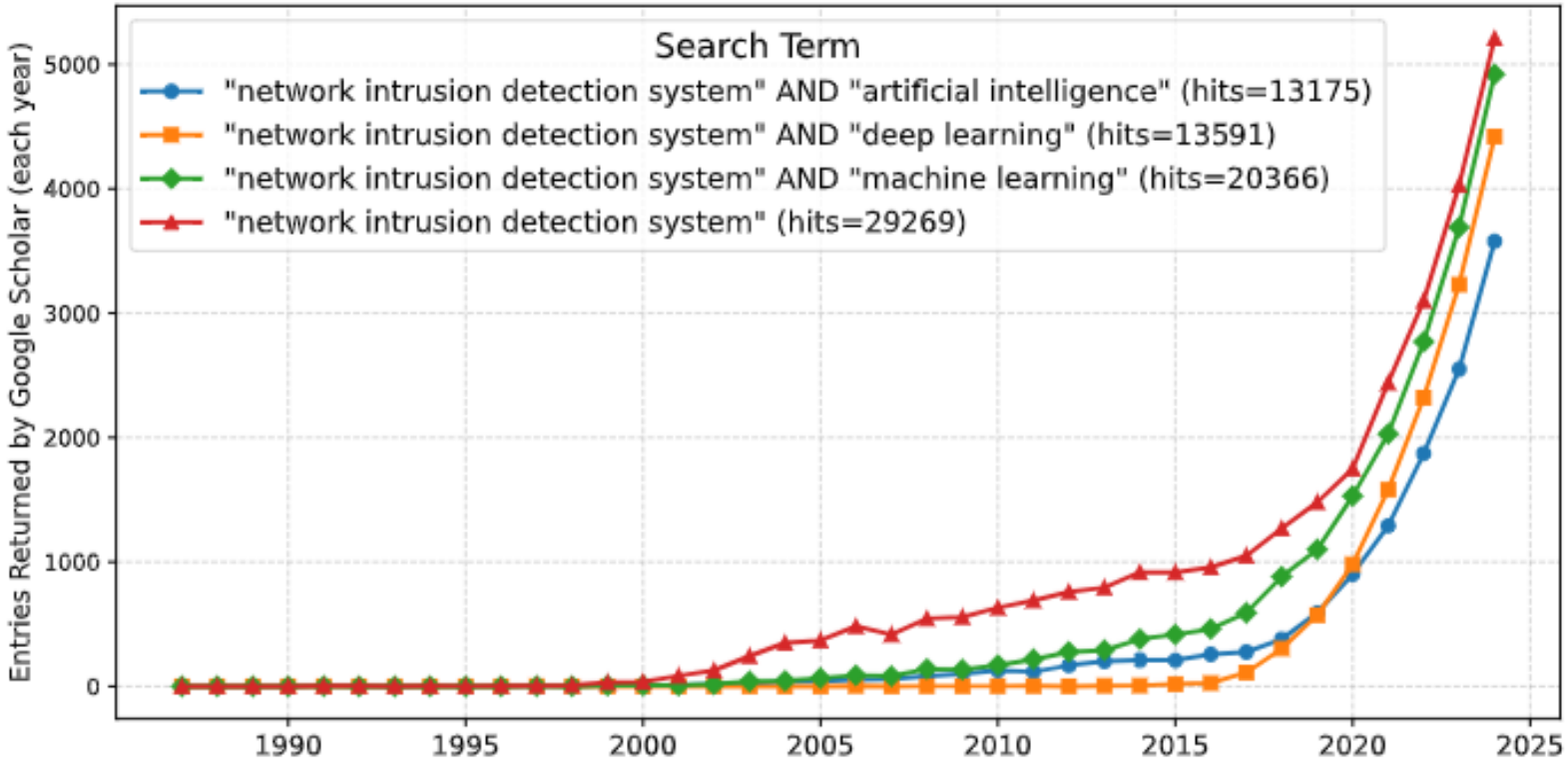
Bangalore, June 4th 2026

21st ACM ASIA Conference on Computer and Communications Security

SoK: Reshaping Research on Network Intrusion Detection Systems

Giovanni Apruzzese

HÁSKÓLINN Í REYKJAVÍK | REYKJAVIK UNIVERSITY



Research interest in NIDS. Queries issued on September 2025.

Giovanni Apruzzese
giovannia@ru.is

Giovanni Apruzzese
giovannia@ru.is

Assumption

Research in NIDS is relatively stagnant, and is hindered by, among others, a superficial treatment of the NIDS domain

This paper wants to change this

Giovanni Apruzzese
giovannia@ru.is

How?

By stating three ASSERTIONS
and devising a practical *vademecum*.

(and showing a simple, novel experiment)

Giovanni Apruzzese
giovannia@ru.is

Rules

- We will use prior work to provide a basis for the arguments and ASSERTIONS stated in the paper.

Giovanni Apruzzese
giovannia@ru.is

Rules

- We will use prior work to provide a basis for the arguments and ASSERTIONS stated in the paper.
- We will use prior work to suggest avenues for future work, or highlight exemplary good practices.

Giovanni Apruzzese
giovannia@ru.is

Rules

- We will use prior work to provide a basis for the arguments and ASSERTIONS stated in the paper.
- We will use prior work to suggest avenues for future work, or highlight exemplary good practices.
- We will never explicitly criticize any prior work—barred those co-authored by the author of this paper.

Giovanni Apruzzese
giovannia@ru.is

Intended Audience

- People who want to carry out research in NIDS

Giovanni Apruzzese
giovannia@ru.is

Intended Audience

- People who want to carry out research in NIDS
- Reviewers of NIDS-related documents.

Giovanni Apruzzese
giovannia@ru.is

Terminology

Giovanni Apruzzese
giovannia@ru.is

Terminology

- What is an *anomaly*?

Giovanni Apruzzese
giovannia@ru.is

Terminology

- What is an *anomaly*?

- What is the *output* of an NIDS?

Giovanni Apruzzese
giovannia@ru.is

Terminology

- What is an *anomaly*?

- What is the *output* of an NIDS?

"Unrealistic"

Giovanni Apruzzese
giovannia@ru.is

Terminology

- What is an *anomaly*?

- What is the *output* of an NIDS?

"Unrealistic"
is a terrible term...

Giovanni Apruzzese
giovannia@ru.is

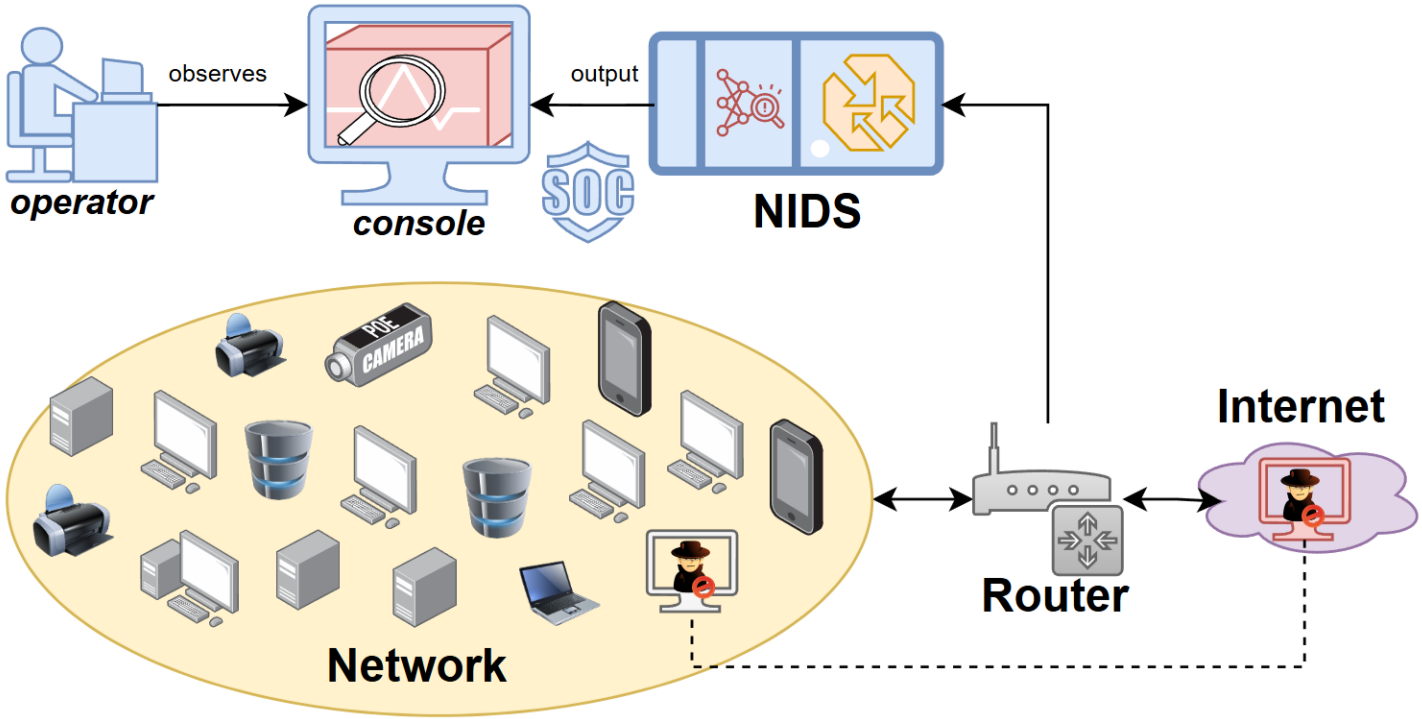


Fig. 2: Typical NIDS scenario. The NIDS receives data from the *router* as input, and shows its output in a *console*. The NIDS expects intrusions to occur in the network (or from the internet).

Giovanni Apruzzese
giovannia@ru.is

Assertion 1. It is non-sensical to assume a scenario in which a network intrusion detection system (NIDS) is expected to work against an attacker that has compromised such a system.

Giovanni Apruzzese
giovannia@ru.is

Assertion 1. It is non-sensical to assume a scenario in which a network intrusion detection system (NIDS) is expected to work against an attacker that has compromised such a system.

Recommendation 1. The threat model must assume attackers who can only control the hosts from the Internet, or the hosts within the network that they have compromised—which should not include the router that feeds data to the NIDS, the NIDS itself, or the NIDS output console.

Giovanni Apruzzese
giovannia@ru.is

Assertion 1. It is non-sensical to assume a scenario in which a network intrusion detection system (NIDS) is expected to work against an attacker that has compromised such a system.

Recommendati assume attackers who ca the hosts withi should not inclu NIDS itself, or the NIDS output

Reflective exercise:
"If the attack's methodology *does* implicitly assume that the NIDS has been compromised, then *how can this be done?*"

Giovanni Apruzzese
giovannia@ru.is

Assertion 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Recommendation 2: Ensure that the deployment scenario described in the paper aligns with the network environment captured by the evaluation dataset.

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Recon nario de-
scribe ent cap-
tured by t

Corollary (note to reviewers):
Criticizing a paper because its evaluation cannot prove that a given result holds in general is not a constructive, and is almost an unfair, remark

Giovanni Apruzzese
giovannia@ru.is

Assertion 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Recommendation 3: the performance of a NIDS-related module should be tied to its practical utility—which requires taking into account also post-processing techniques applied to the output of such a module, and the overarching operational context.

Giovanni Apruzzese
giovannia@ru.is

Assertion 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Recommend ... ted module should ... g into account ... utput of such a module, ... context.

Note:
These additional investigations can reveal some additional properties which may overturn some empirical results

Giovanni Apruzzese
giovannia@ru.is

Experiment – Threat Model

- Small scale network of a dozen hosts.
- The network traffic received by the router is transformed into NetFlows which are forwarded to an ML-based classifier integrated in an NIDS

Giovanni Apruzzese
giovannia@ru.is

Experiment – Threat Model

- Small scale network of a dozen hosts.
- The network traffic received by the router is transformed into NetFlows which are forwarded to an ML-based classifier integrated in an NIDS
- The classifier was trained on various network attacks (e.g., SSH-bruteforcing, or DoS attempts).
- The NIDS receives the NetFlows from the router, and shows the output of the classifier to a network operator, who must decide on a daily basis whether some hosts require some manual triaging.

Giovanni Apruzzese
giovannia@ru.is

Experiment – Threat Model

- Small scale network of a dozen hosts.
- The network traffic received by the router is transformed into NetFlows which are forwarded to an ML-based classifier integrated in an NIDS
- The classifier was trained on various network attacks (e.g., SSH-bruteforcing, or DoS attempts).
- The NIDS receives the NetFlows from the router, and shows the output of the classifier to a network operator, who must decide on a daily basis whether some hosts require some manual triaging.
- The attacker has gained access to one host, and wants to launch attacks not included in the training set of the classifier (e.g., setting up a botnet)
- The attacker has no access whatsoever to the NIDS infrastructure

Giovanni Apruzzese
giovannia@ru.is

Assertion 1. It is non-sensical to assume a scenario in which a network intrusion detection system (NIDS) is expected to work against an attacker that has compromised such a system.

Experiment – Threat Model

- Small scale network of a dozen hosts.
- The network traffic received by the router is transformed into NetFlows which are forwarded to an ML-based classifier integrated in an NIDS
- The classifier was trained on various network attacks (e.g., SSH-bruteforcing, or DoS attempts).
- The NIDS receives the NetFlows from the router, and shows the output of the classifier to a network operator, who must decide on a daily basis whether some hosts require some manual triaging.
- The attacker has gained access to one host, and wants to launch attacks not included in the training set of the classifier (e.g., setting up a botnet)
- The attacker has no access whatsoever to the NIDS infrastructure

Compliance with ASSERTION 1

...anni Apruzzese
giovannia@ru.is

Experiment – Setup

- Dataset: CICIDS17 [116] (but we use the fixed variant [82])

[116] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[82] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021.

Giovanni Apruzzese
giovannia@ru.is

Experiment – Setup

- Dataset: CICIDS17 [116] (but we use the fixed variant [82])
  - It has NetFlow data captured in a network of ~20 hosts over 5 days
  - On the fifth day, attacks of the ‘ARES’ botnet were carried out
  - → We use the first four days to train the ML model, and test on the fifth

[116] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[82] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021.

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Experiment – Setup

- Dataset: CICIDS17 [116] (but we use the fixed variant [82])
  - It has NetFlow data captured in a network of ~20 hosts over 5 days
  - On the fifth day, attacks of the 'ARES' botnet were carried out
  - → We use the first four days to train the ML model, and test on the fifth

Compliance with ASSERTION 2

[116] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[82] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021.

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Experiment – Setup

- Dataset: CICIDS17 [116] (but we use the fixed variant [82])
  - It has NetFlow data captured in a network of ~20 hosts over 5 days
  - On the fifth day, attacks of the 'ARES' botnet were carried out
  - → We use the first four days to train the ML model, and test on the fifth
- We extract and label the NetFlows from CICIDS17
- We train four ML classifiers (DT, RF, HFB, RF) by applying an 80:20 split

Compliance with ASSERTION 2

[116] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[82] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021.

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Experiment – Setup

- Dataset: CICIDS17 [116] (but we use the fixed variant [82])
  - It has NetFlow data captured in a network of ~20 hosts over 5 days
  - On the fifth day, attacks of the 'ARES' botnet were carried out
  - → We use the first four days to train the ML model, and test on the fifth
- We extract and label the NetFlows from CICIDS17
- We train four ML classifiers (DT, RF, HFB, RF) by applying an 80:20 split

Classifier	Validation set (NetFlows) TPR	FPR	#Errors
DT	0.9996	0.0003	45
RF	0.9996	<0.0001	16
HGB	0.9998	<0.0001	9
LR	0.9373	0.0711	8,068

Compliance with ASSERTION 2

[116] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[82] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021.

anni Apruzzese
giovannia@ru.is

ASSERTION 2. The evaluation testbed must account for the real-world characteristics and implications (for both "benign" and "malicious" hosts) of the NIDS' deployment scenario *envisioned in the research paper.*

Experiment – Setup

- Dataset: CICIDS17 [116] (but we use the fixed variant [82])
  - It has NetFlow data captured in a network of ~20 hosts over 5 days
  - On the fifth day, attacks of the 'ARES' botnet were carried out
  - → We use the first four days to train the ML model, and test on the fifth
- We extract and label the NetFlows from CICIDS17
- We train four ML classifiers (DT, RF, HFB, RF) by applying an 80:20 split

Classifier	Validation set (NetFlows) TPR	FPR	#Errors
DT	0.9996	0.0003	45
RF	0.9996	<0.0001	16
HGB	0.9998	<0.0001	9
LR	0.9373	0.0711	8,068

Compliance with ASSERTION 2

[116] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[82] Engelen, Gints, Vera Rimmer, and Wouter Joosen. "Troubleshooting an intrusion detection dataset: the CICIDS2017 case study." 2021 IEEE Security and Privacy Workshops (SPW). IEEE, 2021.

...anni Apruzzese
giovannia@ru.is

Experiment – Results

- TPR and FPR on the test set

Classifier	Validation set (NetFlows) TPR	FPR	#Errors	Test set (NetFlows) TPR	FPR	#Errors
DT	0.9996	0.0003	45	0.0636	0.0002	238,807
RF	0.9996	<0.0001	16	0.3706	0.0000	160,515
HGB	0.9998	<0.0001	9	0.3729	<0.0001	159,913
LR	0.9373	0.0711	8,068	0.3734	0.0564	164,722

Giovanni Apruzzese
giovannia@ru.is

Experiment – Results

- TPR and FPR on the test set

Classifier	Validation set (NetFlows) TPR	FPR	#Errors	Test set (NetFlows) TPR	FPR	#Errors
DT	0.9996	0.0003	45	0.0636	0.0002	238,807
RF	0.9996	<0.0001	16	0.3706	0.0000	160,515
HGB	0.9998	<0.0001	9	0.3729	<0.0001	159,913
LR	0.9373	0.0711	8,068	0.3734	0.0564	164,722

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Experiment – Results

- TPR and FPR on the test set
- …but what if we simply look at the hosts that triggered at least one alarm?

Compliance with ASSERTION 3

Classifier	Validation set (NetFlows) TPR	FPR	#Errors	Test set (NetFlows) TPR	FPR	#Errors
DT	0.9996	0.0003	45	0.0636	0.0002	238,807
RF	0.9996	<0.0001	16	0.3706	0.0000	160,515
HGB	0.9998	<0.0001	9	0.3729	<0.0001	159,913
LR	0.9373	0.0711	8,068	0.3734	0.0564	164,722

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Experiment – Results

- TPR and FPR on the test set
- …but what if we simply look at the hosts that triggered at least one alarm?

Compliance with ASSERTION 3

Classifier	Validation set (NetFlows) TPR	FPR	#Errors	Test set (NetFlows) TPR	FPR	#Errors	Test set (Hosts) TPR	FPR	#Errors
DT	0.9996	0.0003	45	0.0636	0.0002	238,807	1.000	0.250	2
RF	0.9996	<0.0001	16	0.3706	0.0000	160,515	0.125	0	7
HGB	0.9998	<0.0001	9	0.3729	<0.0001	159,913	0.125	0.125	8
LR	0.9373	0.0711	8,068	0.3734	0.0564	164,722	1.000	1.000	8

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Experiment – Results

- TPR and FPR on the test set
- …but what if we simply look at the hosts that triggered at least one alarm?
  - The DT flagged 10 hosts, of which 8 are involved in malicious NetFlows
  - The RF flagged only one 'malicious' host
  - The HGB flagged two hosts – one malicious, the other not.
  - The LR flagged 16 hosts: 8 malicious, and 8 not malicious

Compliance with ASSERTION 3

Classifier	Validation set (NetFlows) TPR	FPR	#Errors	Test set (NetFlows) TPR	FPR	#Errors	Test set (Hosts) TPR	FPR	#Errors
DT	0.9996	0.0003	45	0.0636	0.0002	238,807	1.000	0.250	2
RF	0.9996	<0.0001	16	0.3706	0.0000	160,515	0.125	0	7
HGB	0.9998	<0.0001	9	0.3729	<0.0001	159,913	0.125	0.125	8
LR	0.9373	0.0711	8,068	0.3734	0.0564	164,722	1.000	1.000	8

Giovanni Apruzzese
giovannia@ru.is

ASSERTION 3. Depending on the specific context, the TPR and the FPR can be incomplete performance metrics: they not necessarily are meaningful indicators of a NIDS' practical utility.

Experiment – Results

Do you think the RF or HGB are still 'the best'?

- TPR and FPR on the test set
- …but what if we simply look at the hosts that triggered at least one alarm?
  - The DT flagged 10 hosts, of which 8 are involved in malicious NetFlows
  - The RF flagged only one 'malicious' host
  - The HGB flagged two hosts – one malicious, the other not.
  - The LR flagged 16 hosts: 8 malicious, and 8 not malicious

Compliance with ASSERTION 3

Classifier	Validation set (NetFlows) TPR	FPR	#Errors	Test set (NetFlows) TPR	FPR	#Errors	Test set (Hosts) TPR	FPR	#Errors
DT	0.9996	0.0003	45	0.0636	0.0002	238,807	1.000	0.250	2
RF	0.9996	<0.0001	16	0.3706	0.0000	160,515	0.125	0	7
HGB	0.9998	<0.0001	9	0.3729	<0.0001	159,913	0.125	0.125	8
LR	0.9373	0.0711	8,068	0.3734	0.0564	164,722	1.000	1.000	8

Giovanni Apruzzese
giovannia@ru.is

Vademecum

#	Question
Q1	Has the notion of “anomaly” been properly defined?
Q2	Has the term “unrealistic” been used unambiguously?
Q3	Does the evasion attack’s methodology *implicitly* assume a compromise of the NIDS (or of its input/output)?
Q4	(if yes to Q3) Has the fact that the NIDS is assumed to be compromised been taken into account, or at least justified?
Q5	Does the evaluation dataset capture the characteristics of the network environment envisioned in the threat model?
Q6	Does the paper claim “generality” of the conclusions—and, if so, is there sufficient evidence to substantiate such a claim?
Q7	Has the TPR/FPR been operationally contextualized (e.g., how does the analyst use the output of a classifier)?

Giovanni Apruzzese
giovannia@ru.is

Vademecum

#	Question
Q1	Has the notion of "anomaly" been properly defined?
Q2	Has the term "unrealistic" been used unambiguously?
Q3	Does the evasion attack's methodology *implicitly* assume a compromise of the NIDS (or of its input/output)?
Q4	(if yes to Q3) Has the fact that the NIDS is assumed to be compromised been taken into account, or at least justified?
Q5	Does the evaluation dataset capture the characteristics of the network environment envisioned in the threat model?
Q6	Does the paper claim "generality" of the conclusions—and, if so, is there sufficient evidence to substantiate such a claim?
Q7	Has the TPR/FPR been operationally contextualized (e.g., how does the analyst use the output of a classifier)?

In our demonstration (§6), we followed our vademecum: "anomaly" and "unrealistic" have never been used; the attacker is not assumed to have compromised the NIDS; the dataset was carefully chosen so as to resemble the envisioned scenario; there is no "generality claim"; and the TPR/FPR have been assessed by considering the post-processed output of a classifier.

Giovanni Apruzzese
giovannia@ru.is

Bangalore, June 4th 2026

21st ACM ASIA Conference on Computer and Communications Security

SoK: Reshaping Research on Network Intrusion Detection Systems

Giovanni Apruzzese

HÁSKÓLINN Í REYKJAVÍK | REYKJAVIK UNIVERSITY

Bangalore, June 4th 2026

21st ACM ASIA Conference on Computer and Communications Security

SoK: Reshaping Research on Network Intrusion Detection Systems

Overall recommendation - Summary of the main reason(s) for your recommendation.

This SoK provides a timely and thought-provoking reflection on long-standing assumptions in Network Intrusion Detection Systems (NIDS) research.

Giovanni Apruzzese

HÁSKÓLINN Í REYKJAVÍK | REYKJAVIK UNIVERSITY

I am hiring ☺

- **Two PhD students (3 years)**
- **One PostDoc (3 years)**

https://giovanniapruzzese.com

Giovanni Apruzzese
giovannia@ru.is

I am hiring ☺

- **Two PhD students (3 years)**
- **One PostDoc (3 years)**

https://giovanniapruzzese.com



PC Co-Chairs: Sahar Abdelnabi, Giovanni Apruzzese, Matthew Jagielski

https://aisec.cc

Looking for PC members

Giovanni Apruzzese
giovannia@ru.is