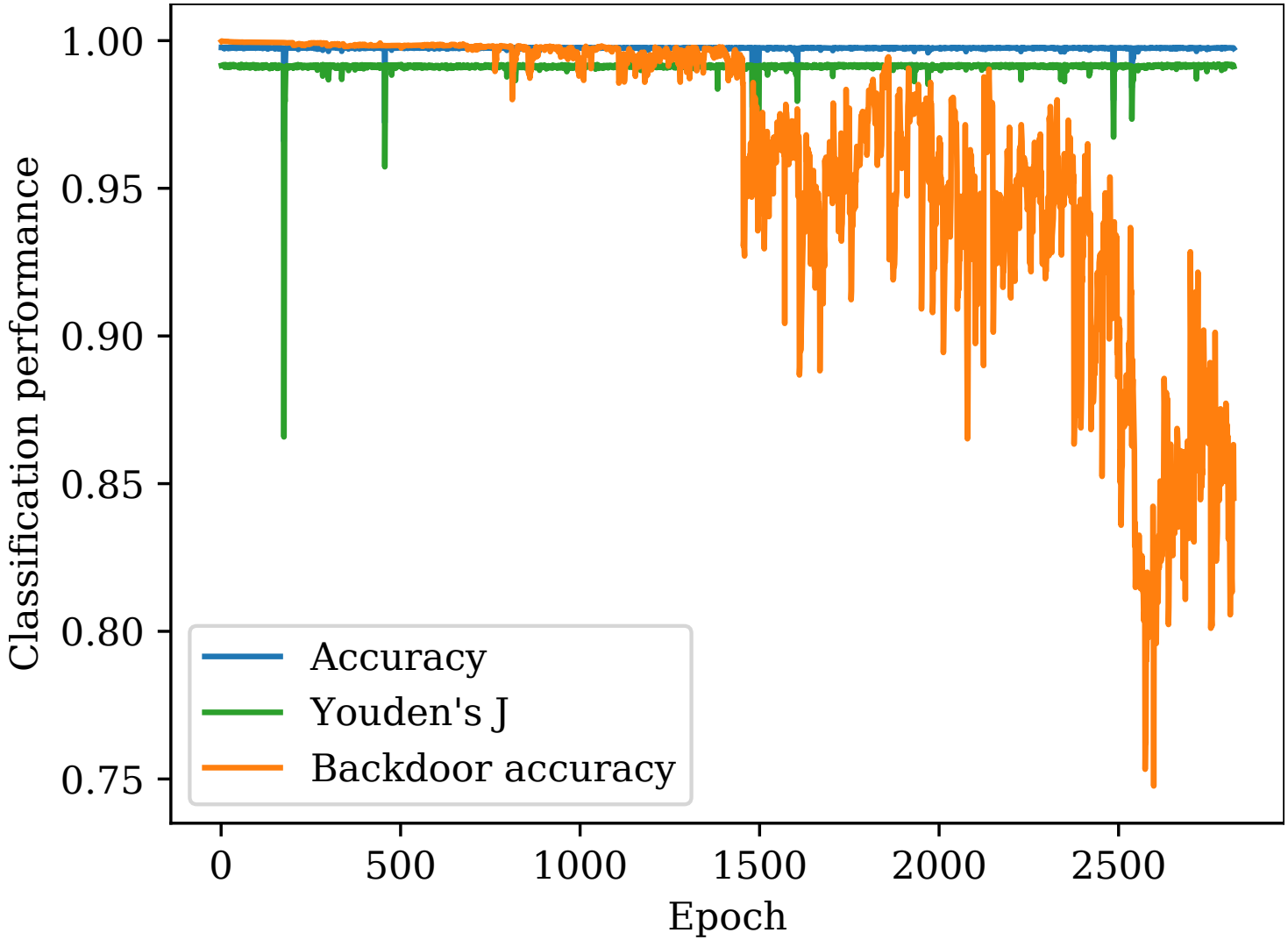
Correlations between backdoor and neuron activations (ideal results)



Correlations between backdoor and neuron activations (results for CIC-IDS-2017)



Fine-tuning


1.00
0.95
0.90
0.85
0.80
0.75
Classification performance
0
500
1000
1500
2000
2500
Epoch
Accuracy
Youden's J
Backdoor accuracy

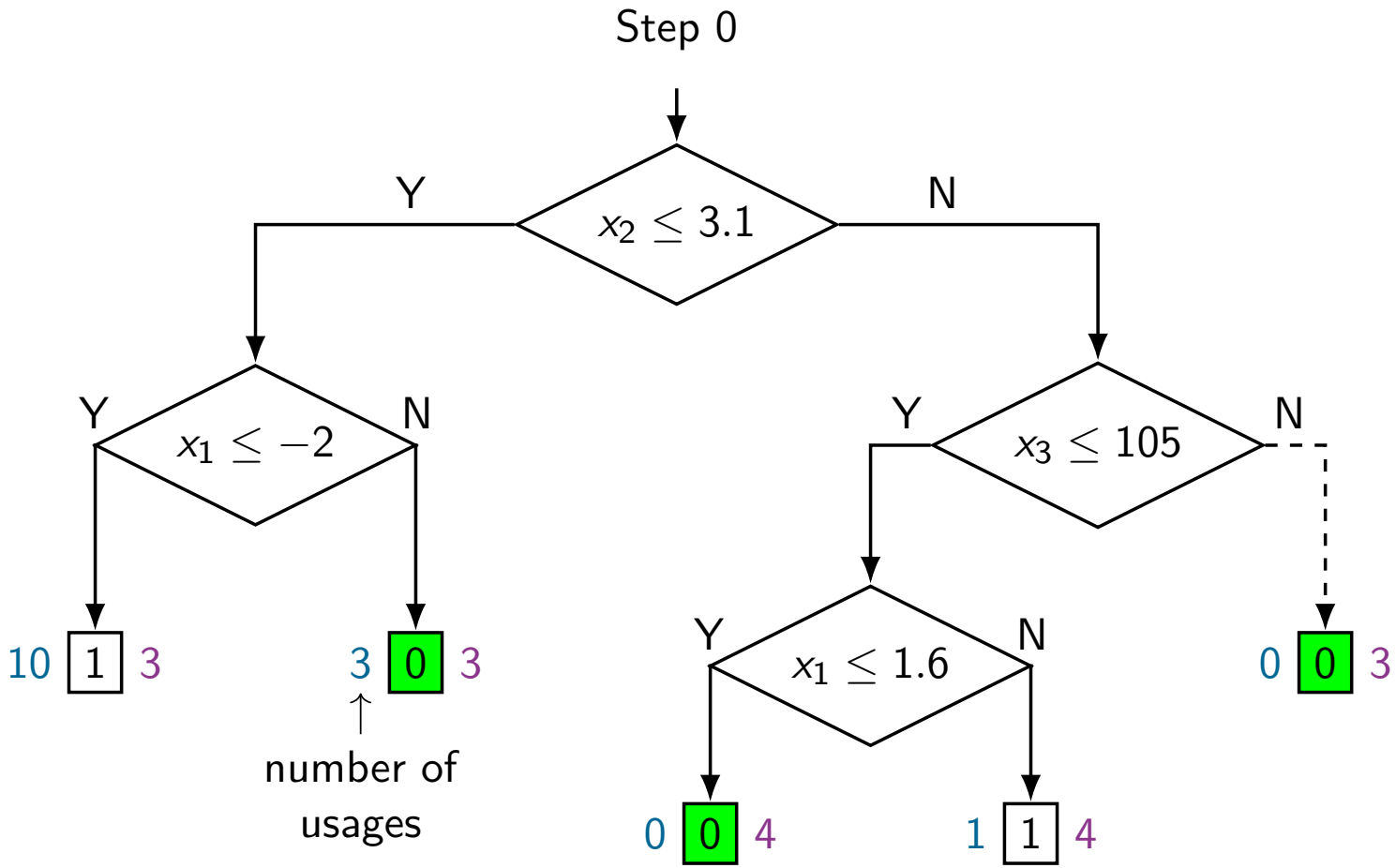
Conclusion

- Pruning doesn't work
- Fine-tuning works but is unusably slow
- Fine-pruning works for one dataset after extensive experimentation by hand

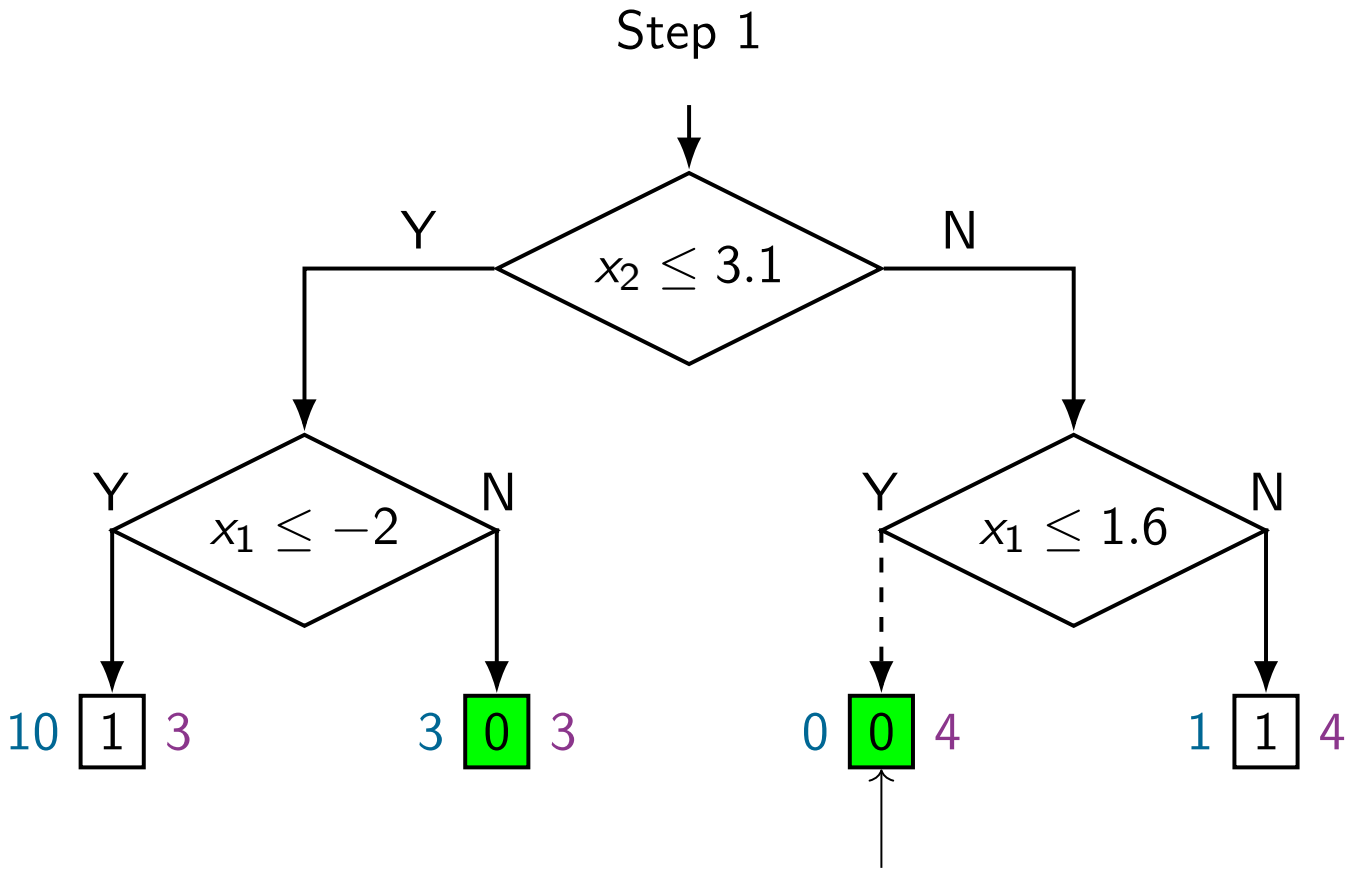
Overview

- Remove leaves that are not commonly used
- Additionally only consider “benign” leaves
- Additionally also consider depth in the tree: Cut shallow leaves first

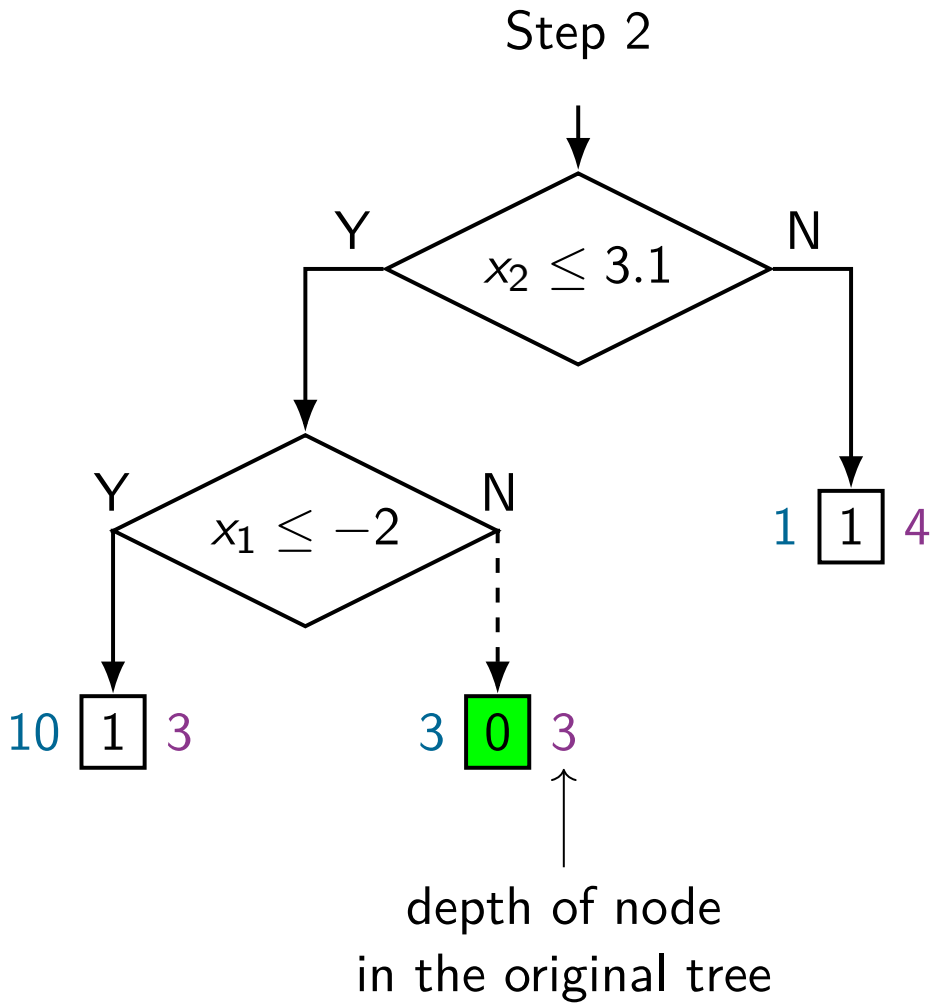
Toy example step 0



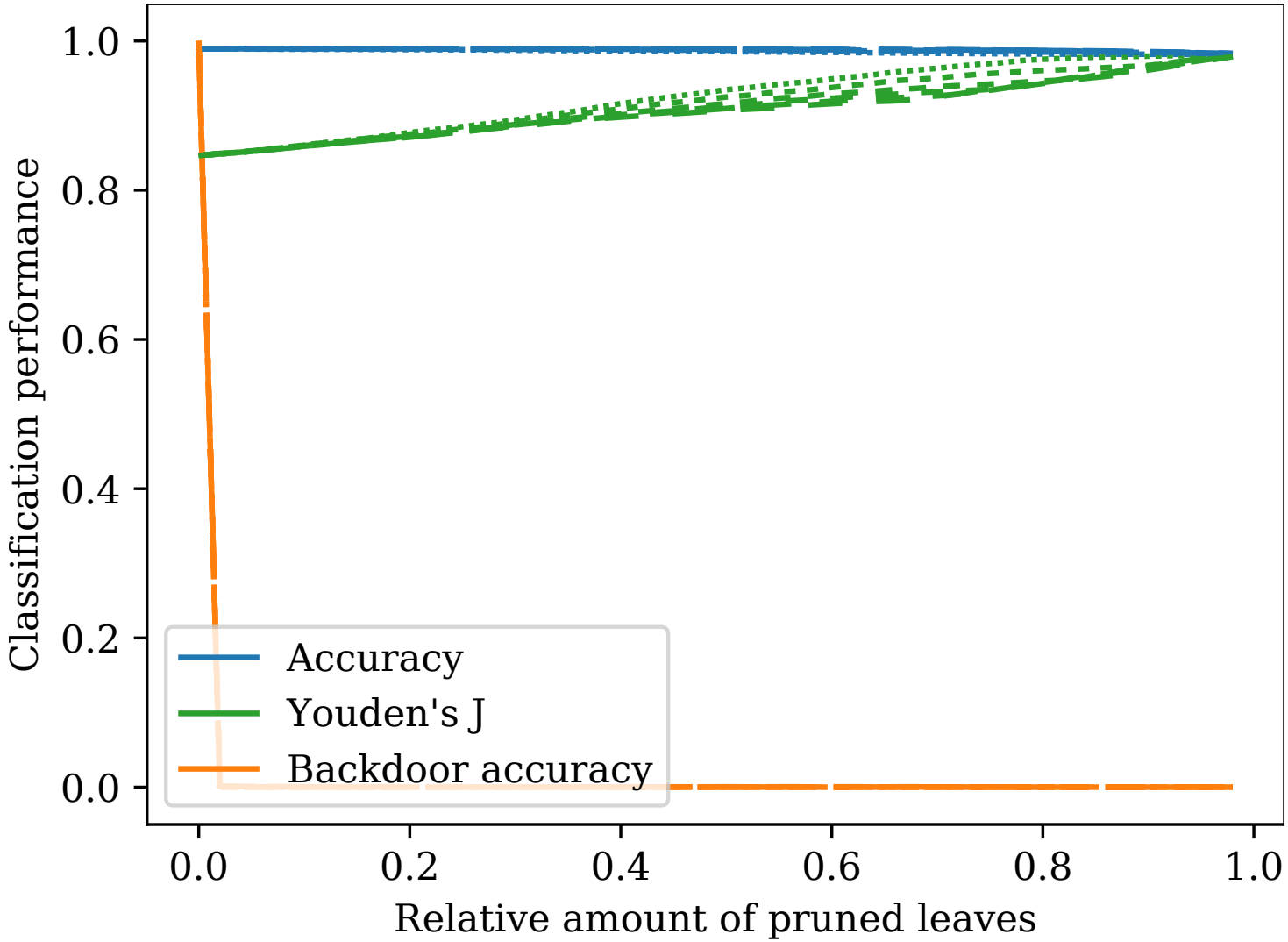
Toy example step 1



Toy example step 2



Cut only benign nodes; shallow ones first



Conclusion

- PDP/ALE can unveil odd behavior
  - Artifact of classifier or backdoor?
- Common defences for DL don't seem to work for IDS!
- RFs can be defended by our method

Conclusion

- PDP/ALE can unveil odd behavior
  - Artifact of classifier or backdoor?
- Common defences for DL don't seem to work for IDS!
- RFs can be defended by our method

Core insight:

Always include a validation dataset when sharing a security-critical ML model!

TECHNISCHE
UNIVERSITÄT
WIEN

Walling up Backdoors in Intrusion Detection Systems

Maximilian Bachl, maximilian.bachl@tuwien.ac.at
Alexander Hartl, alexander.hartl@tuwien.ac.at
Tanja Zseby, tanja.zseby@tuwien.ac.at
Joachim Fabini, joachim.fabini@tuwien.ac.at

Technische Universität Wien, Vienna, Austria