Bad Design Smells in Benchmark NIDS Datasets

Robert Flood, Gints Engelen, David Aspinall, Lieven Desmet

University of Edinburgh, UK; KU Leuven, Belgium

NIDS Datasets

- Researchers want network datasets for ML!
- Synthetic datasets emerge as a solution to many issues
  - Often, scripted interactions across a cyber range
  - PCAP: Raw network traffic. Large, unwieldy.
  - Flow summaries: Ready for ML pipelines. Bespoke feature sets.
- Much criticism of research network datasets [9, 2]!
  - Mislabelling/Artefacts [8, 6], Ben/Mal Ratio [3], Trivial Classification [7], Poor Generalisation [1] . . .
- Unlike prior work, we scrutinise *design decisions*

Our Work

- To support good research, NIDS datasets require careful design
- '*Design*' $\Rightarrow$ (Un)conscious decisions
- Investigate the impact of questionable design choices on dataset structure and downstream research

Table 1: Dataset Summary

Dataset	Year	Class	Cit.
CIC IDS 2017	2017	14	3264
CIC IDS 2018	2018	16	3264
ICSX 2012	2012	5	1365
UNSW-NB15	2015	10	2817
Ton_IoT	2019	10	254
Bot-IoT	2021	5	1217
CTU-13	2014	13	866

Bad Data Design Smells

Why ‘Smells’?!

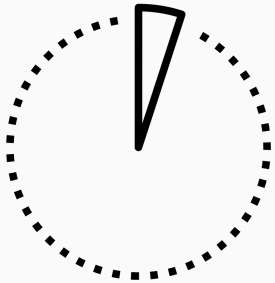
Analogous to design smells in software engineering —- signals of questionable design practises —- we define ***data design smells***.

NB: Public NIDS datasets are precious and we do not claim to invalidate the methodologies/contributions of work relying on them.

Bad Data Design Smells

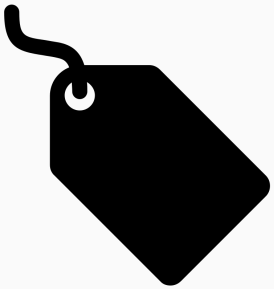
Highly Repetitive

- Low Data Diversity
- Traffic Collapse



Simulation Artefacts

- Highly Dependent Features
- Artificial Diversity



Mislabelled

- Wrong Label
- Unclear Ground Truth

Many Important Questions!

Many Important Questions!

RQ1: Do data smells affect downstream research?

RQ2: Does research use datasets ‘off-the-shelf’?

RQ3: How do we find bad smells?

RQ4: Can we detect bad smells automatically?

RQ5: How do we minimise impact of smelly data?

Research Question 1

Do data smells affect downstream research?

Recreate the methodologies of NIDS papers; uncover potential bad smell bias

RQ1 - Case Study 1: LUCID DoS IDS [5]

LUCID

Packet Level Features

[0.223, 0.322 . . . 0.0]

[0.456, 0.654 . . . 1.0]

[0.333, 0.000 . . . 1.0]

...

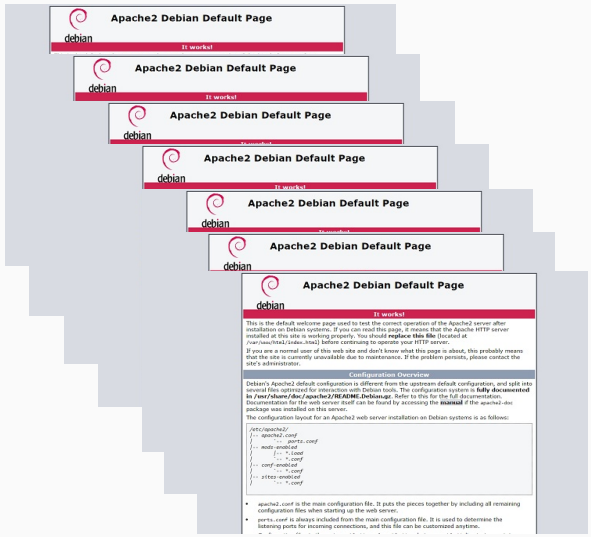


Table 2: LUCID vs Baseline Random Forest

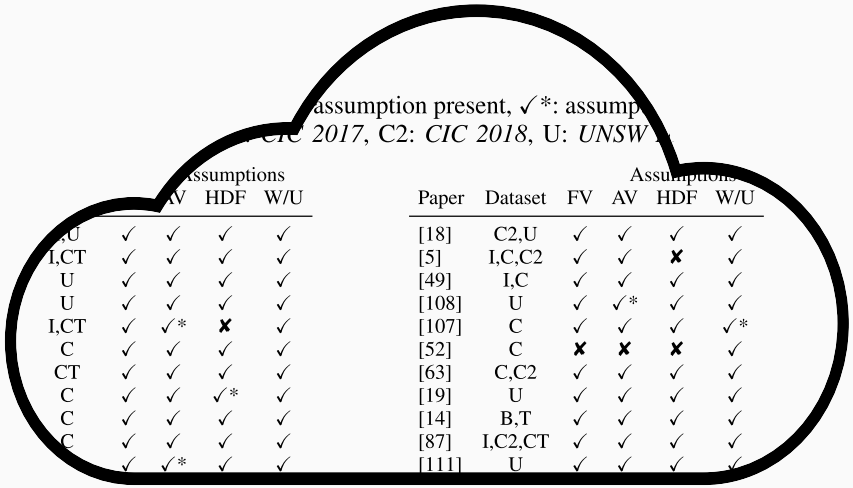
Model	Acc.	F1
LUCID	0.997	0.997
Baseline	0.997	0.997
Baseline (Corrected)	1.000	1.000

Research Question 2

Do papers assume that datasets can be used ‘off-the-shelf’? Do they discuss sources of potential bias?

We investigate assumptions reflecting our smells in papers published in 14 security/networking conferences

RQ2 - NIDS data assumptions



assumption present, ✓*: assump
CIC 2017, C2: *CIC 2018*, U: *UNSW*

	Assumptions			
		AV	HDF	W/U
U	✓	✓	✓	✓
I,CT	✓	✓	✓	✓
U	✓	✓	✓	✓
U	✓	✓	✓	✓
I,CT	✓	✓*	✘	✓
C	✓	✓	✓	✓
CT	✓	✓	✓	✓
C	✓	✓	✓*	✓
C	✓	✓	✓	✓
C	✓	✓	✓	✓
	✓	✓*	✓	✓

Paper	Dataset	FV	Assumptions AV	HDF	W/U
[18]	C2,U	✓	✓	✓	✓
[5]	I,C,C2	✓	✓	✘	✓
[49]	I,C	✓	✓	✓	✓
[108]	U	✓	✓*	✓	✓
[107]	C	✓	✓	✓	✓*
[52]	C	✘	✘	✘	✓
[63]	C,C2	✓	✓	✓	✓
[19]	U	✓	✓	✓	✓
[14]	B,T	✓	✓	✓	✓
[87]	I,C2,CT	✓	✓	✓	✓
[111]	U	✓	✓	✓	✓

Little evidence of raw data auditing ⇒ questionable methods; high (unjustified?) complexity; mislabelled data

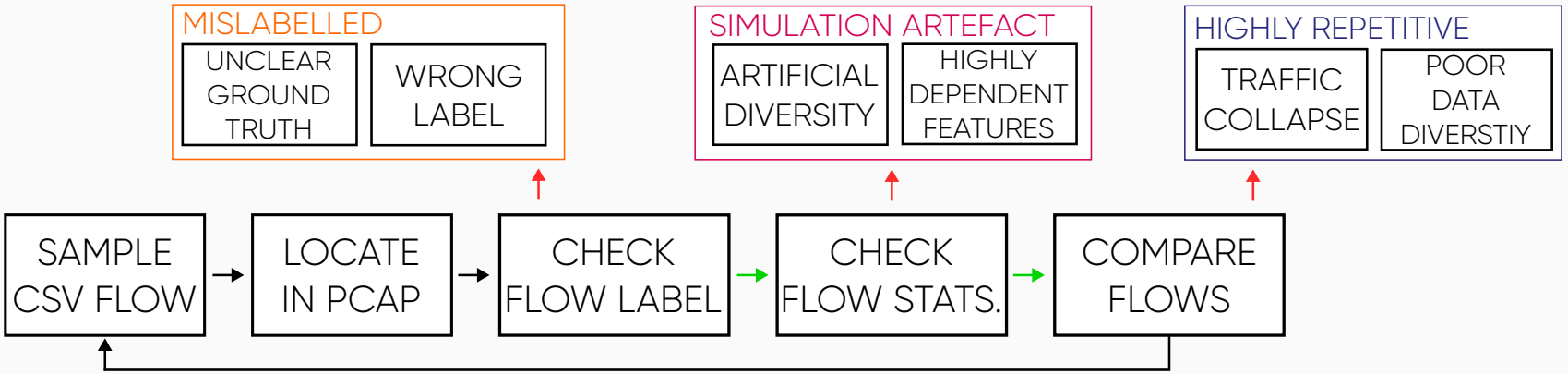
(Some Exceptions!)

Research Question 3

How do we find bad smells? What do they look like in practise?

Exhaustive manual analysis of seven benchmark datasets

RQ3 - Manual Analysis



- CTU-13: 99.9% of a malicious class is mislabelled/misprocessed
- UNSW NB15: Unidirectional flows; no clear effect
- CIC IDS 2018: Attacks launched against closed ports

Research Question 4

Can we detect bad smells automatically?

We develop *heuristic* bad smell measures;
compare with datasets from other domains.

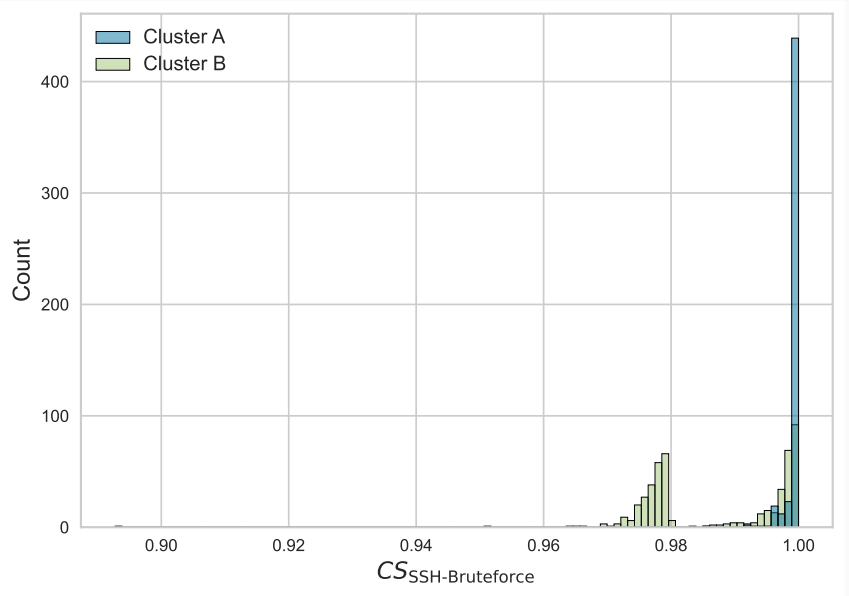
RQ4 - Highly Repetitive Measures

For class C, we find clusters C_i and calculate a weighted cosine similarity as:

$$CS_{C_i} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \quad \text{where} \quad \mathbf{A} \sim C_i,\ \mathbf{B} \sim C_i,$$

$$CS_{C_i} \in [0, 1]$$

Frequently find trivial diversity across all datasets examined (≈33%)



CS_C of two clusters of *CIC IDS 2018 SSH-BruteForce* class, partially launched against a closed port.

Research Question 5

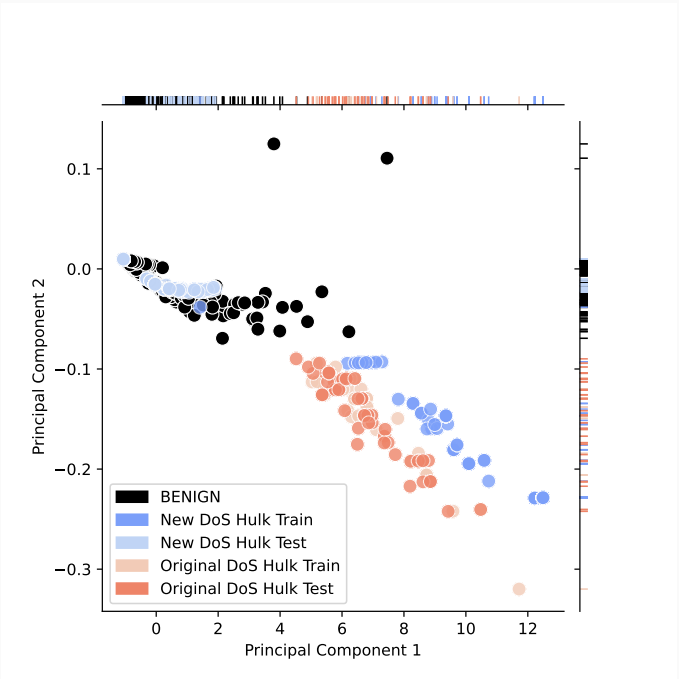
How do we (a) use smelly NIDS data and (b) avoid developing smelly NIDS data?

We offer some recommendations for both using and developing NIDS data.

RQ5 - Recommendations

Explicit Generalisation

- Standard Train/Test Pipeline
  $\Rightarrow$ Training and Evaluating on **Single** Attack
- Instead, success conditions should be outlined clearly
- Use additional datasets/generation frameworks



Overlap between train/test sets on *CIC 17* via standard evaluation pipeline (Original) vs. DetGen [4] data (New).

Caveats & Conclusion

- We investigate seven benchmark datasets, uncover six bad data design smells, undertake an impact analysis and systematic literature overview. We exhaustively analyse seven benchmark NIDS datasets manually, cataloguing design smells. We develop six heuristic metrics and perform an automated prevalence analysis. We recommend some best practises for developing/using NIDS datasets.
- We stress that these datasets are important tools, despite our critiques, and completely sufficient for other tasks
- Caveats: heterogeneity of these datasets $\Rightarrow$ subjectivity; heuristics must be applied sensibly to prevent false positives

Bibliography

Apruzzese, G., Pajola, L., and Conti, M.
The Cross-evaluation of Machine Learning-based Network Intrusion Detection Systems.
IEEE Transactions on Network and Service Management 19, 4 (2022), 5152–5169.

Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., and Rieck, K.
Dos and Don'ts of Machine Learning in Computer Security.
In *Proc. of the USENIX Security Symposium* (2022).

Catillo, M., Del Vecchio, A., Pecchia, A., and Villano, U.
A Critique on the Use of Machine Learning on Public Datasets for Intrusion Detection.
In *International Conference on the Quality of Information and Communications Technology* (2021), Springer, pp. 253–266.

Clausen, H., Flood, R., and Aspinall, D.
Traffic generation using containerization for machine learning.
In *Proceedings of the 2019 Workshop on DYnamic and Novel Advances in Machine Learning and Intelligent Cyber Security* (2019), pp. 1–12.

Doriguzzi-Corin, R., Millar, S., Scott-Hayward, S., Martinez-del Rincon, J., and Siracusa, D.
LUCID: A Practical, Lightweight Deep Learning Solution for DDoS Attack Detection.
IEEE Transactions on Network and Service Management 17, 2 (2020), 876–889.

Engelen, G., Rimmer, V., and Joosen, W.
Troubleshooting an Intrusion Detection Dataset: the CICIDS2017 Case Study.
In *2021 IEEE Security and Privacy Workshops (SPW)* (2021), IEEE, pp. 7–12.

Jacobs, A. S., Beltiukov, R., Willinger, W., Ferreira, R. A., Gupta, A., and Granville, L. Z.
AI/ML for Network Security: The Emperor has no Clothes.
In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (2022), pp. 1537–1551.

Liu, L., Engelen, G., Lynar, T., Essam, D., and Joosen, W.
Error Prevalence in NIDS Datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018.
In *2022 IEEE Conference on Communications and Network Security (CNS)* (2022), IEEE, pp. 254–262.

Sommer, R., and Paxson, V.
Outside the Closed World: On using Machine Learning for Network Intrusion Detection.
In *2010 IEEE symposium on security and privacy* (2010), IEEE, pp. 305–316.