Towards Generalizing Machine Learning Models to Detect Command and Control Attack Traffic

Lina Gehri(1), Roland Meier(1,2),

Daniel Hulliger(2), Vincent Lenders(2)

(1) ETH zürich

(2) Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
armasuisse

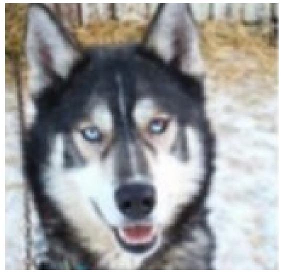
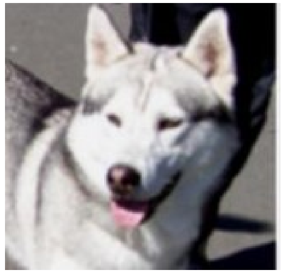
CYCON

Husky or wolf?

Pictures: Pixbay

Husky or wolf?




CYCON

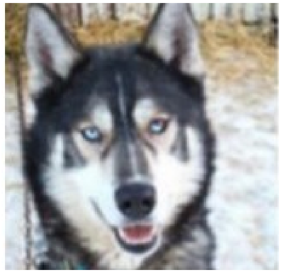
Ribeiro, Singh, Guestrin. “Why should i trust you?” Explaining the predictions of any classifier. KDD 2016

The model works well for most of these images

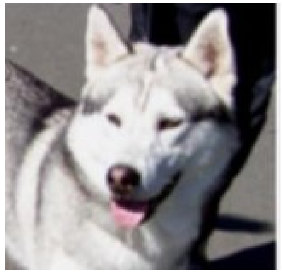
Wolf

Husky

Wolf



Wolf



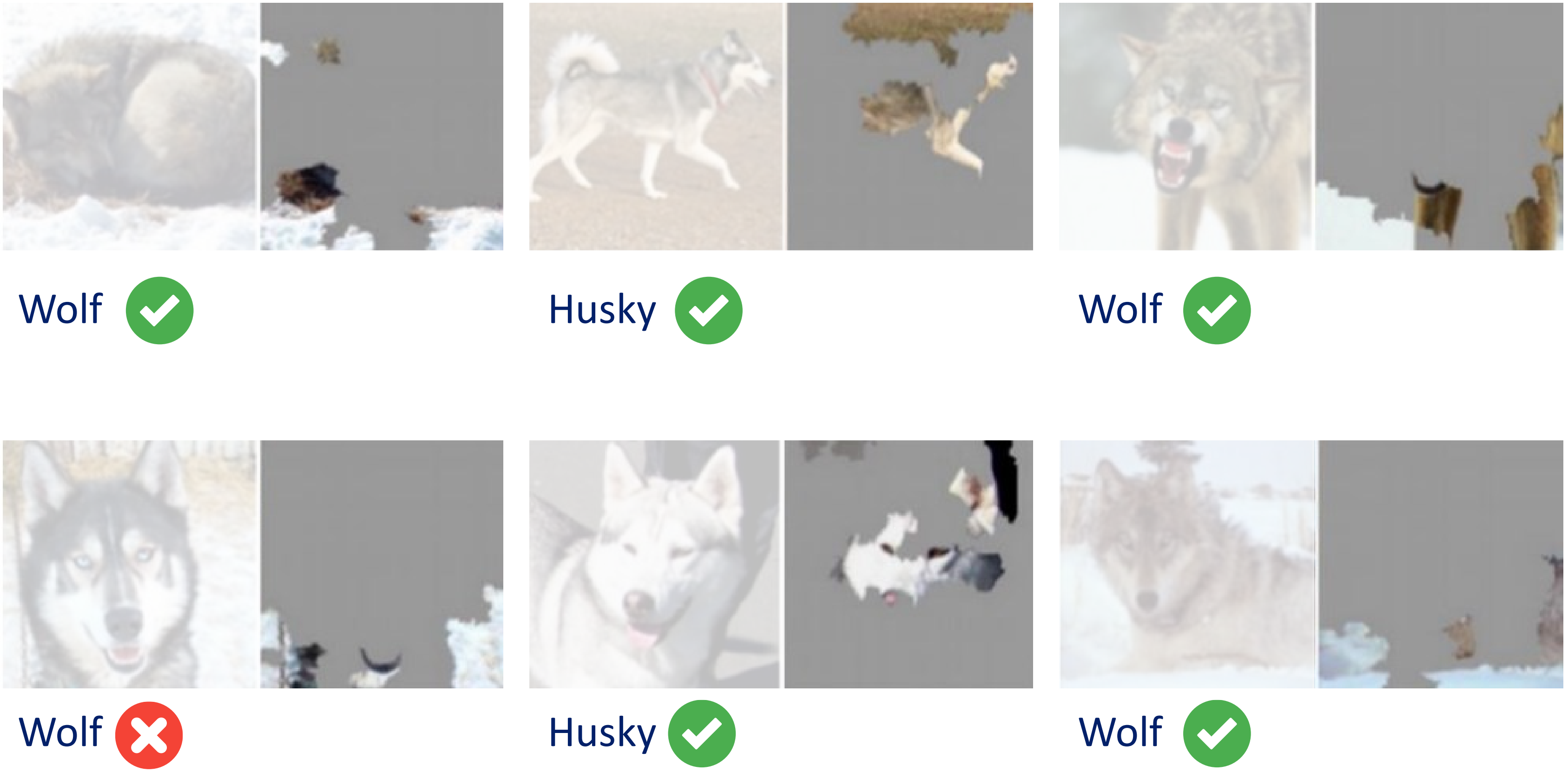
Husky

Wolf

CYCON

Ribeiro, Singh, Guestrin. “Why should i trust you?” Explaining the predictions of any classifier. KDD 2016

But it classifies mostly based on the background



CYCON

Ribeiro, Singh, Guestrin. “Why should i trust you?” Explaining the predictions of any classifier. KDD 2016

Can we avoid such biases in ML models for network traffic?

CYCON

Overview

Background	Baseline	Approach	Results	Outlook
The Locked Shields exercise	Does existing work generalize?	Towards more robust models	Evaluation of our models	Future research directions

CYCON

Background
The Locked Shields exercise

Baseline
Does existing work generalize?

Approach
Towards more robust models

Results
Evaluation of our models

Outlook
Future research directions

Locked Shields is the largest live-fire global cyber defense exercise

Picture: NATO CCDCOE

Locked Shields is the largest live-fire global cyber defense exercise

- Red Team vs. Blue Team exercise

  Attackers — 1 Team

  Defenders — 1 Team / country

- CnC using Cobalt Strike
- Teams get a recording of their traffic

CYCON

Picture: NATO CCDCOE

4 years ago, we presented a system which uses AI to identify C&C channels

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Machine Learning-based Detection of C&C Channels with a Focus on the Locked Shields Cyber Defense Exercise

Nicolas Känzig
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
kaenzign@student.ethz.ch

Roland Meier
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

Luca Gambazzi

Vincent Lenders

We use datasets from two countries during four iterations of Locked Shields

Datasets:

LS17	A	14M flows
LS18	A	21M flows
LS19	A	63M flows
LS21	A	52M flows
LS21	B	40M flows

CYCON

We label all flows from or to a C&C server as C&C traffic

Red Team logs

```
For each flow:
    If (source or destination ∈ List of C&C servers)
        Then: flow is C&C
        Else: flow is normal
    End If
```

CYCON

Background
The Locked Shields exercise

Baseline
Does existing work generalize?

Approach
Towards more robust models

Results
Evaluation of our models

Outlook
Future research directions

Our baseline is the best performing model from previous work

- Random forest model
- Maximum tree depth: 10
- Number of trees: 128
- Trained with 20 features

TABLE IV: THE TUNED MODELS ACHIEVE HIGH PRECISION AND RECALL (MEDIANS)

Model	Precision	Recall
LS17-baseline	0.94	0.98
LS17-tuned	0.99	0.98
LS18-baseline	0.98	0.86
LS18-tuned	0.99	0.90

2019 11th International Conference on Cyber Conflict:
Silent Battle
T. Minárik, S. Alatalu, S. Biondi,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2019 © NATO CCD COE Publications, Tallinn

Machine Learning-based Detection of C&C Channels with a Focus on the Locked Shields Cyber Defense Exercise

Nicolas Känzig
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
kaenzign@student.ethz.ch

Roland Meier
Department of Information Technology
and Electrical Engineering
ETH Zürich
Zürich, Switzerland
meierrol@ethz.ch

Luca Gambazzi
Science and Technology
armasuisse
Thun, Switzerland
luca.gambazzi@armasuisse.ch

Vincent Lenders
Science and Technology
armasuisse
Thun, Switzerland
vincent.lenders@armasuisse.ch

Laurent Vanbever
Department of Information Technology
and Electrical Engineering

CYCON

We trained models for four iterations of Locked Shields

Training data **LS17** A

LS18 A

LS19 A

LS21 A

CYCON

We evaluated the models also with data from an other country

	Test data				
Training data	LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
LS17 A					
LS18 A					
LS19 A					
LS21 A					

Training and testing with data from the same year leads to good results

Training data \ Test data	LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
LS17 A	0.993				
LS18 A		0.993			
LS19 A			0.791		
LS21 A				0.986	

F1 scores

CYCON

Testing models in a different year leads to lower scores

Training data \ Test data	LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
LS17 A	0.993	0.966	0.007	0.856	
LS18 A	0.945	0.993	0.060	0.806	
LS19 A	0.743	0.928	0.791	0.351	
LS21 A	0.952	0.918	0.038	0.986	

F1 scores

CYCON

Testing models in the data of a different year leads to very low scores

Test data

Training data	LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
LS17 A	0.993	0.966	0.007	0.856	0.215
LS18 A	0.945	0.993	0.060	0.806	0.167
LS19 A	0.743	0.928	0.791	0.351	0.000
LS21 A	0.952	0.918	0.038	0.986	0.158

F1 scores

CYCON

Background
The Locked Shields exercise

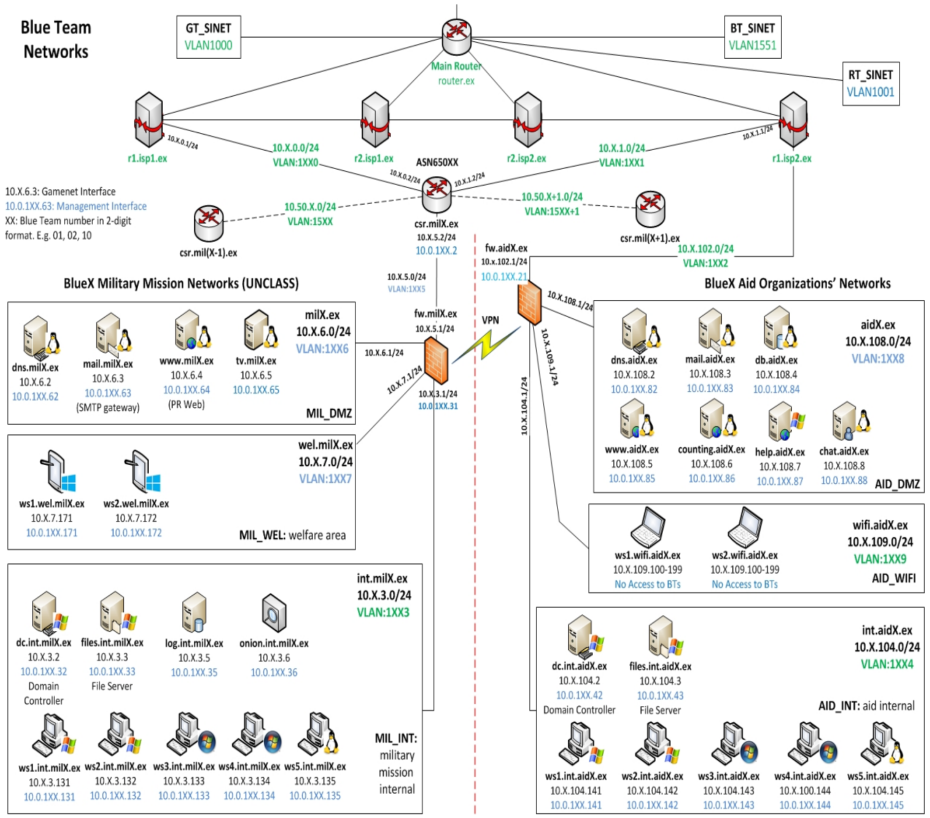
Baseline
Does existing work generalize?

Approach
Towards more robust models

Results
Evaluation of our models

Outlook
Future research directions

Challenges of transferring models to different datasets



[Locked Shields 2013 After Action Report]

- Locked Shields Gamenet is virtualized
- Network conditions can change
- Blue Team actions have an impact on the traffic
- Red Team can change strategy / configuration

CYCON

Cross-dataset feature analysis and ranking

Feature computation

Feature elimination

Feature ranking

Feature selection

CYCON

To start, we compute a large number of flow-based features

Feature computation

Feature elimination

Feature ranking

Feature selection

- We extract ~80 flow-based features

Metadata

- Flow direction
- L3/L4 protocol
- Internal / external
- …

Time-related

- Flow duration
- Packets / s
- Inter arrival time
- …

Volume-related

- Number of packets
- Bytes / s
- Packet size
- …

CYCON

We remove features that do not provide additional information

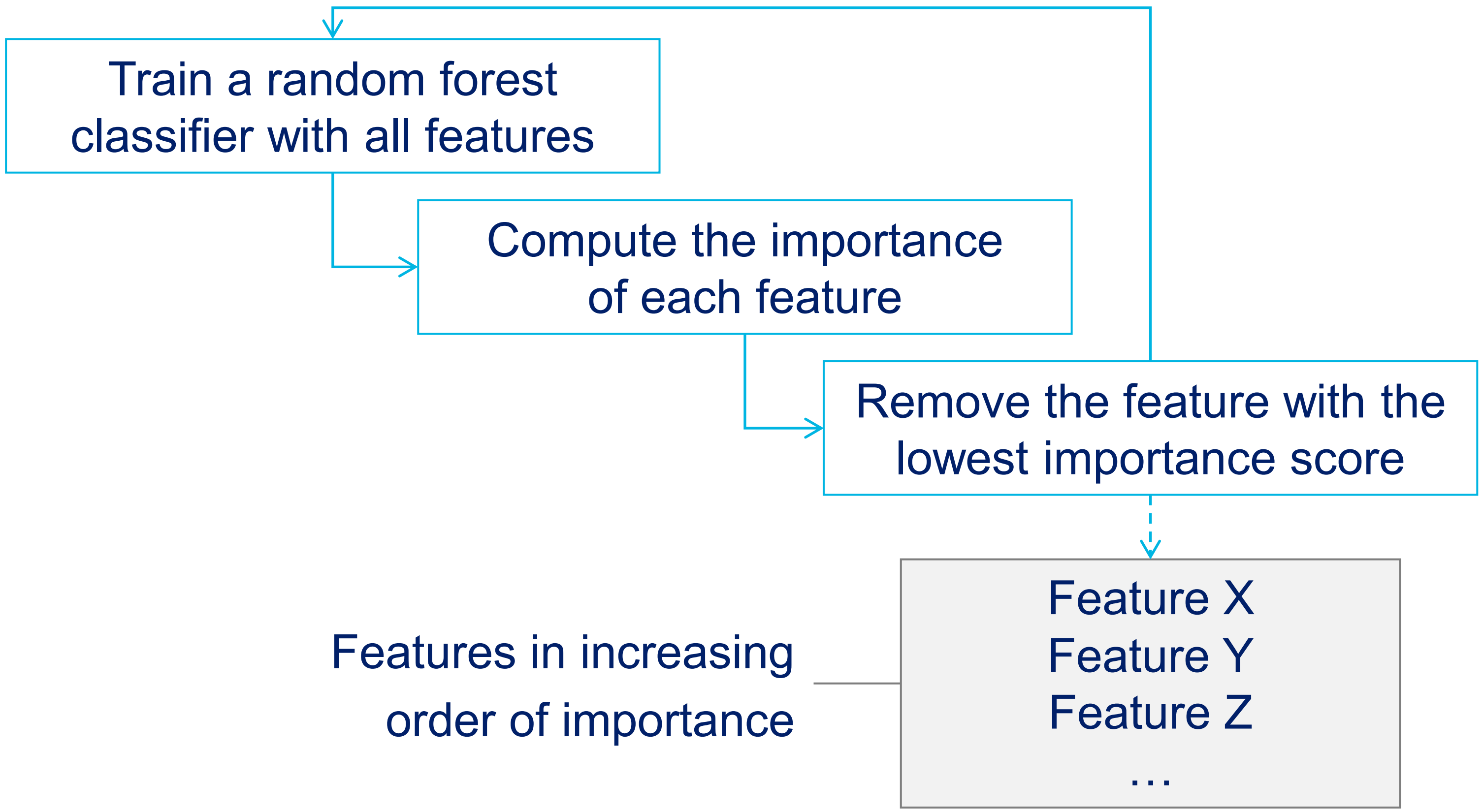
- Feature computation
- **Feature elimination**
- Feature ranking
- Feature selection

- Remove constant features
- Remove highly correlated features
- Remove features with a low RMI (i.e., features that do not contain information about the label)

CYCON

We rank features through recursive feature elimination

- Feature computation
- Feature elimination
- **Feature ranking**
- Feature selection



CYCON

We focus on time-independent features because they are less affected by the environment

- Feature computation
- Feature elimination
- Feature ranking
- **Feature selection**

Feature	Average rank	Rank in LS17A	Rank in LS18A	Rank in LS19A	Rank in LS21A
Pkt Len Max	1	8	8	2	5
Init Fwd Win Byts	2	1	18	4	1
Fwd Pkt Len Max	3	7	10	9	4
Bwd Pkt Len Std	4	4	17	8	6
Pkt Len Var	5	2	11	17	7
Bwd Pkt Len Max	6	18	14	1	8
Fwd Pkt Len Std	7	3	13	20	10
Pkt Len Mean	8	13	5	15	13
Bwd Header Len	9	9	4	12	23
Init Bwd Win Byts	10	10	19	7	12

CYCON

We developed two types of models

Flow-based models

Goal is to detect malicious flows

Random forest model
with 10 or 20 features

Trained on A datasets

Host-based models

Goal is to identify infected hosts

Classification using
the flow-based model

CYCON

Background The Locked Shields exercise	Baseline Does existing work generalize?	Approach Towards more robust models	**Results** **Evaluation of our models**	Outlook Future research directions

CYCON

We trained models with the top 10 or 20 (time-independent) features

Training data

A Generic, 10 Feat.

A Generic, 10 t.-i. Feat.

A Generic, 20 Feat.

A Generic, 20 t.-i. Feat.

CYCON

We evaluate the models on all available datasets

Test data

Training data	LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
A **Generic, 10 Feat.**					
A **Generic, 10 t.-i. Feat.**					
A **Generic, 20 Feat.**					
A **Generic, 20 t.-i. Feat.**					

CYCON

The models generally perform well on data of Country A

Test data

Training data	LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
A Generic, 10 Feat.	0.980	0.991	0.426	0.975	
A Generic, 10 t.-i. Feat.	0.985	0.992	0.554	0.971	
A Generic, 20 Feat.	0.991	0.992	0.621	0.967	
A Generic, 20 t.-i. Feat.	0.992	0.993	0.638	0.989	

F1 score

CYCON

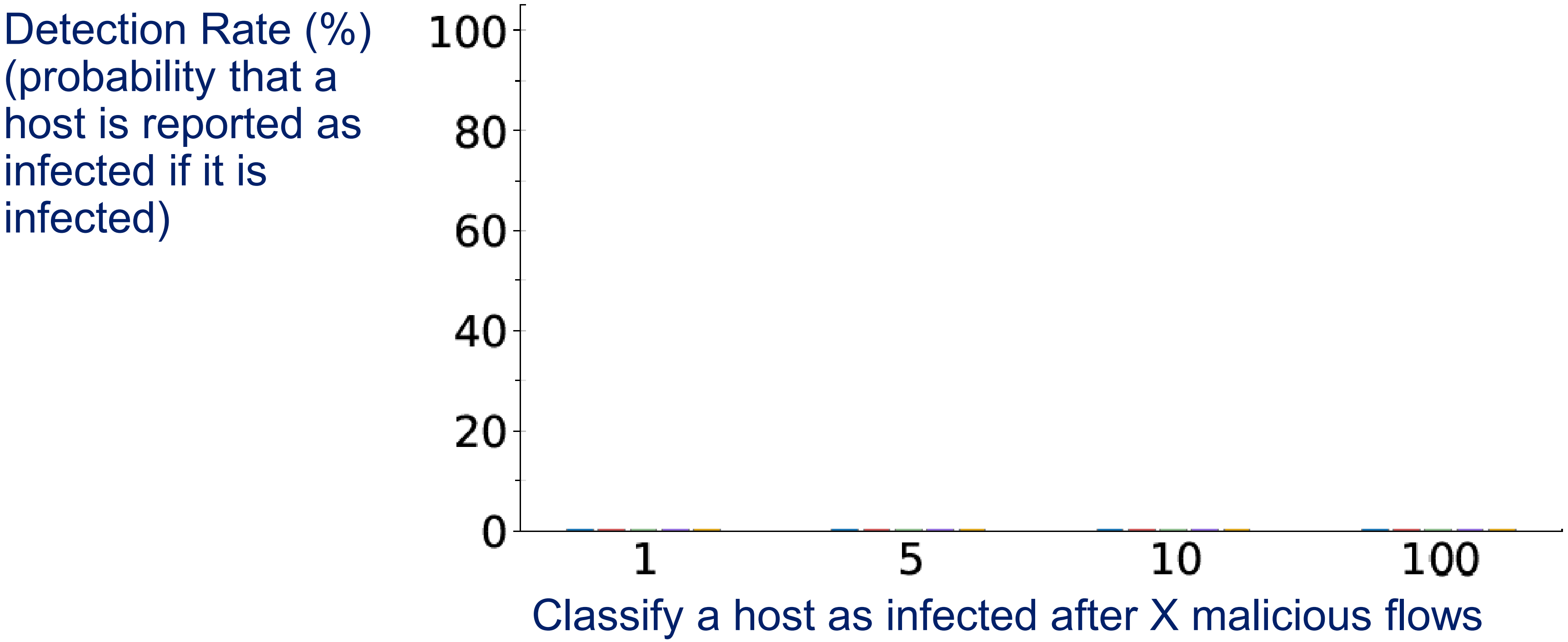
The models do not perform well on data of Country B

Training data	Test data: LS17 A	LS18 A	LS19 A	LS21 A	LS21 B
A Generic, 10 Feat.	0.980	0.991	0.426	0.975	0.116
A Generic, 10 t.-i. Feat.	0.985	0.992	0.554	0.971	0.162
A Generic, 20 Feat.	0.991	0.992	0.621	0.967	0.135
A Generic, 20 t.-i. Feat.	0.992	0.993	0.638	0.989	0.185

F1 score

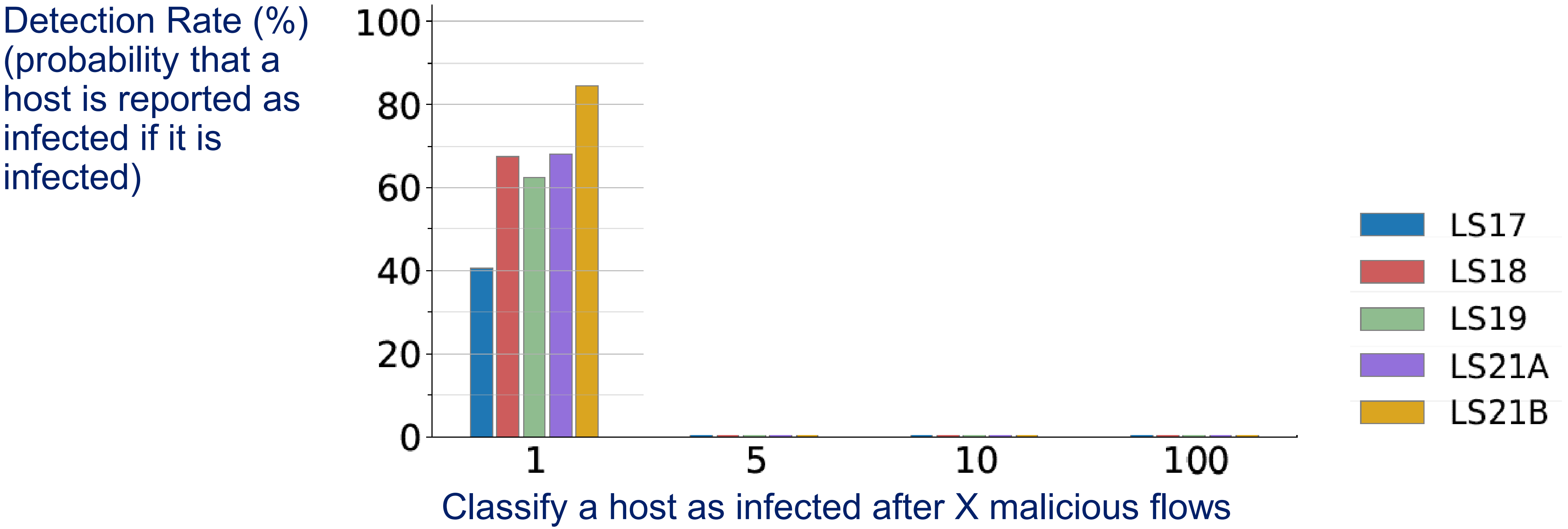
CYCON

The host-based model identifies compromised hosts



CYCON

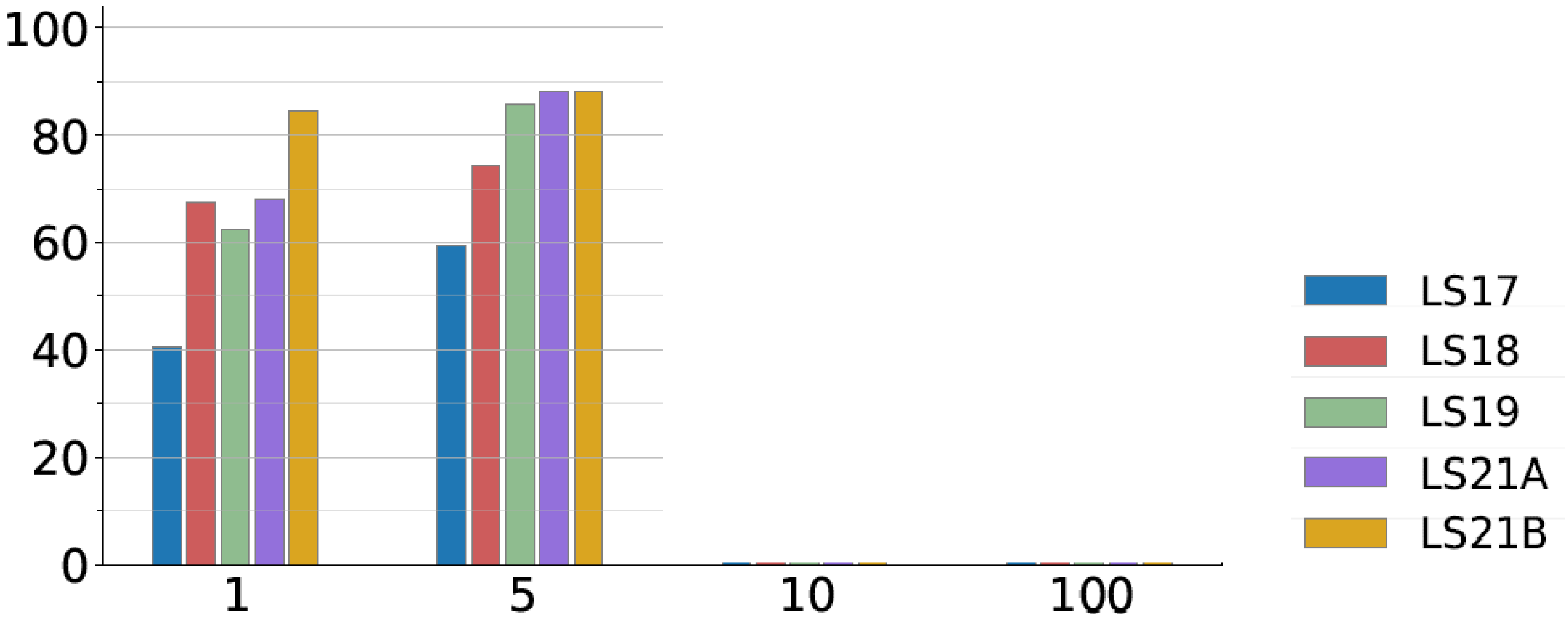
Reporting a host as compromised after 1 flow is prone to errors



CYCON

Waiting for multiple malicious flows makes the detection more robust

Detection Rate (%) (probability that a host is reported as infected if it is infected)

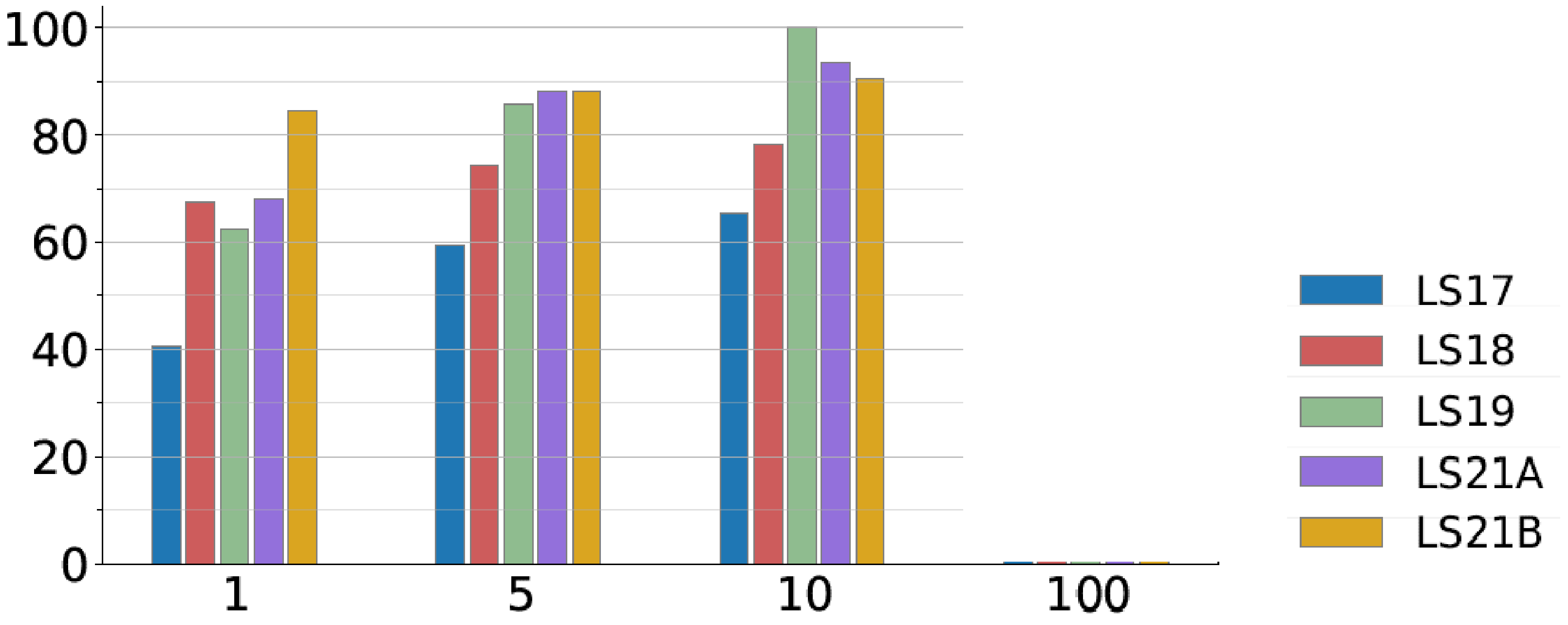


Classify a host as infected after X malicious flows

CYCON

Waiting for multiple malicious flows makes the detection more robust

Detection Rate (%) (probability that a host is reported as infected if it is infected)

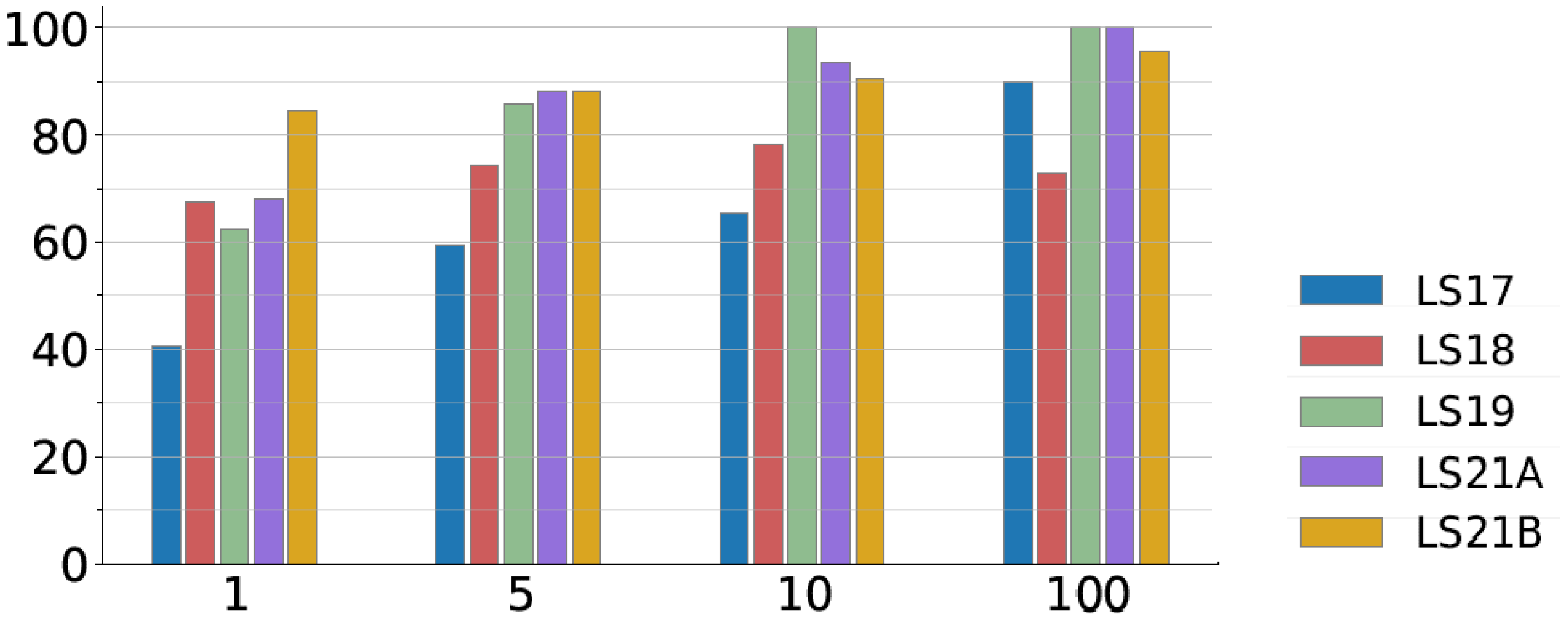


Classify a host as infected after X malicious flows

CYCON

Waiting for multiple malicious flows makes the detection more robust

Detection Rate (%) (probability that a host is reported as infected if it is infected)



Classify a host as infected after X malicious flows

CYCON

Background
The Locked Shields exercise

Baseline
Does existing work generalize?

Approach
Towards more robust models

Results
Evaluation of our models

Outlook
Future research directions

Robust traffic classification across multiple environments remains challenging

Recently published work shows that many models classify based on the “background”

- Automatically generated explanations of ML models show problems in the datasets
- Example: VPN vs. Non-VPN classification based on three bytes (that have nothing to do with VPN or Non-VPN traffic):

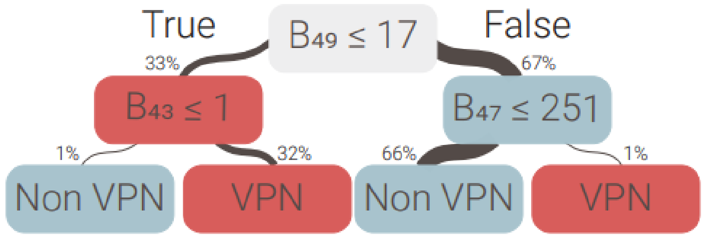
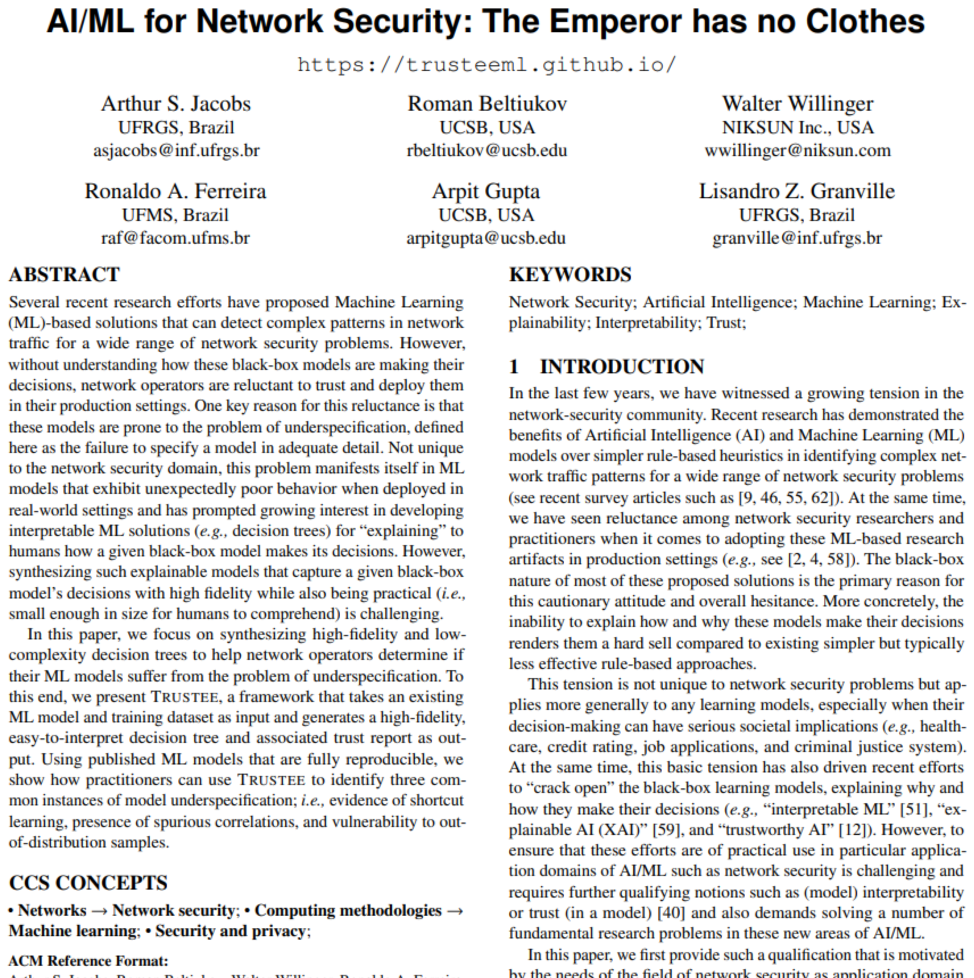


Figure 2: Decision tree for 1D-CNN model. The percentage of samples that follow each branch is presented above each node. Line widths are proportional to the percentage of samples.



AI/ML for Network Security: The Emperor has no Clothes

https://trusteeml.github.io/

Arthur S. Jacobs, UFRGS, Brazil, asjacobs@inf.ufrgs.br
Roman Beltiukov, UCSB, USA, rbeltiukov@ucsb.edu
Walter Willinger, NIKSUN Inc., USA, wwillinger@niksun.com
Ronaldo A. Ferreira, UFMS, Brazil, raf@facom.ufms.br
Arpit Gupta, UCSB, USA, arpitgupta@ucsb.edu
Lisandro Z. Granville, UFRGS, Brazil, granville@inf.ufrgs.br

ABSTRACT

Several recent research efforts have proposed Machine Learning (ML)-based solutions that can detect complex patterns in network traffic for a wide range of network security problems. However, without understanding how these black-box models are making their decisions, network operators are reluctant to trust and deploy them in their production settings. One key reason for this reluctance is that these models are prone to the problem of underspecification, defined here as the failure to specify a model in adequate detail. Not unique to the network security domain, this problem manifests itself in ML models that exhibit unexpectedly poor behavior when deployed in real-world settings and has prompted growing interest in developing interpretable ML solutions (*e.g.*, decision trees) for “explaining” to humans how a given black-box model makes its decisions. However, synthesizing such explainable models that capture a given black-box model’s decisions with high fidelity while also being practical (*i.e.*, small enough in size for humans to comprehend) is challenging.

In this paper, we focus on synthesizing high-fidelity and low-complexity decision trees to help network operators determine if their ML models suffer from the problem of underspecification. To this end, we present TRUSTEE, a framework that takes an existing ML model and training dataset as input and generates a high-fidelity, easy-to-interpret decision tree and associated trust report as output. Using published ML models that are fully reproducible, we show how practitioners can use TRUSTEE to identify three common instances of model underspecification; *i.e.*, evidence of shortcut learning, presence of spurious correlations, and vulnerability to out-of-distribution samples.

CCS CONCEPTS

• Networks → Network security; • Computing methodologies → Machine learning; • Security and privacy;

ACM Reference Format:

KEYWORDS

Network Security; Artificial Intelligence; Machine Learning; Explainability; Interpretability; Trust;

1 INTRODUCTION

In the last few years, we have witnessed a growing tension in the network-security community. Recent research has demonstrated the benefits of Artificial Intelligence (AI) and Machine Learning (ML) models over simpler rule-based heuristics in identifying complex network traffic patterns for a wide range of network security problems (see recent survey articles such as [9, 46, 55, 62]). At the same time, we have seen reluctance among network security researchers and practitioners when it comes to adopting these ML-based research artifacts in production settings (*e.g.*, see [2, 4, 58]). The black-box nature of most of these proposed solutions is the primary reason for this cautionary attitude and overall hesitance. More concretely, the inability to explain how and why these models make their decisions renders them a hard sell compared to existing simpler but typically less effective rule-based approaches.

This tension is not unique to network security problems but applies more generally to any learning models, especially when their decision-making can have serious societal implications (*e.g.*, health-care, credit rating, job applications, and criminal justice system). At the same time, this basic tension has also driven recent efforts to “crack open” the black-box learning models, explaining why and how they make their decisions (*e.g.*, “interpretable ML” [51], “explainable AI (XAI)” [59], and “trustworthy AI” [12]). However, to ensure that these efforts are of practical use in particular application domains of AI/ML such as network security is challenging and requires further qualifying notions such as (model) interpretability or trust (in a model) [40] and also demands solving a number of fundamental research problems in these new areas of AI/ML.

In this paper, we first provide such a qualification that is motivated by the needs of the field of network security as application domain

Directions for future research

Better datasets

Better features

Better models

Understand limitations

Directions for future research

- Better datasets
- Better features
- Better models
- Understand limitations

- Today: hard (or impossible) to distinguish between malicious activities and “background”
- Large synthetic datasets would allow to learn the actual characteristics of malicious traffic
- Missing labels (attack traffic that is marked as normal traffic) might confuse a model
- Virtual environments are not representative w.r.t. many features

Directions for future research

- Better datasets
- **Better features**
- Better models
- Understand limitations

- Currently, the focus in on flow-based features. But other abstractions would provide additional information.
- For example: Host-based features to capture periodic connections to CnC server

Directions for future research

Better datasets

Better features

Better models

Understand limitations

- Our focus was on random forest models (as in previous work)
- Other types of models might perform better
- But main limitation is likely the amount/quality of the datasets

CYCON

Directions for future research

Better datasets

Better features

Better models

Understand limitations

- Currently, we assume that the attackers do not try to circumvent our model
- Realistically, attackers would adapt their behavior depending on the defense tools
- Many features can be manipulated in order to conceal malicious traffic

Feature	Average rank
Pkt Len Max	1
Init Fwd Win Byts	2
Fwd Pkt Len Max	3
Bwd Pkt Len Std	4
Pkt Len Var	5
Bwd Pkt Len Max	6
Fwd Pkt Len Std	7
Pkt Len Mean	8
Bwd Header Len	9
Init Bwd Win Byts	10

CYCON

Thank you for your attention

Background	Baseline	Approach	Results	Outlook
The Locked Shields exercise	Does existing work generalize?	Towards more robust models	Evaluation of our models	Future research directions

CYCON

Roland Meier, roland.meier@ar.admin.ch

CYCON