TU Delft – July 5th, 2023

IEEE European Symposium on Security and Privacy

SoK: Pragmatic Assessment of Machine Learning for Network Intrusion Detection

Giovanni Apruzzese, Pavel Laskov, Johannes Schneider

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

GOAL

Changing the way research on Network Intrusion Detection (NID) based on Machine Learning (ML) is carried out.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

GOAL

Changing the way research on Network Intrusion Detection (NID) based on Machine Learning (ML) is carried out.

WHY?

In research (20 years ago)…

An application of **machine learning** to network **intrusion detection**
C Sinclair, L Pierce, S Matzner - Proceedings 15th annual …, 1999 - ieeexplore.ieee.org
… **machine learning** techniques, we also intend to research other artificial intelligence methods applicable to **intrusion detection**… can **detect** will improve as our **machine learning** techniques …
☆ Save 99 Cite Cited by 416 Related articles All 12 versions

[PDF] HIDE: a hierarchical **network intrusion detection** system using statistical preprocessing and **neural network** classification
Z Zhang, J Li, CN Manikopoulos… - Proc. IEEE Workshop …, 2001 - cs.rhodes.edu
… **Intrusion DEtection** (HIDE) system, which detects **network**-based attacks as anomalies using statistical preprocessing and **neural network** … We tested five different types of **neural network** …
☆ Save 99 Cite Cited by 380 Related articles All 10 versions

Intrusion detection using **neural networks** and support vector machines
S Mukkamala, G Janoski, A Sung - … on **Neural Networks**. IJCNN' …, 2002 - ieeexplore.ieee.org
… standard benchmark for **intrusion detection** evaluations. Our goal for **intrusion detection** is to **detect** both anomalies and misuses. The approach is to train the **neural networks** or support …
☆ Save 99 Cite Cited by 1159 Related articles All 5 versions

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

GOAL

Changing the way research on Network Intrusion Detection (NID) based on Machine Learning (ML) is carried out.

WHY?

"Application of ML in intrusion detection has been uneven at best, with deep and widespread (and generally justified) skepticism among subject matter experts" [9].

Markus de Shon (Lead of Detection Engineering at **NetFlix**)

...in practice (in 2020s)

According to a recent survey, over 75% of companies employ ML solutions for network security [65]. Most of such companies, however, *delegate* their cybersecurity to third-party vendors [66]. Indeed, several commercial products for NID actively leverage ML (e.g., [67]–[69]). Yet, all such products adopt ML methods that are decades old and mostly in their unsupervised form (e.g., the one-class SVM of [50] was proposed in 2002 [70]). Simply put, the integration of research endeavours into operational environments is slow in the context of ML-NIDS.

UNIVERSITÄT LIECHTENSTEIN

[9] M. De Shon. "Information Security Analysis as Data Fusion." In *IEEE International Conference on Information Fusion*, 2019
[50] D. Ucci et al. "Near-real-time Anomaly Detection in Encrypted Traffic using Machine Learning Techniques". *IEEE Symposium Series on Computational Intelligence*, 2021

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(Meanwhile, in Computer Vision…)

Hey, I have a new algorithm to generate synthetic images!

2014

2017

2022

UNIVERSITÄT
LIECHTENSTEIN

Source: Nicholas Carlini, "A crisis in adversarial machine learning" Art of Robustness, 2022.

…BUT WHY SO?

Lack of an "Universal" Dataset

2010 IEEE Symposium on Security and Privacy

**Outside the Closed World:
On Using Machine Learning For Network Intrusion Detection**

Robin Sommer
International Computer Science Institute, and
Lawrence Berkeley National Laboratory

Vern Paxson
International Computer Science Institute, and
University of California, Berkeley

This was the first SoK!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

…BUT WHY SO?

Lack of an "Universal" Dataset

2010 IEEE Symposium on Security and Privacy

Outside the Closed World:
On Using Machine Learning For Network Intrusion Detection

Robin Sommer
International Computer Science Institute, and
Lawrence Berkeley National Laboratory

Vern Paxson
International Computer Science Institute, and
University of California, Berkeley

This was the first SoK!

Instead, we address another shortcoming…

The focus is on the ML model

1. Propose a "new" solution
2. Choose a given metric
3. Show that you "outperform" the state-of-the-art

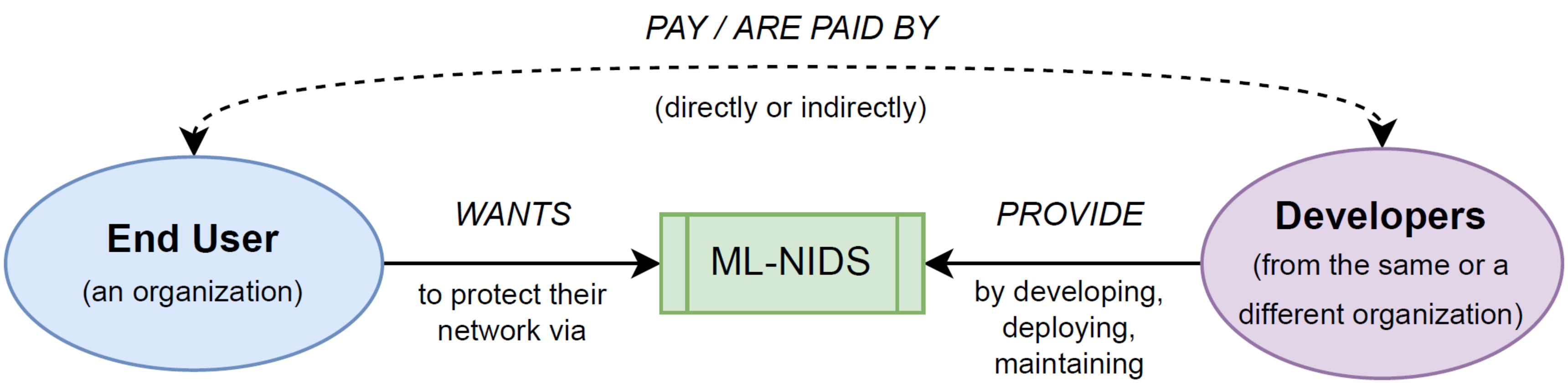
…what about the *rest*?

This may not be what practitioners want!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(1) Practical Factors**



Deployment of ML in NIDS must account for several factors *before* developing any ML model

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(1) Practical Factors** – cont'd

The "value" of an ML model can be seen as a function of five factors:

1. **System Infrastructure**

   (how does the ML model interact with the overarching system?)

An ML model is just a single component within a NIDS

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(1) Practical Factors** – cont'd

The "value" of an ML model can be seen as a function of five factors:

1. **System Infrastructure**

   (how does the ML model interact with the overarching system?)

   An ML model is just a single component within a NIDS

2. **Preprocessing**

   (what data is passed as input to the ML model?)

   There exist dozens of tools to preprocess data

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(1) Practical Factors** – cont'd

The "value" of an ML model can be seen as a function of five factors:

1. **System Infrastructure**
   (how does the ML model interact with the overarching system?)

An ML model is just a single component within a NIDS

2. **Preprocessing**
   (what data is passed as input to the ML model?)

There exist dozens of tools to preprocess data

3. **Data Availability**
   (how much data is required to train the ML model?)

Even "unsupervised" ML algorithms need training data
(sometimes requiring weeks to collect! [75])

UNIVERSITÄT LIECHTENSTEIN

[75] G. Apruzzese et al. "The role of machine learning in cybersecurity." *ACM Digital Threats: Research and Practice*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(1) Practical Factors** – cont'd

The "value" of an ML model can be seen as a function of five factors:

1. **System Infrastructure**
   (how does the ML model interact with the overarching system?)
   An ML model is just a single component within a NIDS

2. **Preprocessing**
   (what data is passed as input to the ML model?)
   There exist dozens of tools to preprocess data

3. **Data Availability**
   (how much data is required to train the ML model?)
   Even "unsupervised" ML algorithms need training data (sometimes requiring weeks to collect! [75])

4. **Hardware**
   (what platform is expected to run the ML model?)
   There can be differences between train and inference hardware

UNIVERSITÄT LIECHTENSTEIN

[75] G. Apruzzese et al. "The role of machine learning in cybersecurity." *ACM Digital Threats: Research and Practice*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(1) Practical Factors** – cont'd

The "value" of an ML model can be seen as a function of five factors:

1. **System Infrastructure**
   (how does the ML model interact with the overarching system?)
2. **Preprocessing**
   (what data is passed as input to the ML model?)
3. **Data Availability**
   (how much data is required to train the ML model?)
4. **Hardware**
   (what platform is expected to run the ML model?)
5. **Unpredictability**
   (how to deal with the concept drift?)

An ML model is just a single component within a NIDS

There exist dozens of tools to preprocess data

Even "unsupervised" ML algorithms need training data (sometimes requiring weeks to collect! [75])

There can be differences between train and inference hardware

The performance **will** deteriorate (eventually)

UNIVERSITÄT LIECHTENSTEIN

[75] G. Apruzzese et al. "The role of machine learning in cybersecurity." *ACM Digital Threats: Research and Practice*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(2) Research Guidelines**

How can researchers meet the needs of practitioners?

1. **System Infrastructure**
   → *Provide a schematic!*

Readers will like it!

UNIVERSITÄT LIECHTENSTEIN

(Some of our guidelines *do* overlap with those of Arp et al. [8])

But some don't!

[8] D. Arp et al. "Dos and don'ts of machine learning in computer security." In *USENIX Security Symp.*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(2) Research Guidelines**

How can researchers meet the needs of practitioners?

1. **System Infrastructure**
   → *Provide a schematic!*

   Readers will like it!

2. **Preprocessing**
   → *Report which tools*

   Try also varying them!

UNIVERSITÄT LIECHTENSTEIN

(Some of our guidelines *do* overlap with those of Arp et al. [8])

But some don't!

[8] D. Arp et al. "Dos and don'ts of machine learning in computer security." In *USENIX Security Symp.*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(2) Research Guidelines**

How can researchers meet the needs of practitioners?

1. **System Infrastructure**

   → *Provide a schematic!*

   Readers will like it!

2. **Preprocessing**

   → *Report which tools*

   Try also varying them!

3. **Data Availability**

   → *Consider different amounts of training data*

   You do not "always" need to outperform SotA

UNIVERSITÄT LIECHTENSTEIN

(Some of our guidelines *do* overlap with those of Arp et al. [8]) But some don't!

[8] D. Arp et al. "Dos and don'ts of machine learning in computer security." In *USENIX Security Symp.*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(2) Research Guidelines**

How can researchers meet the needs of practitioners?

1. **System Infrastructure**

   → *Provide a schematic!*

   Readers will like it!

2. **Preprocessing**

   → *Report which tools*

   Try also varying them!

3. **Data Availability**

   → *Consider different amounts of training data*

   You do not "always" need to outperform SotA

4. **Hardware**

   → *Report the specifications of the evaluation platform*

   Measure the runtime!

UNIVERSITÄT LIECHTENSTEIN

(Some of our guidelines *do* overlap with those of Arp et al. [8]) But some don't!

[8] D. Arp et al. "Dos and don'ts of machine learning in computer security." In *USENIX Security Symp.*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(2) Research Guidelines**

How can researchers meet the needs of practitioners?

1. **System Infrastructure**
   → *Provide a schematic!*
   Readers will like it!
2. **Preprocessing**
   → *Report which tools*
   Try also varying them!
3. **Data Availability**
   → *Consider different amounts of training data*
   You do not "always" need to outperform SotA
4. **Hardware**
   → *Report the specifications of the evaluation platform*
   Measure the runtime!
5. **Unpredictability**
   → *Assess as many "likely" operational scenarios as possible*
   Also report the statistical significance

UNIVERSITÄT LIECHTENSTEIN

(Some of our guidelines *do* overlap with those of Arp et al. [8]) But some don't!

[8] D. Arp et al. "Dos and don'ts of machine learning in computer security." In *USENIX Security Symp.*, 2022.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(3) State-of-the-Art?**

How does the SotA “comply” with our recommendations?

Venues: S&P, EuroS&P, SEC, NDSS, CCS, AsiaCCS, RAID, DIMVA, ACSAC.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(3) State-of-the-Art?**

How does the SotA "comply" with our recommendations?

Venues: S&P, EuroS&P, SEC, NDSS, CCS, AsiaCCS, RAID, DIMVA, ACSAC.

TABLE 2: State-of-the-Art: papers published since 2017 in top cybersecurity conferences that consider applications of ML linked with NID.

Paper	Year	Hardware	Runtime	Adaptive	Stat. Sign.	Avail.	Pub. Data
Bortolamelotti [113]	2017	✗	✗	✓	✗	✗	✗ (1)
Ho [120]	2017	✗	✗	◐	✗	✗	✗ (1)
Cho [121]	2017	✗	✗	✓	✗	✗	✗ (1)
Siadati [122]	2017	✗	✗	◐	✗	✗	✗ (1)
Oprea [46]	2018	✗	$\mathbb{T}$	◐	✗	✗	✗ (1)
Pereira [95]	2018	◐	$\mathbb{T}$	◐	✗	✓	◐ (1)
Kheib [123]	2018	✗	✗	◐	✗	✗	✗ (1)
Araujo [124]	2019	✗	$\mathbb{E}$	✗	✗	✓	✗ (1)
Mudgerikar [112]	2019	✗	✓	✗	✗	✗	✗ (1)
Mirsky [60]	2019	◐	✓	◐	✗	✗	✓ (1)
Feng [125]	2019	✗	✗	◐	✗	✗	✓ (2)
Milajerdi [114]	2019	◐	✓	◐	✗	✗	✓ (1)
Liu [126]	2019	◐	✗	◐	✗	✗	✓ (2)
Du [127]	2019	✗	$\mathbb{T}$	◐	✗	✗	✓ (3)
Erba [116]	2020	◐	$\mathbb{E}$	✓	✗	✓	✓ (2)
Bowman [98]	2020	◐	$\mathbb{E}$	✗	✗	✗	✓ (2)
Leichtnam [128]	2020	◐	✗	✗	✗	✗	✓ (1)
Singla [129]	2020	✗	✗	✗	✗	✓	✓ (2)
Han [130]	2020	✓	✓	◐	✗	✗	✓ (2)
Jan [131]	2020	✗	✗	✓	✓	✓	✗ (1)
Ghorbani [132]	2021	✓	$\mathbb{E}$	◐	✗	✗	✗ (1)
Nabeel [133]	2021	✗	✗	◐	✗	✗	✗ (1)
Wang [115]	2021	✗	$\mathbb{E}$	✓	✗	✗	✓ (2)
Piszkozub [134]	2021	✗	✗	◐	✗	✗	◐ (2)
Yuan [135]	2021	✗	✗	◐	✗	✓	✓ (1)
Yang [136]	2021	✗	✗	◐	✓	✗	✓ (1)
Barradas [137]	2021	◐	✓	◐	✗	✗	✓ (1)
Han [138]	2021	✓	✓	✓	✗	✓	✓ (2)
Liang [139]	2021	✗	$\mathbb{T}$	◐	✓	✓	✓ (1)
Fu [140]	2021	◐	✓	✓	✗	✗	✓ (3)

TABLE 5: State-of-the-Art (2022): papers published in top cybersecurity conferences that consider applications of ML linked with NID.

Paper	Year	Hardware	Runtime	Adaptive	Stat. Sign.	Avail.	Pub. Data
Apruzzese [79]	2022	✓	$\mathbb{T}$	✗	✓	✓	✓ (3)
Arp [8]	2022	✗	✗	◐	✗	✗	✓ (1)
D'hooge [179]	2022	✗	✗	✗	✗	✓	✓ (8)
Dodia [170]	2022	✗	✗	✗	✓	✗	✓ (1)
Erba [177]	2022	✗	✗	✓	✗	✗	✓ (1)
Feng [180]	2022	✓	✓	◐	✗	✓	✓ (1)
Fu [181]	2022	✓	$\mathbb{E}$	◐	✗	✗	✓ (2)
Jacobs [178]	2022	✗	✗	✗	✗	✗	✓ (6)
King [182]	2022	✓	✓	✗	✗	✓	✓ (3)
Landen [183]	2022	✗	$\mathbb{T}$	✓	✗	✓	✗ (1)
Sharma [184]	2022	◐	✗	◐	✗	✗	✗ (1)
Tekiner [185]	2022	✓	$\mathbb{E}$	✓	✓	✓	✓ (3)
Van Ede [61]	2022	✓	✓	✓	✗	✓	✓ (1)
Wang [186]	2022	✓	✓	✓	✗	✓	✓ (1)
Wang [187]	2022	✗	✗	✗	✗	✗	✓ (3)
Wolsing [169]	2022	✗	✗	✗	✗	✗	✓ (3)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(3) State-of-the-Art?**

How does the SotA "comply" with our recommendations?

TABLE 2: State-of-the-Art: papers published since 2017 in top cybersecurity conferences that consider applications of ML linked with NID.

Paper	Year	Hardware	Runtime	Adaptive	Stat. Sign.	Avail.	Pub. Data
Bortolamelotti [113]	2017	✗	✗	✓	✗	✗	✗ (1)
Ho [120]	2017	✗	✗	◐	✗	✗	✗ (1)
Cho [121]	2017	✗	✗	✓	✗	✗	✗ (1)
Siadati [122]	2017	✗	✗	◐	✗	✗	✗ (1)
Oprea [46]	2018	✗	T	◐	✗	✗	✗ (1)
Pereira [95]	2018	◐	T	◐	✗	✓	◐ (1)
Kheib [123]	2018	✗	✗	◐	✗	✗	✗ (1)
Araujo [124]	2019	✗	E	✗	✗	✓	✗ (1)
Mudgerikar [112]	2019	✗	✓	✗	✗	✗	✗ (1)
Mirsky [60]	2019	◐	✓	◐	✗	✗	✓ (1)
Feng [125]	2019	✗	✗	◐	✗	✗	✓ (2)
Milajerdi [114]	2019	◐	✓	◐	✗	✗	✓ (1)
Liu [126]	2019	◐	✗	◐	✗	✗	✓ (2)
Du [127]	2019	✗	T	◐	✗	✗	✓ (3)
Erba [116]	2020	◐	E	✓	✗	✓	✓ (2)
Bowman [98]	2020	◐	E	✗	✗	✗	✓ (2)
Leichtnam [128]	2020	◐	✗	✗	✗	✗	✓ (1)
Singla [129]	2020	✗	✗	✗	✗	✓	✓ (2)
Han [130]	2020	✓	✓	◐	✗	✗	✓ (2)
Jan [131]	2020	✗	✗	✓	✓	✓	✗ (1)
Ghorbani [132]	2021	✓	E	◐	✗	✗	✗ (1)
Nabeel [133]	2021	✗	✗	◐	✗	✗	✗ (1)
Wang [115]	2021	✗	E	✓	✗	✗	✓ (2)
Piszkozub [134]	2021	✗	✗	◐	✗	✗	◐ (2)
Yuan [135]	2021	✗	✗	◐	✗	✓	✓ (1)
Yang [136]	2021	✗	✗	◐	✓	✗	✓ (1)
Barradas [137]	2021	◐	✓	◐	✗	✗	✓ (1)
Han [138]	2021	✓	✓	✓	✗	✓	✓ (2)
Liang [139]	2021	✗	T	◐	✓	✓	✓ (1)
Fu [140]	2021	◐	✓	✓	✗	✗	✓ (3)

Venues: S&P, EuroS&P, SEC, NDSS, CCS, AsiaCCS, RAID, DIMVA, ACSAC.

TABLE 5: State-of-the-Art (2022): papers published in top cybersecurity conferences that consider applications of ML linked with NID.

Paper	Year	Hardware	Runtime	Adaptive	Stat. Sign.	Avail.	Pub. Data
Apruzzese [79]	2022	✓	T	✗	✓	✓	✓ (3)
Arp [8]	2022	✗	✗	◐	✗	✗	✓ (1)
D'hooge [179]	2022	✗	✗	✗	✗	✓	✓ (8)
Dodia [170]	2022	✗	✗	✗	✓	✗	✓ (1)
Erba [177]	2022	✗	✗	✓	✗	✗	✓ (1)
Feng [180]	2022	✓	✓	◐	✗	✓	✓ (1)
Fu [181]	2022	✓	E	◐	✗	✗	✓ (2)
Jacobs [178]	2022	✗	✗	✗	✗	✗	✓ (6)
King [182]	2022	✓	✓	✗	✗	✓	✓ (3)
Landen [183]	2022	✗	T	✓	✗	✓	✗ (1)
Sharma [184]	2022	◐	✗	◐	✗	✗	✗ (1)
Tekiner [185]	2022	✓	E	✓	✓	✓	✓ (3)
Van Ede [61]	2022	✓	✓	✓	✗	✓	✓ (1)
Wang [186]	2022	✓	✓	✓	✗	✓	✓ (1)
Wang [187]	2022	✗	✗	✗	✗	✗	✓ (3)
Wolsing [169]	2022	✗	✗	✗	✗	✗	✓ (3)

We added this during the peer-review!
(There is an improvement over the previous 5 years)

UNIVERSITÄT LIECHTENSTEIN

All papers consider "open-world" scenarios

No paper changes the preprocessing tool

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(4) Practitioners' opinion – A**

User study with 12 practitioners with hands-on experience on ML and NID, who are acquainted with research and work in renown security companies.

"How important is this factor?"

Factor	Not important	Important	Crucial
System Infrastructure			
Preprocessing			
Data Availability			
Hardware			
Unpredictability			

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(4) Practitioners’ opinion – A**

User study with 12 practitioners with hands-on experience on ML and NID, who are acquainted with research and work in renown security companies.

“How important is this factor?”

Factor	Not important	Important	Crucial
System Infrastructure	9%	27%	64%
Preprocessing	0%	9%	91%
Data Availability	9%	18%	73%
Hardware	9%	64%	27%
Unpredictability	9%	18%	73%

- Preprocessing is the most relevant
- Hardware is the least relevant

We made them change their mind!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(4) Practitioners' opinion – B**

User study with 12 practitioners with hands-on experience on ML and NID, who are acquainted with research and work in renown security companies.

TABLE 2: State-of-the-Art: papers published since 2017 in top cybersecurity conferences that consider applications of ML linked with NID.

Paper	Year	Hardware	Runtime	Adaptive	Stat. Sign.	Avail.	Pub. Data
	2017	✗	✗	✓	✗	✗	✗ (1)
	2017	✗	✗	◐	✗	✗	✗ (1)
	2017	✗	✗	✓	✗	✗	✗ (1)
	2017	✗	✗	◐	✗	✗	✗ (1)
	2018	✗	T	◐	✗	✗	✗ (1)
	2018	◐	T	◐	✗	✓	◐ (1)
	2018	✗	✗	◐	✗	✗	✗ (1)
	2019	✗	E	✗	✗	✓	✗ (1)
	2019	✗	✓	✗	✗	✗	✗ (1)
	2019	◐	✓	◐	✗	✗	✓ (1)
	2019	✗	✗	◐	✗	✗	✓ (2)
	2019	◐	✓	◐	✗	✗	✓ (1)
	2019	◐	✗	◐	✗	✗	✓ (2)
	2019	✗	T	◐	✗	✗	✓ (3)
	2020	◐	E	✓	✗	✓	✓ (2)
	2020	◐	E	✗	✗	✗	✓ (2)
	2020	◐	✗	✗	✗	✗	✓ (1)
	2020	✗	✗	✗	✗	✓	✓ (2)
	2020	✓	✓	◐	✗	✗	✓ (2)
	2020	✗	✗	✓	✓	✓	✗ (1)
	2021	✓	E	◐	✗	✗	✗ (1)
	2021	✗	✗	◐	✗	✗	✗ (1)
	2021	✗	E	✓	✗	✗	✓ (2)
	2021	✗	✗	◐	✗	✗	◐ (2)
	2021	✗	✗	◐	✗	✓	✓ (1)
	2021	✗	✗	◐	✓	✗	✓ (1)
	2021	◐	✓	◐	✗	✗	✓ (1)
	2021	✓	✓	✓	✗	✓	✓ (2)
	2021	✗	T	◐	✓	✓	✓ (1)
	2021	◐	✓	✓	✗	✗	✓ (3)

We did this in 2022

"How problematic is it that…"

Column (Issue)	Not very Problematic	Problematic (but OK)	Very problematic
Poor Hardware			
Poor Runtime			
Poor Adaptive atk.			
Poor Stat. Sign.			
Poor Data Availab.			
Poor Pub. Data			

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(4) Practitioners' opinion – B**

User study with 12 practitioners with hands-on experience on ML and NID, who are acquainted with research and work in renown security companies.

TABLE 2: State-of-the-Art: papers published since 2017 in top cybersecurity conferences that consider applications of ML linked with NID.

Paper	Year	Hardware	Runtime	Adaptive	Stat. Sign.	Avail.	Pub. Data
	2017	✗	✗	✓	✗	✗	✗ (1)
	2017	✗	✗	◐	✗	✗	✗ (1)
	2017	✗	✗	✓	✗	✗	✗ (1)
	2017	✗	✗	◐	✗	✗	✗ (1)
	2018	✗	T	◐	✗	✗	✗ (1)
	2018	◐	T	◐	✗	✓	◐ (1)
	2018	✗	✗	◐	✗	✗	✗ (1)
	2019	✗	E	✗	✗	✓	✗ (1)
	2019	✗	✓	✗	✗	✗	✗ (1)
	2019	◐	✓	◐	✗	✗	✓ (1)
	2019	✗	✗	◐	✗	✗	✓ (2)
	2019	◐	✓	◐	✗	✗	✓ (1)
	2019	◐	✗	◐	✗	✗	✓ (2)
	2019	✗	T	◐	✗	✗	✓ (3)
	2020	◐	E	✓	✗	✓	✓ (2)
	2020	◐	E	✗	✗	✗	✓ (2)
	2020	◐	✗	✗	✗	✗	✓ (1)
	2020	✗	✗	✗	✗	✓	✓ (2)
	2020	✓	✓	◐	✗	✗	✓ (2)
	2020	✗	✗	✓	✓	✓	✗ (1)
	2021	✓	E	◐	✗	✗	✗ (1)
	2021	✗	✗	◐	✗	✗	✗ (1)
	2021	✗	E	✓	✗	✗	✓ (2)
	2021	✗	✗	◐	✗	✗	◐ (2)
	2021	✗	✗	◐	✗	✓	✓ (1)
	2021	✗	✗	◐	✓	✗	✓ (1)
	2021	◐	✓	◐	✗	✗	✓ (1)
	2021	✓	✓	✓	✗	✓	✓ (2)
	2021	✗	T	◐	✓	✓	✓ (1)
	2021	◐	✓	✓	✗	✗	✓ (3)

We did this in 2022

"How problematic is it that…"

Column (Issue)	Not very Problematic	Problematic (but OK)	Very problematic
Poor Hardware	25%	75%	0%
Poor Runtime	0%	75%	25%
Poor Adaptive atk.	8%	67%	25%
Poor Stat. Sign.	0%	10%	90%
Poor Data Availab.	16%	42%	42%
Poor Pub. Data	0%	41%	59%

Note2: we made them change their mind on hardware and runtime!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets ("large" and "small")

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets (“large” and “small”)
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets ("large" and "small")
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)
- 4 ML algorithms (no deep learning!) DL is impractical!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets ("large" and "small")
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)
- 4 ML algorithms (no deep learning!) DL is impractical!
- 6 Hardware platforms (from a Raspberry Pi4B to an HPC)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets (“large” and “small”)
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)
- 4 ML algorithms (no deep learning!) DL is impractical!
- 6 Hardware platforms (from a Raspberry Pi4B to an HPC)

We evaluate all of the above in: (i) open-world, (ii) closed-world, and (iii) adversarial settings.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets ("large" and "small")
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)
- 4 ML algorithms (no deep learning!) DL is impractical!
- 6 Hardware platforms (from a Raspberry Pi4B to an HPC)

We evaluate all of the above in: (i) open-world, (ii) closed-world, and (iii) adversarial settings.

We consider *random* split and *temporal* splits. There is no difference!

- We repeat all the random splits 100 times (to compute statistically significant results)

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
  - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets (“large” and “small”)
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)
- 4 ML algorithms (no deep learning!) *DL is impractical!*
- 6 Hardware platforms (from a Raspberry Pi4B to an HPC)

We evaluate all of the above in: (i) open-world, (ii) closed-world, and (iii) adversarial settings.

We consider *random* split and *temporal* splits. *There is no difference!*

- We repeat all the random splits 100 times (to compute statistically significant results)

We measure the *true positive* rate, *false positive* rate, *inference* time, *training* time.

Our paper is 36 pages long!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

What we do: **(5) Pragmatic Assessment**

We showcase how to apply *all* our guidelines in research.

We do so by re-assessing existing methods for *Netflow classification*.

We consider:

- 5 well-known public datasets (from diverse network enviroments)
    - Each generated with a different NetFlow tool
- 4 amounts of data availability (from 100s to 80% of total dataset)
- 2 feature sets (“large” and “small”)
- 6 ML pipelines (single classifiers, ensembles, and even a cascade)
- 4 ML algorithms (no deep learning!) DL is impractical!
- 6 Hardware platforms (from a Raspberry Pi4B to an HPC)

We evaluate all of the above in: (i) open-world, (ii) closed-world, and (iii) adversarial settings.

We consider *random* split and *temporal* splits. There is no difference!

- We repeat all the random splits 100 times (to compute statistically significant results)

We measure the *true positive* rate, *false positive* rate, *inference* time, *training* time.

(Source code available at https://github.com/hihey54/pragmaticAssessment)

Our paper is 36 pages long!

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Let’s talk about Hardware…

Hardware is largely neglected in past research.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Let’s talk about Hardware…

Hardware is largely neglected in past research.

- Some <u>do not provide any</u> hardware specs
  - Interestingly, some report the runtime without specifying the hardware…

There is a huge difference between a Raspberry Pi4B and an HPC…

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

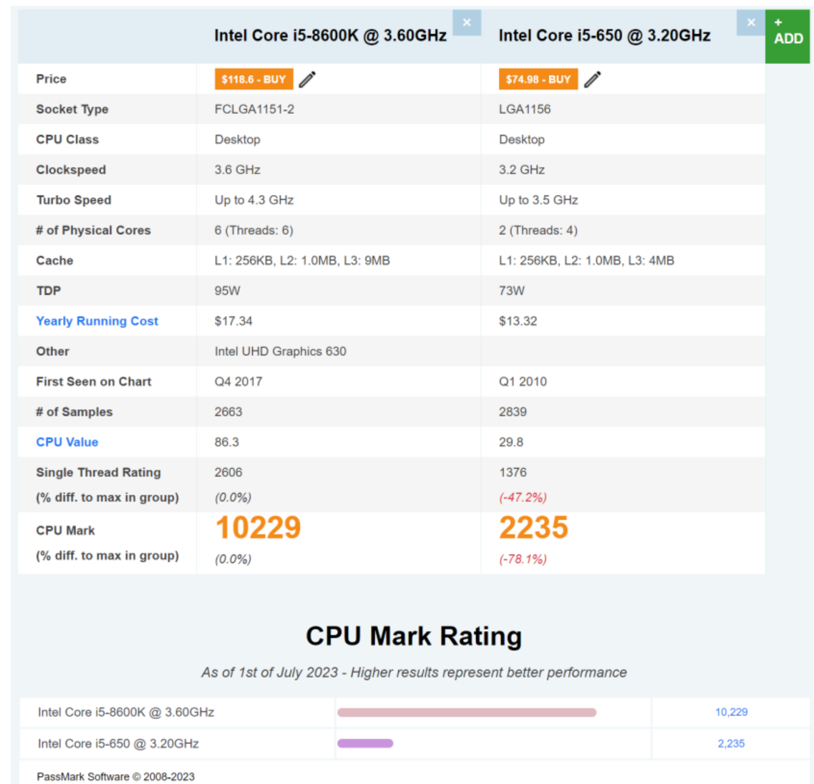
Let’s talk about Hardware…

Hardware is largely neglected in past research.

- Some do not provide any hardware specs
  - Interestingly, some report the runtime without specifying the hardware…
- Most papers report incomplete hardware specs
  - Some stated (in 2018) “the CPU is an Intel Core i5”

There is a huge difference between a Raspberry Pi4B and an HPC…

In 2018, there were 80 different “Intel Core i5” available on the market [A]



	Intel Core i5-8600K @ 3.60GHz	Intel Core i5-650 @ 3.20GHz
Price	$118.6 - BUY	$74.98 - BUY
Socket Type	FCLGA1151-2	LGA1156
CPU Class	Desktop	Desktop
Clockspeed	3.6 GHz	3.2 GHz
Turbo Speed	Up to 4.3 GHz	Up to 3.5 GHz
# of Physical Cores	6 (Threads: 6)	2 (Threads: 4)
Cache	L1: 256KB, L2: 1.0MB, L3: 9MB	L1: 256KB, L2: 1.0MB, L3: 4MB
TDP	95W	73W
Yearly Running Cost	$17.34	$13.32
Other	Intel UHD Graphics 630	
First Seen on Chart	Q4 2017	Q1 2010
# of Samples	2663	2839
CPU Value	86.3	29.8
Single Thread Rating (% diff. to max in group)	2606 (0.0%)	1376 (-47.2%)
CPU Mark (% diff. to max in group)	10229 (0.0%)	2235 (-78.1%)

UNIVERSITÄT LIECHTENSTEIN

[A] https://en.wikipedia.org/wiki/List_of_Intel_Core_i5_processors

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Let’s talk about Hardware…

Hardware is largely neglected in past research.

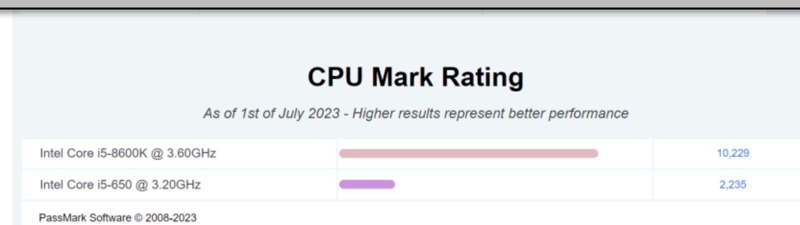
- Some do not provide any hardware specs
  - Interestingly, some report the runtime without specifying the hardware…
- Most papers report incomplete hardware specs
  - Some stated (in 2018) “the CPU is an Intel Core i5”

There is a huge difference between a Raspberry Pi4B and an HPC…

In 2018, there were 80 different “Intel Core i5” available on the market [A]

	Intel Core i5-8600K @ 3.60GHz	Intel Core i5-650 @ 3.20GHz
Price	$118.6 - BUY	$74.98 - BUY
Socket Type	FCLGA1151-2	LGA1156
CPU Class	Desktop	Desktop
Clockspeed	3.6 GHz	3.2 GHz
Turbo Speed	Up to 4.3 GHz	Up to 3.5 GHz
# of Physical Cores	6 (Threads: 6)	2 (Threads: 4)
Cache	L1: 256KB, L2: 1.0MB, L3: 9MB	L1: 256KB, L2: 1.0MB, L3: 4MB
TDP	95W	73W
Yearly Running Cost	$17.34	$13.32

Reporting the complete specifications can determine the “winner” among 2+ ML methods



UNIVERSITÄT LIECHTENSTEIN

[A] https://en.wikipedia.org/wiki/List_of_Intel_Core_i5_processors

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

REMARK

We do a massive re-assessment, but not all research <u>must</u> do <u>all of</u> what we suggest

There is value even in "small" evaluations, if appropriate to test a given hypothesis!

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

TAKEAWAY

We want to see our research have a better impact to the (practical) real world.

In our user-study with practitioners, we asked a final question:

"In general, do you think that research papers facilitate the practitioners' job in determining the real value of the proposed ML methods?"

- 92% are "uncertain"
- 8% are "left with more questions than answers after reading a research paper"

Our paper can hopefully inspire the change we want to see.

UNIVERSITÄT LIECHTENSTEIN

TU Delft – July 5th, 2023

IEEE European Symposium on Security and Privacy

SoK: Pragmatic Assessment of Machine Learning for Network Intrusion Detection

Giovanni Apruzzese, Pavel Laskov, Johannes Schneider

UNIVERSITÄT LIECHTENSTEIN