AUSTRIAN INSTITUTE OF TECHNOLOGY

W|W|T|F
VIENNA SCIENCE AND TECHNOLOGY FUND

Machine Learning based Approaches for Anomaly Detection and Classification in Cellular Networks

Pedro Casas (*), Pierdomenico Fiadino, Alessandro D'Alconzo

(*) AIT Austrian Institute of Technology, Vienna

Anomaly Detection in Cellular Traffic

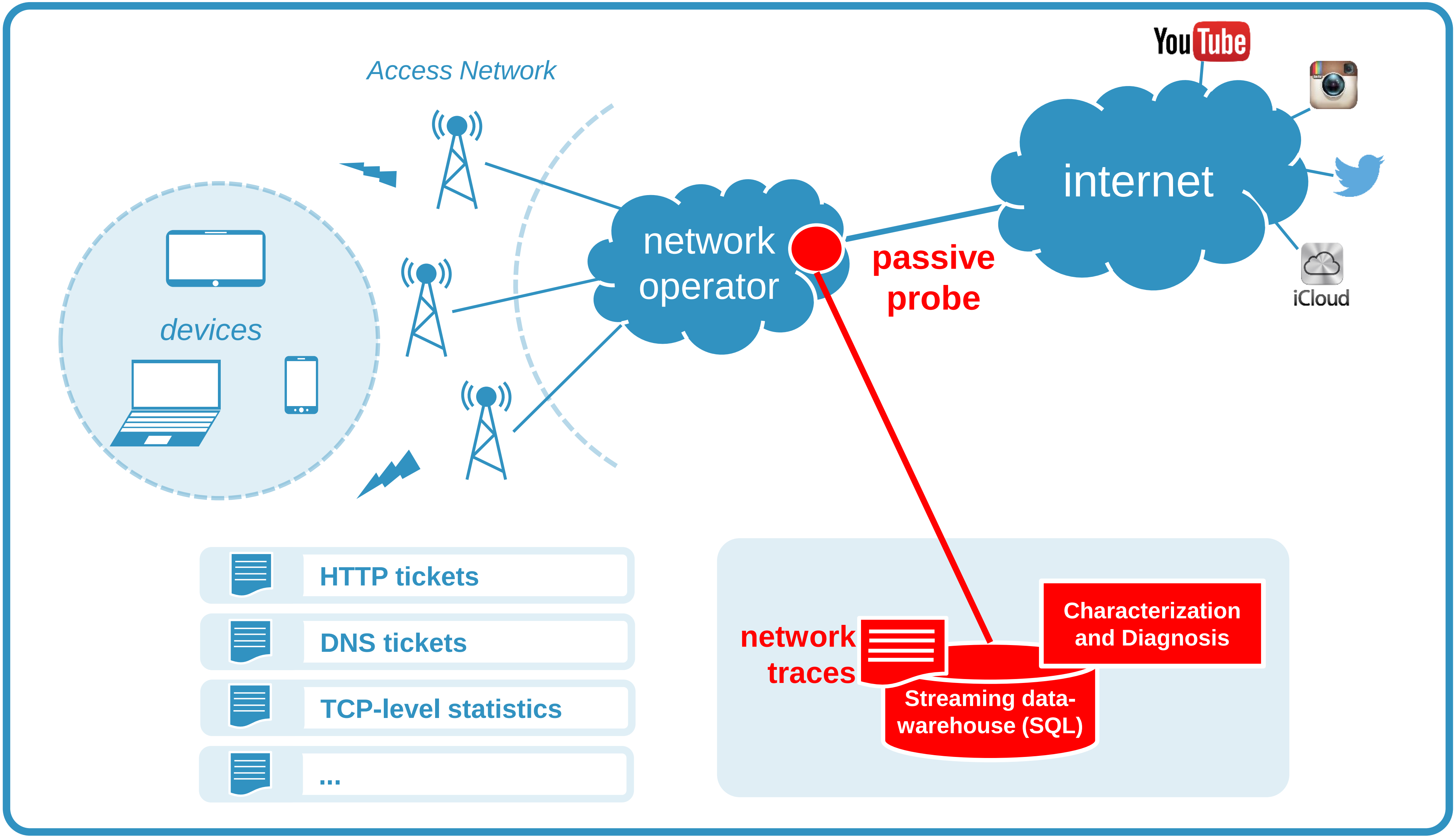
We study the problem of **detecting and classifying** certain types of network **anomalies** in cellular networks, **relying on Machine Learning approaches**

Outline of the Talk

- ***Cellular Network Monitoring and Synthetic Datasets***
- ***Anomaly Detection and Classification Approaches***
- ***Evaluation Results***
- ***Impact of Feature Selection and OOS Testing***

cellular network monitoring

Passive Measurements at Core of EU Cellular ISP



Automatic Diagnosis Framework

key idea 1

distinguish between:

- **symptomatic features**
  notify occurrence of an anomaly
- **diagnostic features**
  provide context for diagnosis

key idea 2

observe significant changes in multiple traffic features

key idea 3

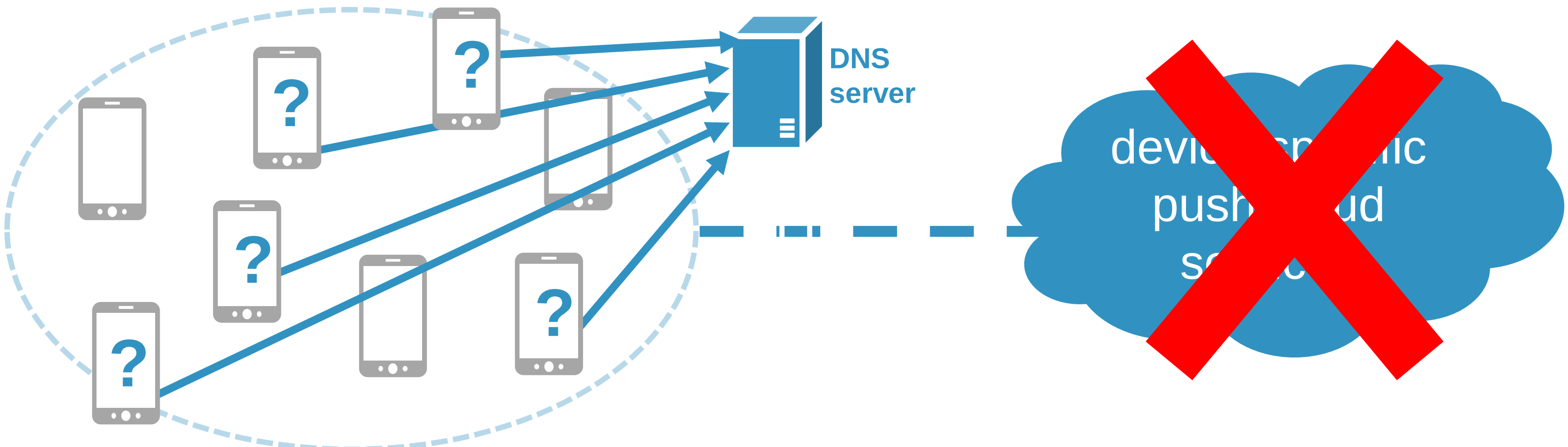
keep status of changes, classify for troubleshooting

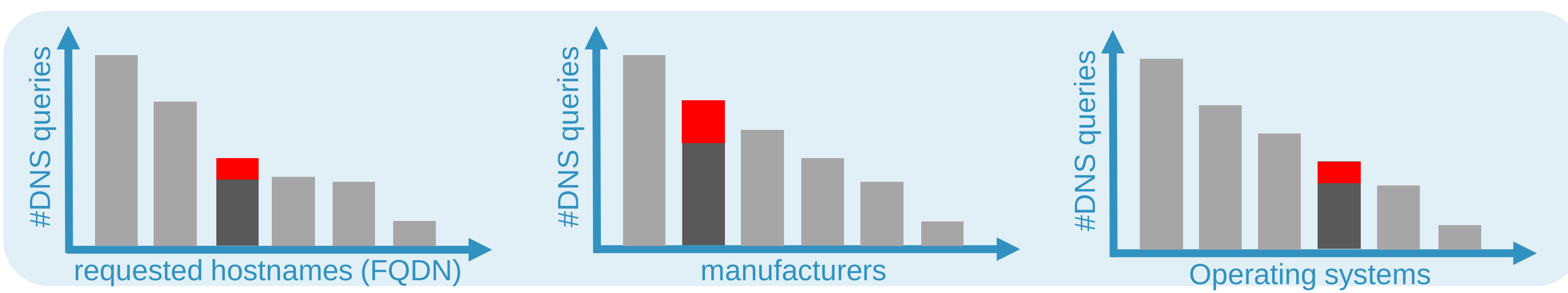
Service Anomalies Visible in DNS Traffic

Device-specific anomalies affecting sub-populations

Observed multiple instances in few months

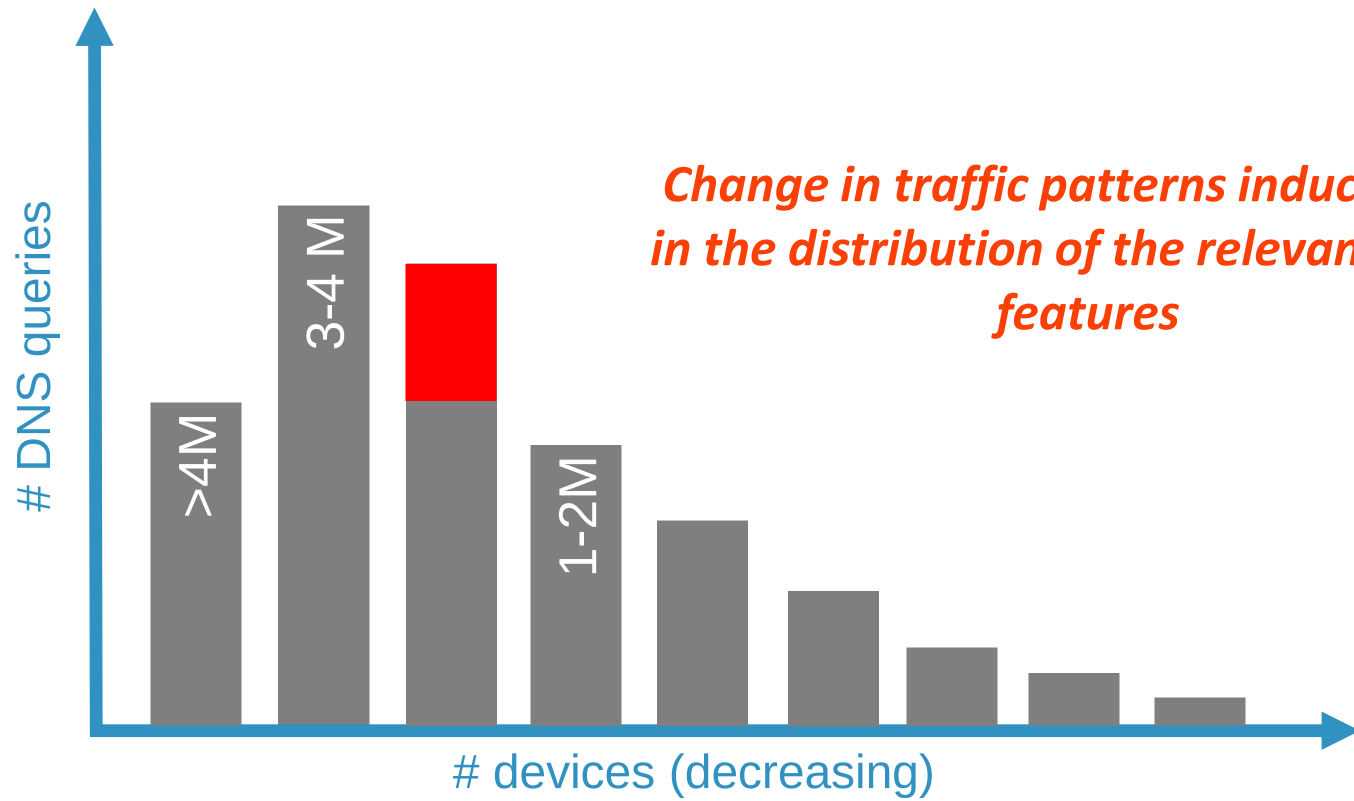
Impacting operator-managed DNS servers and signaling plane





Traffic Feature Distributions for Change Detection

Empirical distribution of # devices across DNS query counts (binning)



Change in traffic patterns induce shifts in the distribution of the relevant traffic features

Symptomatic and Diagnostic Features

- ***Symptomatic Feature (the trigger)*** → *distribution of # devices across DNS query counts (10' time-bin basis)*
  - *counting of devices issuing a given number of DNS queries within each time-bin.*

- ***Diagnostic Features (troubleshooting-support)*** → *distribution of # devices across field in Tab. I (10' time-bin basis)*

Field Name	**Description**
Manufacturer	Device manufacturer
OS	Device operating system
APN	Access Point Name
FQDN	Fully Qualified Domain Name of remote service
Error Flag	Status of the DNS transaction

Table I
FEATURES USED IN THE ANALYSIS.

Anomaly Templates and Synthetic Datasets

- *Privacy: we use* ***synthetically generated datasets****, derived from the real celular ISP measurements (details in the paper)*
- ***Anomaly Templates****, derived from the real anomalies observed in the celular traffic → in this paper, anomaly types E1, E2 and E3*
- ***Traffic measurements*** *collected during* ***6-months in 2014***
- ***Evaluation labelled dataset****: 1 month of normal operation traffic, and 16 different anomaly instances of E1, E2 and E3 types, with different intensity (number of involved devices varies from 0.1% to 20%)*

Type	E_1	E_2	E_3
Start time t_1	9:00	13:00	18:00
Duration d	2h	1 day	1h
Involved devices D	10%	5%	3%
Back-off time	5 sec	180 sec	20 sec
Manufacturer	single popular	multiple	multiple
OS	single	single	multiple
Error flag	+5% timeout	—	—
FQDN	top-2LD	top-2LD	top-2LD

Table III
ANOMALOUS DNS TRAFFIC FEATURES FOR TYPES E_1, E_2, E_3.

anomaly detection & classification

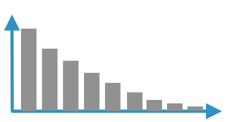
Statistical Anomaly Detection

Distribution-based detector

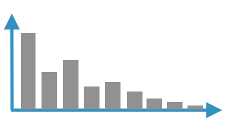
Consider whole empirical distribution
instead of just its entropy (as usually done)

Compute distance between distribution and normality reference
i.e. a set of historical anomaly-free distributions

Self-adapting sliding window algorithm for reference identification
my extension to the algorithm (originally based on simple sliding window mechanism)



current distribution at t_i



reference distribution r

Distribution distance metric

$$D(p\,||\,q) = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)}$$

Kullback-Leibler divergence

$$L(p,q) = \frac{D(p\,||\,q)}{H(p)} + \frac{D(q\,||\,p)}{H(q)}$$

Entropy normalized KL metric

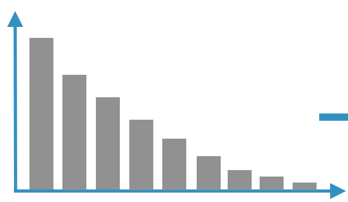
$D_f(t_i) = 3$

Measures how different current distribution is from normality

Entropy–based Anomaly Detection

Entropy-based Detection

Represent entire distribution as Entropy



feature *f* distribution at t_0
2015-08-18 08:00

$$H(x) = -\sum_{i=1}^{n} p(x_i) log(p(x_i))$$

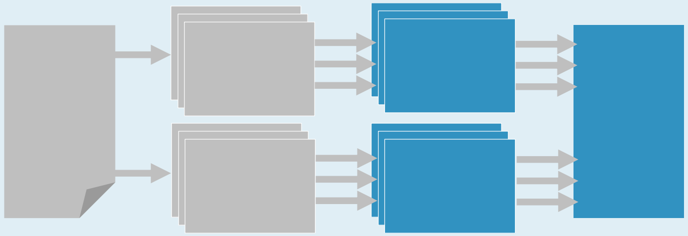
Entropy

$H_f(t_0) = 0.42$

Entropy captures the dispersion of the corresponding distribution

A change in traffic patterns induces a change (spike/notch) in the entropy of the relevant traffic feature

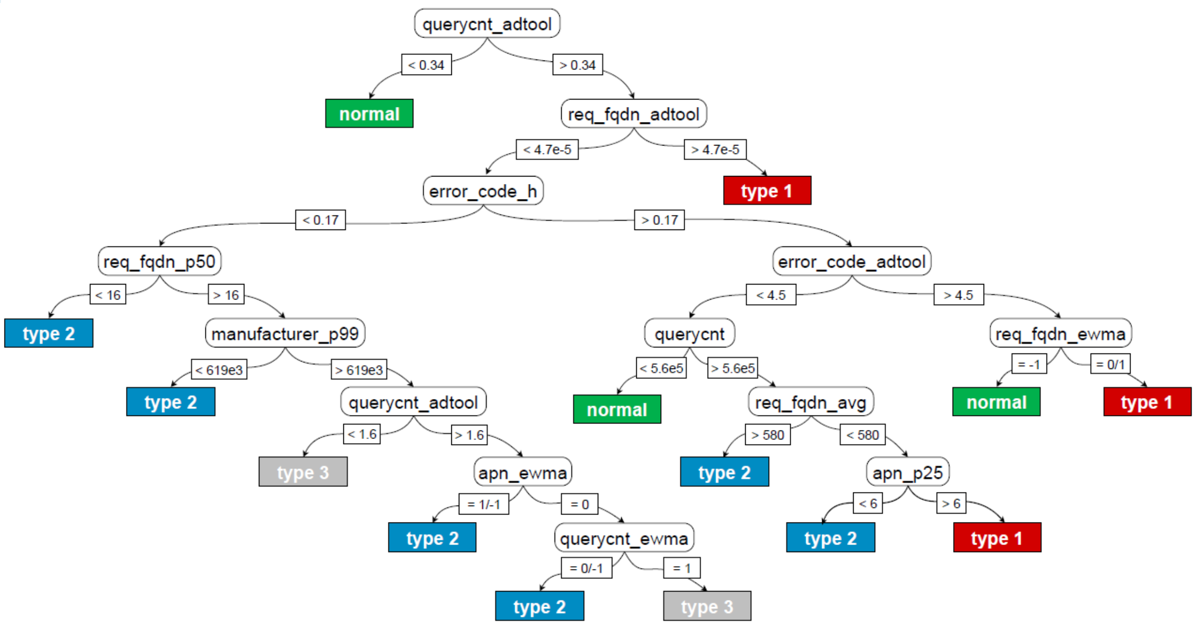
Machine – Learning based Anomaly Detection and Classification



Support Signal Correlation through ML

- Classify detected anomalies (assign event lable)
- Enrich final event reporting and support troubleshooting

Supervised classification techniques



Evaluation on synthetic dataset (from real DNS anomaly templates)

Datasets includes detector outputs as features

Several tested algorithms (SVN, NB, DT, RF, MLP)

C4.5 (Decision Tree) best performing*

*Considering feature selection approaches

C4.5 Decision Tree-based Detection and Classification

- *A decisi … instances by* ***repe*** *… a* ***tree with lea … e.***
- *They are … speed is paramo…*
- *They are … g rules.*
- *They* ***ex|*** *… as the learning … on.*
- *They ter… **oisy or loosely …***

Field	Feature	Description
DNS_query	querycnt	total num of DNS requests
APN	apn_h	$H(\text{APN})$
	apn_avg	$\overline{\text{APN}}$
	apn_p{99,75,50,25,05}	percentiles
Error_flag	error_code_h	$H(\text{Error\_flag})$
	error_code_avg	$\overline{\text{Error\_flag}}$
	error_code_p{99,75,50,25,05}	percentiles
Manufacturer	manufacturer_h	$H(\text{Manufacturer})$
	manufacturer_avg	$\overline{\text{Manufacturer}}$
	manufacturer_p{99,75,50,25,05}	percentiles
OS	os_h	$H(\text{OS})$
	os_avg	$\overline{\text{OS}}$
	os_p{99,75,50,25,05}	percentiles
FQDN	req_fqdn_h	$H(\text{FQDN})$
	req_fqdn_avg	$\overline{\text{FQDN}}$
	req_fqdn_p{99,75,50,25,05}	percentiles

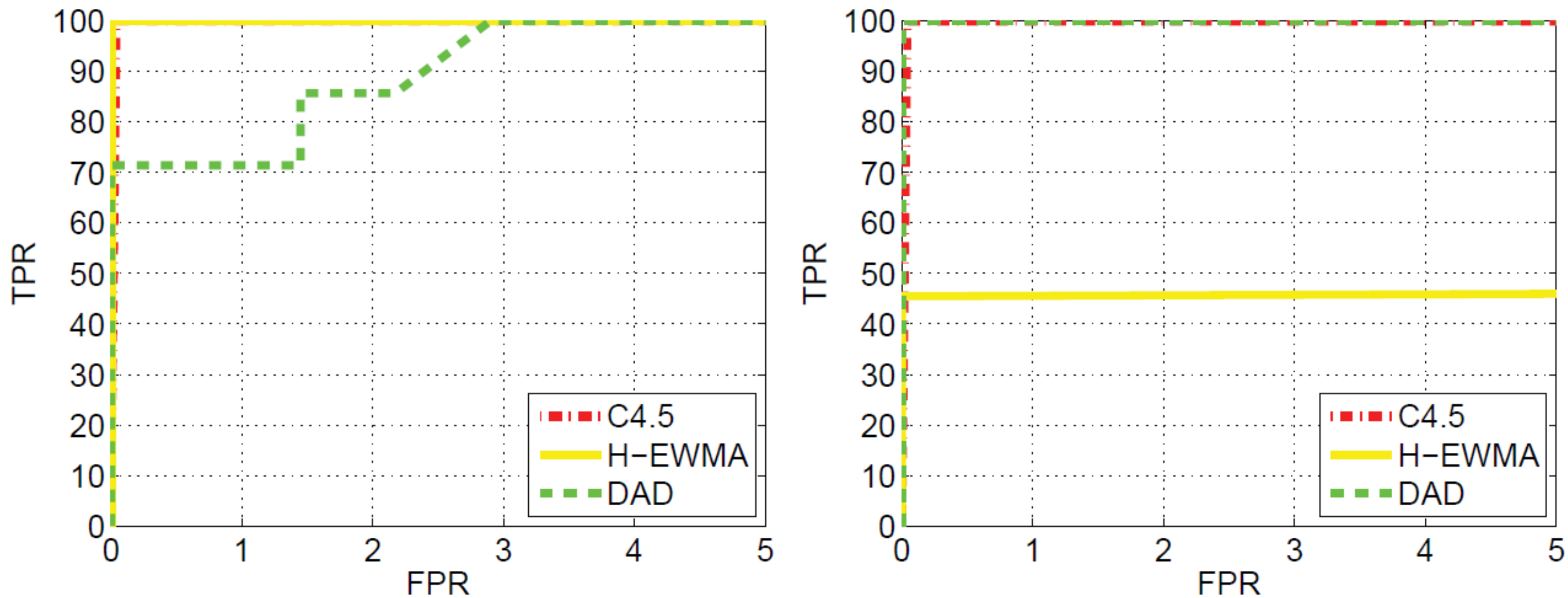
Table III
INPUT FEATURES FOR THE C4.5 DT-BASED DETECTOR/CLASSIFIER.

- ***We additionally use the output of the statistical and entropy-based detectors as input for anomaly classification purposes***

Evaluation Results

Anomaly Detection: C4.5 vs Statistical & Entropy

- *First evaluation:* ***C4.5 with full-input features*** *(Tab. III) vs Distribution-based AD (DAD) and Entropy-based (H-EWMA) for* ***E1 and E2 anomalies***
- *DAD and H-EWMA working only on symptomatic feature*
- *Take away:* ***the C4.5 has comparable detection capabilities to SotA ADs***

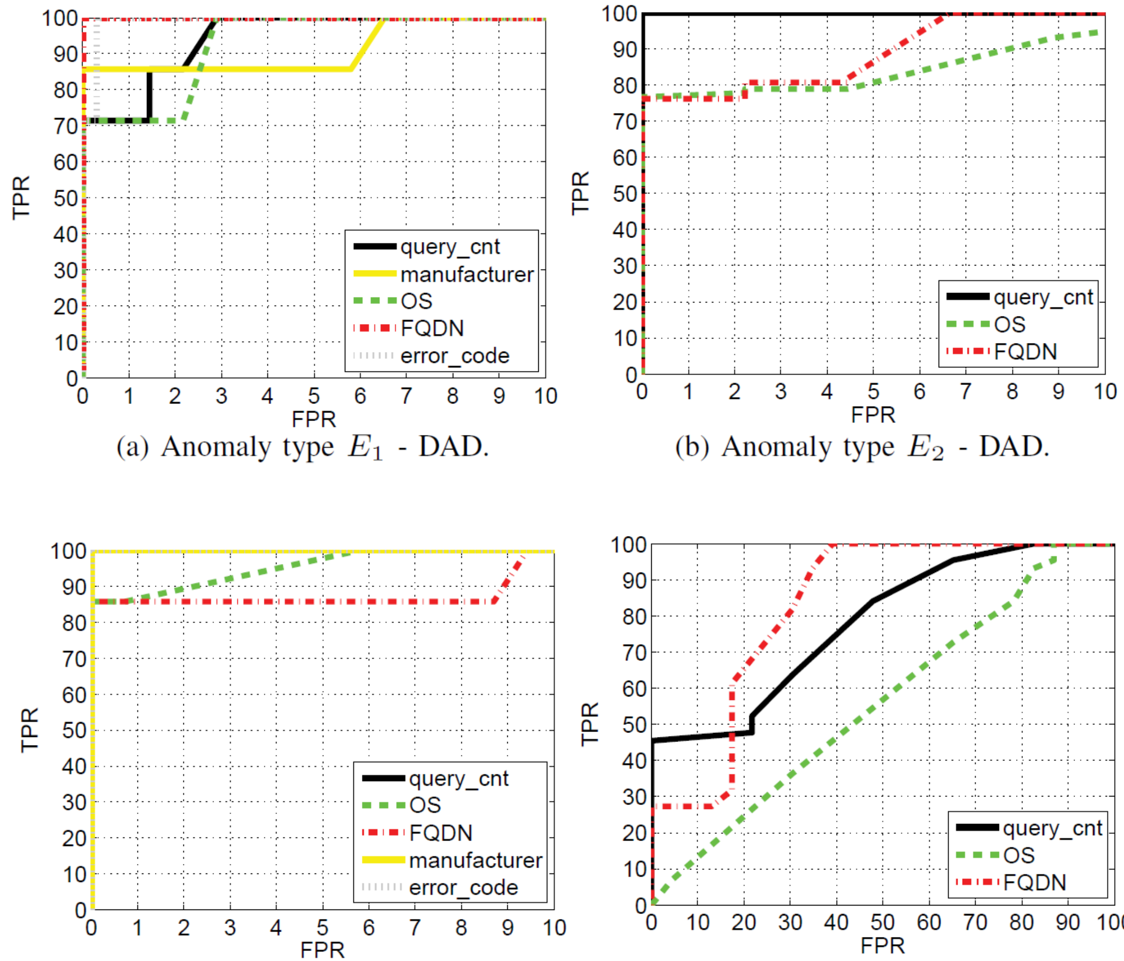


(a) Anomaly type E_1.
(b) Anomaly type E_2.

Figure 1. ROC curves for the detection of anomalies type E_1 and E_2.

Performance of DAD and H-EWMA with other Features

- *We also* ***evaluate DAD and H-EWMA with other input features****, to be closer to the C4.5 inout space*
- *Conclusions remain the same*
- *Note* ***that H-EWMA completely fails to detect the E2 anomalies*** *(supremacy of DAD-like approaches)*

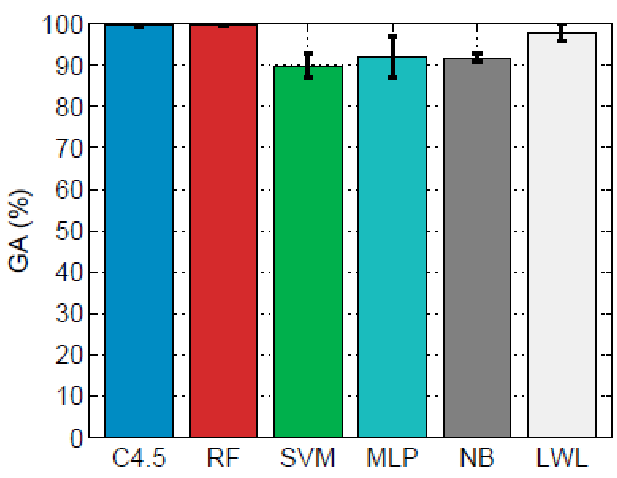


(a) Anomaly type E_1 - DAD. (b) Anomaly type E_2 - DAD.

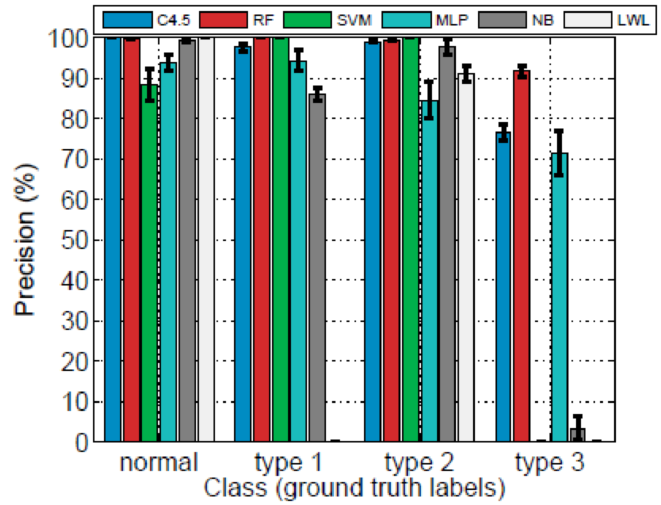
(c) Anomaly type E_1 - H-EWMA. (d) Anomaly type E_2 - H-EWMA.

Figure 2. ROC curves for the detection of anomalies type E_1 and E_2 for DAD and H-EWMA anomaly detectors, considering all the impacted features.

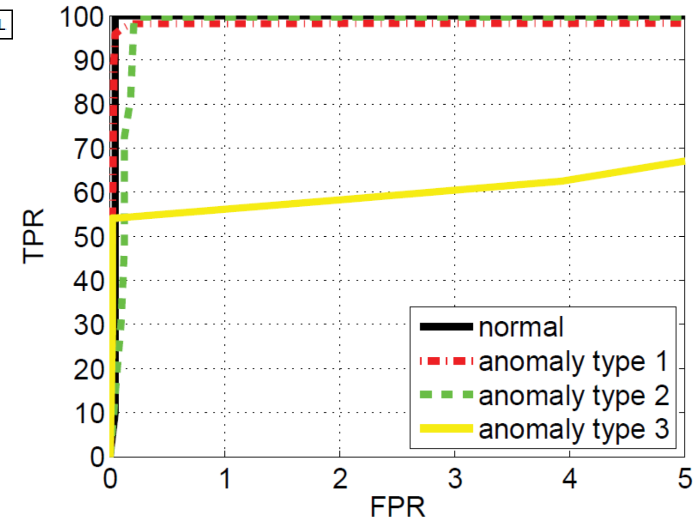
Machine-learning based Classification Benchmarking



(a) Global Accuracy.



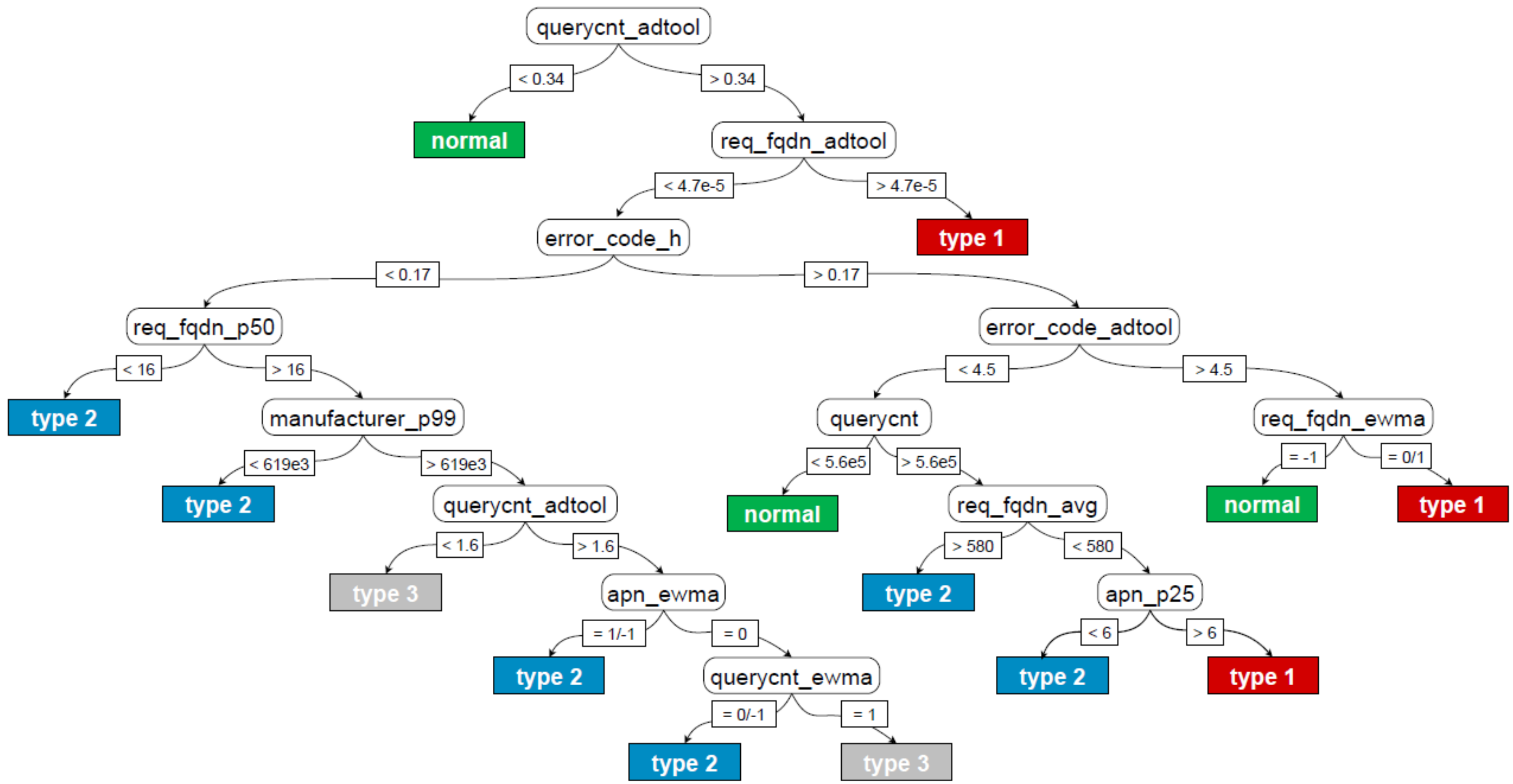
(b) Precision.



Per-class ROC curves for the C4.5 tree.

- *Compare **C4.5** to different ML-based classifiers (**SVM, ANN, Random Forrest, Naïbe Bayes, LWL**)*
- ***Classification Accuracy, Precision, and Recall** for normal operation instances and anomaly-types E1, E2, E3.*
- *The **performance of C4.5 DT is almost perfect** for normal traffic and anomalies of type E1 and E2, but **quality significantly drops for the anomalies of type E3** (also the RF fails)*

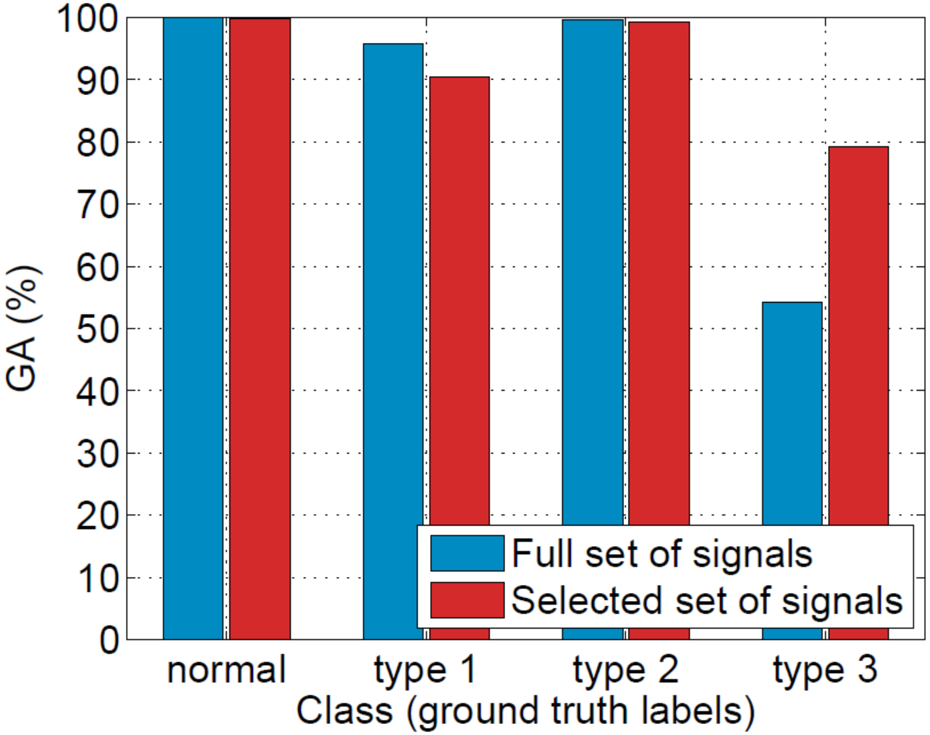
C4.5 Decision Tree-based Detection and Classification



- *Pruned C4.5 DT model for anomaly diagnosis (classification).*
- *The tree fails to track E3 anomalies → **the issue can be solved by** performing pre-filtering on the input features, by **feature selection***

impact of feature selection and OOS testing

Improving Performance for E3 type Anomalies

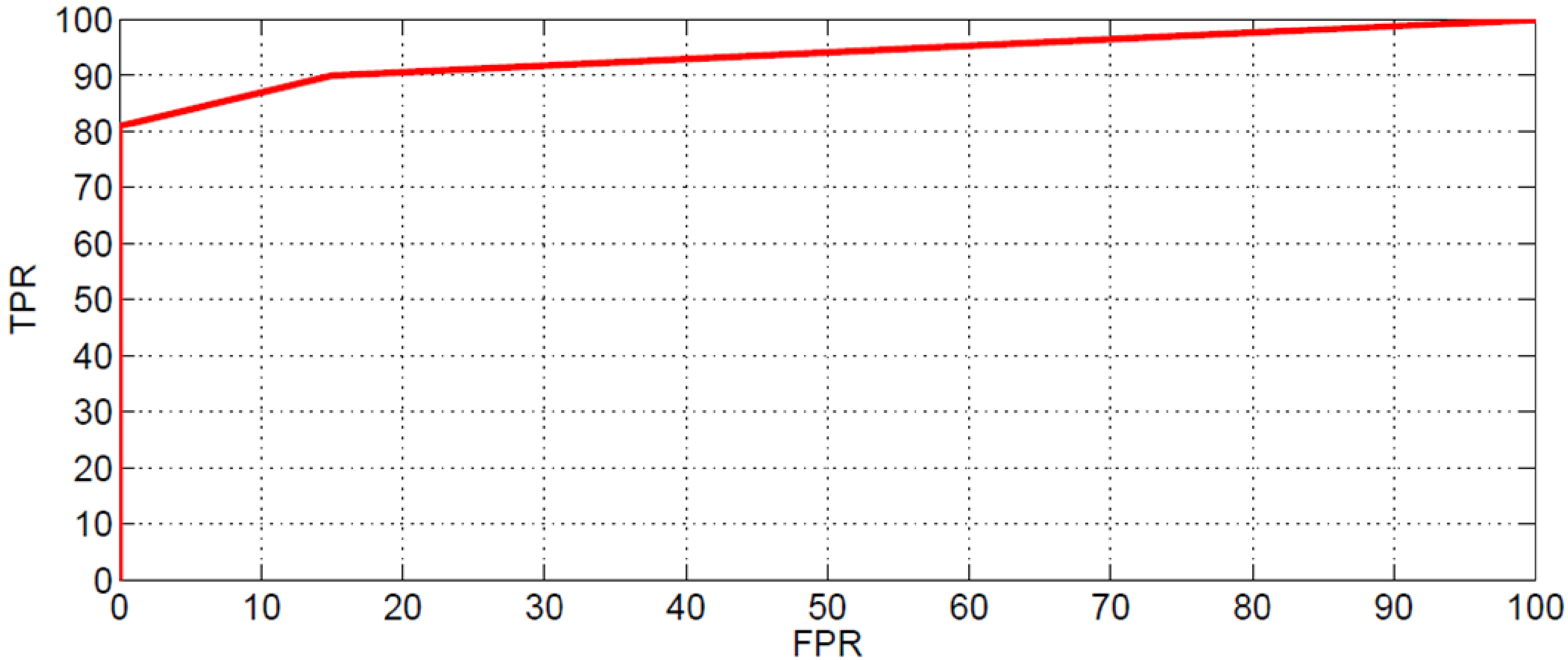


Improving accuracy by feature selection

- ***Irrelevant features introduce noise i****n the classification process*
- ***Select the most relevant ones by correlation****-based approaches*
- *Using* ***Best-First search****: greedy exploration with back-tracking*

- ***Selected features are highly correlated to E3 anomalies***
- *Performance increases for E3 type, with a slight reduction in E1*

Generalization of Results – Out of Sample Testing



- *OOS testing with* ***anomalies of type E4*** *(**flashcrowd-generated**)*
- *The C4.5 model is trained with instances of E1, E2, and E3 only*
- *Performance slightly degrades, but* ***the underlying characteristics of the DNS anomaly class are captured → trees are powerful for generalization***

Thanks for Your Attention!

Pedro Casas
pedro.casas@ait.ac.at