




Dagstuhl Seminar  
Tuesday, July 12th, 2022

**When too good is bad**  
**On the re-use of Datasets in ML Security**

Giovanni Apruzzese



DISCLAIMER:  
Informal talk!



on

ity

Giovanni Apruzzese

Has anybody ever reviewed a NIPS/ICML/ICLR paper?

[BACK] STORY 1

## It all started when...

- Some time ago, I was reviewing the papers for NeurIPS 2022, a total of 5.
- All these papers had a similar structure:
  - An **Introduction**, typically followed with a
  - **Background**, anticipating the main
  - **Method**, which is then subject to the
  - **Experiments** 😊 😊 😊 😊 😊 😊

## It all started when...

- Some time ago, I was reviewing the papers for NeurIPS 2022, a total of 5.
- All these papers had a similar structure:
  - An **Introduction**, typically followed with a
  - **Background**, anticipating the main
  - **Method**, which is then subject to the
  - **Experiments** 😊 😊 😊 😊 😊 😊
- Four (out of five) of these experimented on the well-known MLP dataset, and the results showed the effectiveness of the proposal (of course).
  - I knew the MLP dataset very-well (who doesn't?), so I found it acceptable that the paper went directly to the results, without providing any data-related information.
- The last paper, however...

The screenshot shows a Google Scholar search interface. At the top left is the Google Scholar logo. A search bar contains the text 'mlp dataset' with a magnifying glass icon to its right. Below the search bar, the word 'Articles' is displayed with a blue diamond icon. On the left side, there are filters for 'Any time' (with sub-options: 'Since 2022', 'Since 2021', 'Since 2018', 'Custom range...') and 'Sort by relevance' (with sub-option: 'Sort by date'). The main search result is titled 'Detecting Ponies in photos: the MLP dataset.' by 'R Dash, T Sparkle, A Bloom...' from 'Proceedings of the ...', 1998 - ieeexplore.ieee.org. The abstract text reads: 'Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as features of ponies, with minimal preprocessing. This paper reviews various methods applied to pony detection and compares them on a standard dataset of wild ponies ...'. At the bottom of the result are links for 'Save', 'Cite', 'Cited by 47212', 'Related articles', and 'All 42 versions'.

- Four (out of five) of these experimented on the well-known MLP dataset, and the results showed the effectiveness of the proposal (of course).
  - I knew the MLP dataset very-well (who doesn't?), so I found it acceptable that the paper went directly to the results, without providing any data-related information.
- The last paper, however...

## ...my mind was blown

- The last paper used a dataset I've never heard about: **MNIST**
- My mind started to go awry. *What is this dataset?*
  - *Is it legitimate for the intended scope?*
  - *Has it been used before? What is the performance?*
  - *What data is in it? How big is it?*
  - *Are there any features or preprocessing?*

## ...my mind was blown

- The last paper used a dataset I've never heard about: **MNIST**
- My mind started to go awry. *What is this dataset?*
  - *Is it legitimate for the intended scope?*
  - *Has it been used before? What is the performance?*
  - *What data is in it? How big is it?*
  - *Are there any features or preprocessing?*
- I kept on reading, but my mind was still full of questions.
  - Some were touched (probably?) in the remainder, but I couldn't find a satisfying answer to all of them.
  - Even when I reached the main results, I was still thinking about this "MNIST"
- Eventually, I looked at my watch: I had already spent 8 hours reviewing the paper, and the review was soon due. I was not convinced, so...
  - **Reject / Weak Reject.** "Promising research direction, but I have concerns on the dataset and evaluation"
- I did my duty 😊



## ...my mind was blown

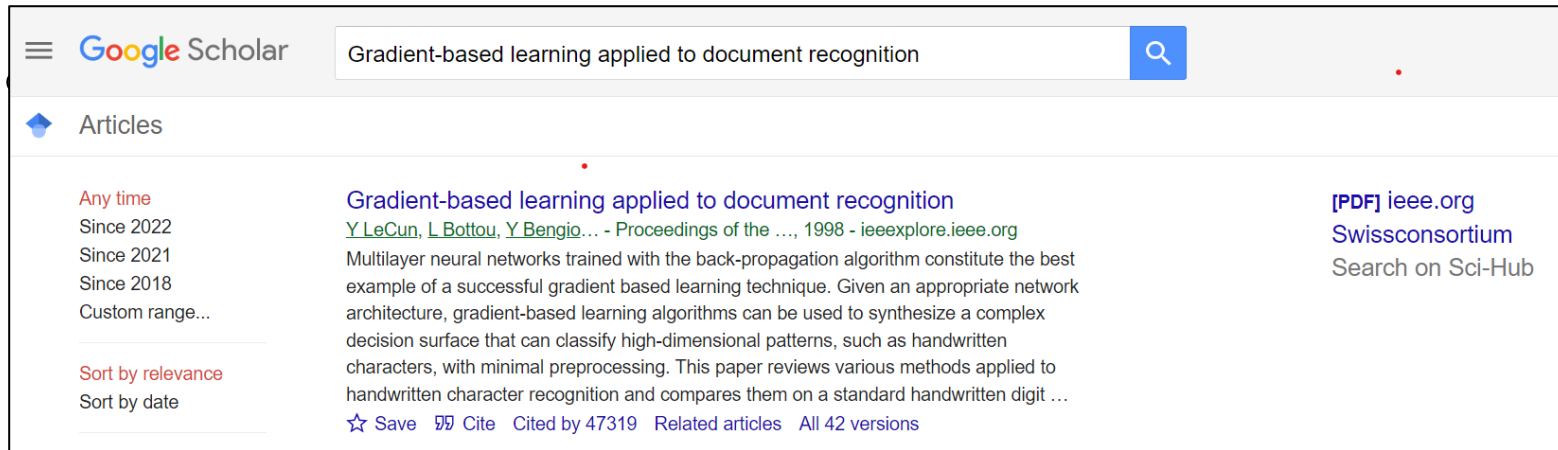
- The last paper used a dataset I've never heard about: **MNIST**
- My mind started to go awry. *What is this dataset?*
  - *Is it legitimate for the intended scope?*
  - *Has it been used before? What is the performance?*
  - *What data is in it? How big is it?*
  - *Are there any features or preprocessing?*
- I kept on reading, but I was still full of questions.
  - Some were touched (probably?) in the paper, but I couldn't find a satisfying answer to all of them.
  - Even when I reached the main results, I was still thinking about this "MNIST"
- Eventually, I looked at my watch: I had already spent 8 hours reviewing the paper, and the review was soon due. I was not convinced, so...
  - **Reject / Weak Reject.** "Promising research direction, but I have concerns on the dataset and validation"
- I did my duty 😞

## ~~It all started when...~~ (TRUTH)

- Some time ago, I was reviewing the papers for NeurIPS 2022, a total of 54.
  - The fifth paper had all authors revealed on the front page and desk rejected
- All these papers had a similar structure:
  - An **Introduction**, typically followed with a
  - **Background**, anticipating the main
  - **Method**, which is then subject to the
  - **Experiments** 😊 😊 😊 😊 😊 😊
- ~~Four (out of five) of these experimented on the well-known MLP dataset, and the results showed the effectiveness of the proposal (of course).~~
  - ~~I knew the MLP dataset very well (who doesn't?), so I found it acceptable that the paper went directly to the results, without providing any data-related information.~~
- The last paper, however...

## ~~It all started when...~~ (TRUTH)

- Some time ago, I was reviewing the papers for NeurIPS 2022, a total of 5 4.
  - The fifth paper had all authors revealed on the front page and desk rejected



The screenshot shows a Google Scholar search result for the paper "Gradient-based learning applied to document recognition" by Y. LeCun, L. Bottou, and Y. Bengio. The search bar contains the text "Gradient-based learning applied to document recognition". The result shows the title, authors, and a brief abstract: "Multilayer neural networks trained with the back-propagation algorithm constitute the best example of a successful gradient based learning technique. Given an appropriate network architecture, gradient-based learning algorithms can be used to synthesize a complex decision surface that can classify high-dimensional patterns, such as handwritten characters, with minimal preprocessing. This paper reviews various methods applied to handwritten character recognition and compares them on a standard handwritten digit ...". The paper is cited by 47319 and has 42 versions. Links for PDF, IEEE.org, Swissconsortium, and Sci-Hub are provided.

- ~~○ Four (out of five) of these experimented on the well-known MLP dataset, and the results showed the effectiveness of the proposal (of course).~~
  - ~~• I knew the MLP dataset very well (who doesn't?), so I found it acceptable that the paper went directly to the results, without providing any data-related information.~~

- The last paper, however...



## So good!

- However, it is true that a lot of papers describe the dataset in just a couple of lines, and nobody complains.

### 5 Experiments

We use image-classification and word-prediction tasks from the federated learning literature.

#### 5.1 Image classification

Following (McMahan et al. 2017), we use CIFAR-10 dataset for our image classification task and train a

**Dataset.** We adopt the dataset that includes 11,613 benign Apps 11,583 malicious Apps from 2011 to 2018 in Malscan [5] to evaluate HRAT (for RQ1-3&5). All Apps are collected from AndroZoo [48] and each sample has been detected by several antivirus systems in

### 6 Evaluation

The algorithms are evaluated by training the Wide-ResNet architecture [38] on the CIFAR-10 and CIFAR-100 datasets [22]. The widening factor is set to 4 and 8 for CIFAR-10 and CIFAR-100 respectively. To facilitate comparison of privacy assured by the two approaches, we train our models

## 3 EXPERIMENTS

### 3.1 DATASETS AND EXPERIMENT SETUP

DBA is evaluated on four classification datasets with non-i.i.d. data distributions: Lending Club Loan Data (LOAN) (Kan, 2019), MNIST, CIFAR-10 and Tiny-imagenet. The data description and parameter setups are summarized in Tb.1. We refer the readers to Appendix A.1 for more details.

## 4 CLASS-WISE ROBUSTNESS ANALYSIS

In this section, we focus on analyzing the class-wise robustness, including class-biased learning and class-relation exploring on six benchmark datasets. Moreover, we investigate the class-wise robustness with different attack and defense models.

We use six benchmark datasets in adversarial training to obtain the corresponding robust model, i.e., MNIST [13], CIFAR-10 & CIFAR-100 [12], SVHN [17], STL-10 [8] and ImageNet [9]. Table 2

### 3.2. Comparison of APGD to usual PGD

We compare our APGD to PGD with Momentum in terms of achieved CE loss and robust accuracy, focusing here on  $l_\infty$ -attacks with perturbation size  $\epsilon$ . We attack the robust models on MNIST and CIFAR-10 from (Madry et al., 2018) and (Zhang et al., 2019b). We run 1000 it-

## So good! (source)

- Xie, Chulin, et al. "Dba: Distributed backdoor attacks against federated learning." *ICLR* 2019.
- Croce, Francesco, and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks." *ICML* 2020.
- Tian, Qi, et al. "Analysis and applications of class-wise robustness in adversarial training." *ACM SIGKDD KDD* 2021.
- Zhao, Kaifa, et al. "Structural attack against graph based android malware detection." *ACM CCS* 2021
- Malek Esmaeili, Mani, et al. "Antipodes of label differential privacy: Pate and alibi." *NeurIPS* 2021
- Bagdasaryan, Eugene, et al. "How to backdoor federated learning." *AISTATS*, 2020.

### 3.2. Comparison of APGD to usual PGD

We compare our APGD to PGD with Momentum in terms of achieved CE loss and robust accuracy, focusing here on  $l_\infty$ -attacks with perturbation size  $\epsilon$ . We attack the robust models on MNIST and CIFAR-10 from (Madry et al., 2018) and (Zhang et al., 2019b). We run 1000 it-

## 3 EXPERIMENTS

### 3.1 DATASETS AND EXPERIMENT SETUP

DBA is evaluated on four classification datasets with non-i.i.d. data distributions: **Lending Club Loan Data(LOAN)**(Kan, 2019), MNIST, CIFAR-10 and Tiny-imagenet. The data description and parameter setups are summarized in Tb.1. We refer the readers to Appendix A.1 for more details.

## 4 CLASS-WISE ROBUSTNESS ANALYSIS

In this section, we focus on analyzing the class-wise robustness, including class-biased learning and class-relation exploring on six benchmark datasets. Moreover, we investigate the class-wise robustness with different attack and defense models.

We use six benchmark datasets in adversarial training to obtain the corresponding robust model, i.e., MNIST [13], CIFAR-10 & CIFAR-100 [12], SVHN [17], STL-10 [8] and ImageNet [9]. Table 2

**Dataset.** We adopt the dataset that includes 11,613 benign Apps and 11,583 malicious Apps from 2011 to 2018 in **Malscan [5]** to evaluate **HRAT (for RQ1-3&5)**. All Apps are collected from **AndroZoo [48]** and each sample has been detected by several antivirus systems in

## 6 Evaluation

The algorithms are evaluated by training the Wide-ResNet architecture [38] on the **CIFAR-10 and CIFAR-100 datasets [22]**. The widening factor is set to 4 and 8 for CIFAR-10 and CIFAR-100 respectively. To facilitate comparison of privacy assured by the two approaches, we train our models

## 5 Experiments

We use image-classification and word-prediction tasks from the federated learning literature.

### 5.1 Image classification

Following (McMahan et al., 2017), we use **CIFAR-10 dataset for our image classification task** and train a

**Was that all science fiction?**

**[TOY] STORY 2**

## A personal, anecdotal, and hence insignificant, Case Study

- Just one week ago, one of my papers was accepted at IEEE Transactions on Network and Service Management.
- The paper performs an extensive analysis of adversarial attacks against 5G Network Infrastructures --- analysis that spans across 6 different case studies.

## A personal, anecdotal, and hence insignificant, Case Study

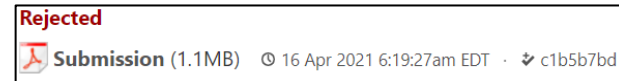
- Just one week ago, one of my papers was accepted at IEEE Transactions on Network and Service Management.
- The paper performs an extensive analysis of adversarial attacks against 5G Network Infrastructures --- analysis that spans across 6 different case studies.
- Before submitting to IEEE TNSM, we submitted the paper:
  - to ACM CCS (Jan 2021)... early reject (Feb 2021)





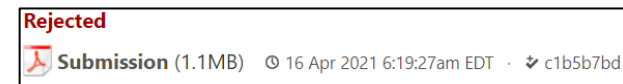
## A personal, anecdotal, and hence insignificant, Case Study

- Just one week ago, one of my papers was accepted at IEEE Transactions on Network and Service Management.
- The paper performs an extensive analysis of adversarial attacks against 5G Network Infrastructures --- analysis that spans across 6 different case studies.
- Before submitting to IEEE TNSM, we submitted the paper:
  - to ACM CCS (Jan 2021)... early reject (Feb 2021)
  - to IEEE SP (Apr 2021)... reject (Jun 2021)



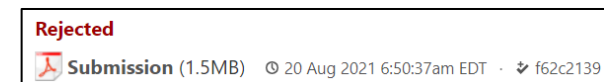
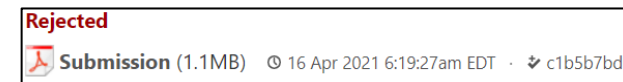
## A personal, anecdotal, and hence insignificant, Case Study

- Just one week ago, one of my papers was accepted at IEEE Transactions on Network and Service Management.
- The paper performs an extensive analysis of adversarial attacks against 5G Network Infrastructures --- analysis that spans across 6 different case studies.
- Before submitting to IEEE TNSM, we submitted the paper:
  - to ACM CCS (Jan 2021)... early reject (Feb 2021)
  - to IEEE SP (Apr 2021)... reject (Jun 2021)
  - to IEEE SP [yes, again!] (Aug 2021)... early reject (Sept 2021)



## A personal, anecdotal, and hence insignificant, Case Study

- Just one week ago, one of my papers was accepted at IEEE Transactions on Network and Service Management.
- The paper performs an extensive analysis of adversarial attacks against 5G Network Infrastructures --- analysis that spans across 6 different case studies.
- Before submitting to IEEE TNSM, we submitted the paper:
  - to ACM CCS (Jan 2021)... early reject (Feb 2021)
  - to IEEE SP (Apr 2021)... reject (Jun 2021)
  - to IEEE SP [yes, again!] (Aug 2021)... early reject (Sept 2021)
  - to USENIX Security (Oct. 2021)... reject and resubmit (Jan 2022)
- We then submitted (as-is) to a journal of a different community.
  - The paper underwent just a single “minor revision” round.



# A personal, anecdotal, and hence insignificant, Case Study

- Just one week ago, one of my papers was accepted at IEEE Transactions on Network and Service Management.
- The paper performs an extensive analysis of adversarial attacks against 5G Network Infrastructures --- analysis that spans across 5 different case studies.
- Before submitting to IEEE TNSM, we submitted the paper to:
  - to ACM CCS (Jan 2021)... early reject (Jan 2021)
  - to IEEE SP (Apr 2021)... reject (Jul 2021)
  - to IEEE SP [yes, I know]... early reject (Sept 2021)
  - to USENIX Security (Oct 2021)... reject and resubmit (Jan 2022)
- We then submitted (as-is) to a journal of a different community.
  - The paper underwent just a single “minor revision” round.

Rejected Round 1  
Submission (1.3MB) 21 Jan 2021 6:31am EST · 5cbc47b2

Rejected  
Submission (1.1MB) 16 Apr 2021 6:19:27am EDT · c1b5b7bd

Rejected  
Submission (1.5MB) 20 Aug 2021 6:50:37am EDT · f62c2139

R2 Reject & Resubmit  
Submission (1.5MB) 13 Oct 2021 4:52:14am PDT · 44822c70

## A personal, anecdotal, and hence insignificant, Case Study [cont'd]

- Over the 4 submissions to security conferences, 12 (+1) people reviewed the paper.

## A personal, anecdotal, and hence insignificant, Case Study [cont'd]

- Over the 4 submissions to security conferences, 12 (+1) people reviewed the paper.
- I realized a few days ago that the same “point” was always raised...

**[CCS]** The use of 5G is misleading, *since the datasets used to build the case studies are rarely 5G. At least 3 out of the 5 datasets are not 5G.* Also, these datasets' volume/quality are highly questionable for any DNN experiments. Specifically: CTU13 is not a 5G network, but a botnet traffic collected before 2014; Deepslice has removed their GitHub data entry so no description to justify for its usage here; RML is a GNU radio raw signal set collected in 2016 (again not 5G, especially since 5G moved to mmWave signals); and Elasticmon [113] contains data from 1 single UE (a single user) based on their entry in CRAWDDAD. Finally, the Irish 5G contains actual traces from Irish 5G deployments, but led to different conclusions from the previous four datasets.

- **[SP1] #A:** Table III provides an overview of the data sets and references the sources, which makes it easier to dig into detail with the different setups. It would be a great improvement to the Section **if the authors described the general structure of these data sets, as right now it's not clear to me what information is part of each set.** Having a rough idea of the dimension, components, and the technological status would be really good.
- **[SP1] #B:** Regarding the experimentation, there is no collected dataset from a reliable open source or closed source software (OAI, free5GC, Open5GCore, AmariSoft, etc). Instead, the authors use available, public datasets. On the one hand, **not all datasets that are used are strictly related to 5G** and, in particular, they are not homogeneous in the sense that they originate from different setups. *An overview of those datasets (number of samples, 4G/5G, how collected, collected by whom and when, etc.) should at least be part of the main text.* On the other hand, there is no evidence provided that these datasets are recorded from correct and reliable experiments. Why can they be blindly trusted? I basically challenge the quality assurance of the input data.

**[SP2]** I am missing an overview of the selected case studies and datasets beyond what is presented in Table II. **Why are these datasets and ML-applications representative for what we can expect to happen in 5G?** Why are parameter choices well justified and how do they impact the results?

- **[USENIX] #A:** I am not sure if this *dataset can appropriately reflect 5G users' application usage/traffic patterns.*
- **[USENIX] #B:** The dataset used for case study 1 is taken from a paper published in 2014 when there were no 5G networks or traces. It is not sure how the experiment results with **such a dataset can faithfully justify the claims** about 5G networks.

## A personal, anecdotal, and hence insignificant, Case Study [cont'd]

- Over the 4 submissions to security conferences, 12 (+1) people reviewed the paper.
- I realized a few days ago that the same “point” was always raised...

**[CCS]** The use of 5G is misleading, since *the datasets used to build the case studies are rarely 5G. At least 3 out of the 5 datasets are not 5G.* Also, these datasets' volume/quality are highly questionable for any DNN experiments. Specifically: CTU13 is not a 5G network, but a botnet traffic collected before 2014; Deepslice has removed their GitHub data entry so no description to justify for its usage here; RML is a GNU radio raw signal set

“What are these datasets???”

sure how the experiment results with such a *dataset* can faithfully justify the claims about 5G networks.

# Abracadabra

*“You never fixed the issue even at the fourth iteration. Your rejection was deserved.”*

- A legitimate observation.



# Abracadabra

*“You never fixed the issue even at the fourth iteration. Your rejection was deserved.”*

- A legitimate observation.

- Unfortunately, the “requested details” were always included in the paper(s),

- be it for CCS...

More details for each CS are provided in Appendix B.

- ...for SP (both iterations)...

Each CS is based on a single dataset. We present an overview of our case studies and corresponding datasets in Table III. Detailed descriptions and additional motivations for these datasets are provided in Appendix A.

of countermeasures or additional comparisons. More technical details on each CS are provided in Appendix A.

- ...and for USENIX...

comparisons. More technical details on each CS are provided in Appendix B.<sup>12</sup>

- ...but most of it in the Appendix 😊

- Note that statements such as “extra details are in the Appendix” were also provided in each of the 6 case studies.

## The dilemma

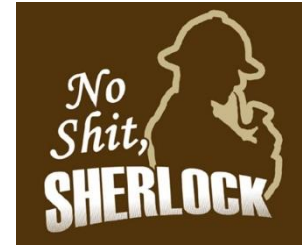
*“But reviewers are not required to read the Appendix!”*

- Another legitimate observation.



# The dilemma

*“But reviewers are not required to read the Appendix!”*



- Another legitimate observation.
- Unfortunately, I do not know how to respond to a similar observation. Some possibilities.
  1. Describe the datasets and all the preprocessing in the main paper.
    - The “main paper” is subject to page limits, and data-preprocessing is (i) “dense” and (ii) hardly passable as a scientific contribution [at least today]
  2. Use only one/two dataset(s): it would require less space.
    - “The attack is evaluated only on one/two dataset(s)! This is not enough!”
  3. What if there are *no* datasets that can be used for a given purpose?
    - Should we write a paper proposing a new dataset, wait for the paper to be accepted, and then use the dataset to validate the “true” contribution?
    - Would you be impressed by an “adversarial attack that works on a self-made dataset”?
  4. Submit to a different community 😊
    - Some reviewers even mentioned this

## So good! ... that is bad?

SOLUTION: Always use well-known “benchmark” datasets.

- Some of our datasets were well-known (e.g., RML2016), but the reviewers still complained

### Radio machine learning dataset generation with gnu radio

[TJ O'shea, N West](#) - Proceedings of the GNU Radio Conference, 2016 - [pubs.gnuradio.org](#)

This paper surveys emerging applications of Machine Learning (ML) to the Radio Signal Processing domain. Provides some brief background on enabling methods and discusses some of the potential advancements for the field. It discusses the critical importance of good datasets for model learning, testing, and evaluation and introduces several public open source synthetic datasets for various radio machine learning tasks. These are intended to provide a robust common baselines for those working in the field and to provide a ...

☆ Save [Cite](#) [Cited by 306](#) [Related articles](#) [»](#)

- (Un)surprisingly (?), we submitted a (different) paper to EuroSP in September 2021: we used 9 “well-known” datasets, covering three diverse security domains, all of which were described in ~20 lines of text. → **None** of the 5 reviewers complained.

For NID, we use: [CTU13](#), [UNB15](#), [IDS17](#). These datasets are well-known in the NID community, and contain network data representing a mixture of simulated and real traffic of large networks. [CTU13](#) is provided as PCAP traces and is focused on *botnet* detection; [UNB15](#) and [IDS17](#) are provided as NetFlows and contain additional malicious activities such as DoS, exploits, or reconnaissance operations.

For PWD, we use: [UCI](#),  [\$\delta\$ Phish](#), [Mendeley](#). These well-known datasets contain information on webpages, such as the URL, the reputation of the website, and the contents of the source HTML. Two ([UCI](#) and [Mendeley](#)) are provided directly as features, while  [\$\delta\$ Phish](#) has raw webpages, from which we extract the features by following established practices [[122](#)].

For MD, we use: [Drebin](#), [Ember](#), [AndMal20](#). These datasets are widely employed for ML-related analyses on malware targeting different OS: [Ember](#) for Windows, [Drebin](#) and [AndMal20](#) for Android. Although [Drebin](#) is becoming outdated (it was collected in 2013), [AndMal20](#) is very recent and serves for a better representation of current trends.

## So good! ... that is bad?



SOLUTION: Always use well-known “benchmark” datasets.

- Some of our datasets were well-known (e.g., RML2016), but the reviewers still complained

### Radio machine learning dataset generation with gnu radio

[TJ O'shea, N West](#) - Proceedings of the GNU Radio Conference, 2016 - [pubs.gnuradio.org](#)

This paper surveys emerging applications of Machine Learning (ML) to the Radio Signal Processing domain. Provides some brief background on enabling methods and discusses some of the potential advancements for the field. It discusses the critical importance of good datasets for model learning, testing, and evaluation and introduces several public open source synthetic datasets for various radio machine learning tasks. These are intended to provide a robust common baselines for those working in the field and to provide a ...

☆ Save  Cite [Cited by 306](#) [Related articles](#) 

- (Un)surprisingly (?), we submitted a (different) paper to EuroSP in September 2021: we used 9 “well-known” datasets, covering three diverse security domains, all of which were described in ~20 lines of text. → **None** of the 5 reviewers complained.

For NID, we use: `CTU13`, `UNB15`, `IDS17`. These datasets are well-known in the NID community, and contain network data representing a mixture of simulated and real traffic of large networks. `CTU13` is provided as PCAP traces and is focused on *botnet* detection; `UNB15` and `IDS17` are provided as NetFlows and contain additional malicious activities such as DoS, exploits, or reconnaissance operations.

For PWD, we use: `UCI`, `δPhish`, `Mendeley`. These well-known datasets contain information on webpages, such as the URL, the reputation of the website, and the contents of the source HTML. Two (`UCI` and `Mendeley`) are provided directly as features, while `δPhish` has raw webpages, from which we extract the features by following established practices [122].

Lesson Learned: if you do research on ML and aim to submit to Security conferences, always use well-known datasets of a well-known security domain (e.g., Malware, or Computer Vision --- the latter being clearly full of attackers).

**Can we argue about “well-known” datasets?**

**[TRUE] STORY 3**



## A short story (this one is real)

- For decades, ML proposals for Network Intrusion Detection (NID) were always evaluated on the same dataset: NSL-KDD (a lot of papers still use it today)

A detailed analysis of the **KDD CUP 99** data set

[M Tavallaee, E Bagheri, W Lu...](#) - 2009 IEEE symposium ..., 2009 - [ieeexplore.ieee.org](http://ieeexplore.ieee.org)



During the last decade, anomaly detection has attracted the attention of many researchers to overcome the weakness of signature-based IDSs in detecting novel attacks, and ...

☆ Save  Cite **Cited by 3711** Related articles All 12 versions 

- The NSL-KDD was collected in 1999 and, obviously:
  - contains attacks that are a solved problem *today*;
  - was captured in a network environment different from those we have *today*.

## A short story (this one is real)

- For decades, ML proposals for Network Intrusion Detection (NID) were always evaluated on the same dataset: NSL-KDD (a lot of papers still use it today)

A detailed analysis of the **KDD CUP 99** data set  
[M Tavallaee, E Bagheri, W Lu...](#) - 2009 IEEE symposium ..., 2009 - [ieeexplore.ieee.org](http://ieeexplore.ieee.org)  
During the last decade, anomaly detection has attracted the attention of many researchers to overcome the weakness of signature-based IDSs in detecting novel attacks, and ...  
☆ Save  Cite **Cited by 3711** Related articles All 12 versions 

- The NSL-KDD was collected in 1999 and, obviously:
  - contains attacks that are a solved problem *today*;
  - was captured in a network environment different from those we have *today*.
- It makes sense (?) that some venues do not accept experiments performed on NSL-KDD anymore.
- According to my sources, evaluations on NSL-KDD are almost stigmatized by experts in NID
  - Especially after the release of new datasets.

*“The data in NSL-KDD is old/flawed, use a more recent dataset, such as CICIDS17”*



## A short story (this one is real) [cont'd]

- Witness the strengths of CICIDS17:
  - Lots of citations.
  - Easy to use:
    1. download (~200MB) and extract,
    2. write (literally) 30 lines of code,
    3. wait 2 minutes (on a laptop),
    4. and enjoy “state-of-the-art” performance.
  - Easy to describe.
    - Just throw a single line in your paper.

[PDF] [Toward generating a new intrusion detection dataset and intrusion traffic characterization.](#)

[I Sharafaldin, AH Lashkari, AA Ghorbani - ICISSp, 2018 - scitepress.org](#)

With exponential growth in the size of computer networks and developed applications, the significant increasing of the potential damage that can be caused by launching attacks is becoming obvious. Meanwhile, Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) are one of the most important defense tools against the sophisticated and ever-growing network attacks. Due to the lack of adequate dataset, anomaly-based approaches in intrusion detection systems are suffering from accurate deployment, analysis ...

☆ Save [Cite](#) [Cited by 1657](#) [Related articles](#) [All 4 versions](#) [»»](#)

Training time:	113.5080296		
Acc:	0.998767		
F1-score:	0.996869		
	<b>col_0</b>	<b>Benign</b>	<b>Malicious</b>
	<b>GT</b>		
	<b>Benign</b>	1134522	836
	<b>Malicious</b>	908	277672

## A short story (this one is real) [cont'd]

- Witness the strengths of CICIDS17:
  - Lots of citations.
  - Easy to use:
    1. download (~200MB) and extract,
    2. write (literally) 30 lines of code,
    3. wait 2 minutes (on a laptop),
    4. and enjoy “state-of-the-art” performance.
  - Easy to describe.
    - Just throw a single line in your paper.



[PDF] Toward generating a new intrusion detection dataset and intrusion traffic characterization.

[I Sharafaldin](#), [AH Lashkari](#), [AA Ghorbani](#) - ICISSp, 2018 - scitepress.org

With exponential growth in the size of computer networks and developed applications, the significant increasing of the potential damage that can be caused by launching attacks is becoming obvious. Meanwhile, Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) are one of the most important defense tools against the sophisticated and ever-growing network attacks. Due to the lack of adequate dataset, anomaly-based approaches in intrusion detection systems are suffering from accurate deployment, analysis ...

☆ Save 📄 Cite Cited by 1657 Related articles All 4 versions ⌕

```
Training time: 113.5080296
Acc: 0.998767
F1-score: 0.996869
```

	col_0	Benign	Malicious
GT			
Benign		1134522	836
Malicious		908	277672

- YES! Finally we have new and public data that can be used for NID!
- We can finally replace the NSL-KDD!

(what was the problem of NSL-KDD again?)

## A short story (this one is real) [END]

- One day, in Summer last year (2021), I had this paper appear on my feed:

[Troubleshooting an intrusion detection dataset: the CICIDS2017 case study](#)

G Engelen, [V Rimmer](#), W Joosen - 2021 IEEE Security and ..., 2021 - [ieeexplore.ieee.org](https://ieeexplore.ieee.org)

Numerous studies have demonstrated the effectiveness of machine learning techniques in application to network intrusion detection. And yet, the adoption of machine learning for securing large-scale network environments remains challenging. The community acknowledges that network security presents unique challenges for machine learning, and the lack of training data representative of modern traffic remains one of the most intractable issues. New attempts are continuously made to develop high quality benchmark datasets ...

★ Save [Cite](#) Cited by 11 [Related articles](#) [All 5 versions](#) [↔](#)

## A short story (this one is real) [END]

- One day, in Summer last year (2021), I had this paper appear on my feed:

[Troubleshooting an intrusion detection dataset: the CICIDS2017 case study](#)  
G Engelen, V Rimmer, W Joosen - 2021 IEEE Security and ..., 2021 - ieeexplore.ieee.org  
Numerous studies have demonstrated the effectiveness of machine learning techniques in application to network intrusion detection. And yet, the adoption of machine learning for securing large-scale network environments remains challenging. The community acknowledges that network security presents unique challenges for machine learning, and the lack of training data representative of modern traffic remains one of the most intractable issues. New attempts are continuously made to develop high quality benchmark datasets ...  
★ Save  Cite Cited by 11 Related articles All 5 versions 

- Should I now question all papers using the original version of CICIDS17?
  - If not, then why all the stigma towards NSL-KDD?
  - If yes, then should I also question all experiments on datasets that have not been “troubleshooted” yet?
  - And what if new research finds “flaws” in the method applied by Engelen et al.?
- SOLUTION: avoid “blindly” using a dataset, and always precisely describe (in the paper) all the preprocessing operations.
  - But what about space limitations?
  - And wouldn’t this defeat the entire purpose of using “benchmark” datasets?

## A short story (this one is real) [END]

- One day, in Summer last year (2021), I had this paper appear on my feed:

[Troubleshooting an intrusion detection dataset: the CICIDS2017 case study](#)

G Engelen, V Rimmer, W Joosen - 2021 IEEE Security and ..., 2021 - ieeexplore.ieee.org

Numerous studies have demonstrated the effectiveness of machine learning techniques in application to network intrusion detection. And yet, the adoption of machine learning for securing large-scale network environments remains challenging. The community acknowledges that network security presents unique challenges for machine learning, and the lack of training data representative of modern traffic remains one of the most intractable issues. New attempts are continuously made to develop high quality benchmark datasets ...

★ Save [Cite](#) Cited by 11 [Related articles](#) [All 5 versions](#) [↗](#)

- Should I now question all papers using the original version of CICIDS17?
  - If not, then why all the stigma towards NSL-KDD?
  - If yes, then should I also question all experiments on datasets that have not been “troubleshooted” yet?
  - And what if new research finds “flaws” in the method applied by Engelen et al.?
- SOLUTION: avoid “blindly” using a dataset, and always precisely describe (in the paper) all the preprocessing operations.
  - But what about space limitations?
  - And wouldn’t this defeat the entire purpose of using “benchmark” datasets?

What are we - as a community - doing???